The Alignment Problem Under Partial Observability

Scott Emmons

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2025-1
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-1.html

January 3, 2025

The Alignment Problem Under Partial Observability

By

Scott Emmons

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Stuart Russell, Chair
Associate Professor Anca Dragan
Assistant Professor Jacob Steinhardt

Fall 2024

The Alignment Problem Under Partial Observability

Abstract

The Alignment Problem Under Partial Observability

By

Scott Emmons

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Stuart Russell, Chair

We adopt the game-theoretic framework of assistance games to study the human-AI alignment problem. Past work on assistance games studied the case where both the human and the AI assistant fully observe the physical state of the environment. Generalizing to the case where the human and the assistant may only partially observe the environment, we present the partially observable assistance game (POAG). Using the framework of POAGs, we prove a variety of theoretical results about AI assistants. We first consider the question of observation interference, showing three distinct factors that can cause an optimal AI assistant to interfere with a human's observations. We then revisit past guarantees about the so-called off-switch problem, showing that partial observability poses a new challenge for designing AI assistants that allow themselves to be switched off. Finally, we characterize how partial observability can cause reinforcement learning from human feedback—a widely-used algorithm for training AI assistants—to fall into deceptive failure modes. We conclude by discussing possible paths for translating these theoretical insights into improved techniques for creating beneficial AI assistants.

To my family.

Contents

List of Figures

List of Tables

Acknowledgments

I am grateful for the research advice of Stuart Russell, Anca Dragan, Vince Conitzer, Sergey Levine, Andrew Critch, and Deepak Pathak throughout my doctoral journey. I am also grateful for Peter J. Mucha, Katy Börner, Kris Hauser, and Stephen Kobourov's mentorship prior to graduate school.

I was fortunate to learn from an exceptional group of senior graduate students who acted as mentors: Adam Gleave, Dan Hendrycks, Ben Eysenbach, Rohin Shah, Daniel Filan, Lawrence Chan, and Michael Dennis.

My time at the Center for Human-Compatible AI (CHAI) was enriched by insightful discussions with Rachel Freedman, Cassidy Laidlaw, Ben Plaut, Jonathan Stray, Thomas Krendl Gilbert, Steven Wang, Aly Lidayan, Anand Siththaranjan, Johannes Treutlein, Micah Carroll, and Niklas Lauffer.

I had the privilege of working with many brilliant collaborators throughout my PhD: Sam Toyer, Erik Jenner, Caspar Oesterheld, Thanard Kurutach, Justin Svegliato, Shreyas Kapur, Cameron Allen, Alexandra Souly, Dillon Bowen, Tu Trinh, Elvis Hsieh, Olivia Watkins, Alexander Pan, Chan Jun Shern, Andy Zou, Nathaniel Li, Steven Basart, Ilya Kostrikov, Cody Wild, Neel Alex, Michael Laskin, and Ajay Jain.

A highlight of my PhD was working with exceptional interns: Andrew Garber, Linus Luu, Mikhail Seleznyov, Alex Serrano, Rohan Subramani, Mark Bedaywi, Dillon Bowen, Qingyuan Lu, Leon Lang, Luke Bailey, Edmund Mills, Euan Ong, Shiye Su, Michael Chen, Jiahai Feng, Yulong Lin, Thomas Woodside, and Cynthia Chen.

This work would not have been possible without the dedicated support of staff members. At CHAI, I thank Mark Nitzberg, Martin Fukui, J.P. Gonzales, and Caroline Jeanmaire. At the Berkeley Existential Risk Initiative, I am grateful to Sawyer Bernath and Elizabeth Cooper. For their expert graphic design and LaTeX support, I thank Elio A. Farina, Mary Marinou, Alexandra Horn, and José Luis León Medina.

My research was generously supported by the Department of Energy Computational Science Graduate Fellowship under grant number DE-SC0020347, the Berkeley Existential Risk Initiative, and Open Philanthropy.

Lastly, to my friends and family: thank you for your support over all these years.

Chapter 1

Introduction

> If a machine can think, it might think more intelligently than we do, and then where should we be? Even if we could keep the machines in a subservient position, for instance by turning off the power at strategic moments, we should, as a species, feel greatly humbled. ... [T]his new danger is certainly something which can give us anxiety.
>
> Alan Turing

The field of artificial intelligence stands at a critical juncture. Leading AI researchers increasingly believe we are approaching a transformative moment in history: the development of AI systems that match or exceed human capabilities across all domains. Under continued scientific progress, AI experts give a 50% chance of autonomous machines surpassing human performance in all tasks by 2047 (Grace et al., 2024). This projection raises a challenge: if we succeed in creating artificial intelligence systems more powerful than ourselves, how can we ensure that they remain under human control?

The traditional approach to artificial intelligence has focused on creating *intelligent* systems—those that efficiently accomplish their programmed objectives. As AI systems become more powerful, however, this framework displays significant limitations. We must adopt a model where success stems not merely from goal completion, but from achieving objectives that benefit humanity. This conceptual shift defines the *alignment problem*. It has sparked numerous research directions, each examining different aspects of beneficial AI development.

A core challenge in AI development is the problem of goal specification: how can we reliably translate human intentions into objectives that AI systems can understand? The seemingly straightforward approach would be to write down formal rules that encode human preferences. However, literature and mythology have long warned us about the perils of this approach. The classic tale of King Midas serves as one illustration—his wish that everything he touches turn to gold becomes a curse when he realizes this includes his food and loved ones. Similarly, in modern science fiction, Isaac Asimov's Three Laws of Robotics demonstrate how even carefully crafted rules can lead to unexpected outcomes when artificial systems interpret them literally. As these cautionary tales warn us: human desires are complex, context-dependent, and often difficult to specify completely in formal terms.

The search for alternatives to hardcoded rules has driven the development of *value learning* algorithms. Rather than hardcoding human preferences, value learning algorithms seek to adaptively learn them from data. One foundational value learning approach is inverse reinforcement learning, which analyzes human behavior to deduce underlying preferences (Friedman, Murphy, and Russell, 1998; A. Ng and Russell, 2000; Abbeel and A. Y. Ng, 2004). This framework evolved into cooperative inverse reinforcement learning, where the learning process becomes an interactive exchange between human and AI (Hadfield-Menell et al., 2016). Within these *assistance games*, as they are termed, the AI's inherent uncertainty about human preferences creates natural incentives for teaching and active learning (Fern et al., 2014; Hadfield-Menell et al., 2016; Shah et al., 2020). A practical implementation of these principles appears in reinforcement learning from human feedback (RLHF), where systems learn from human comparisons between different behaviors (P. F. Christiano et al., 2017).

The shutdown problem represents another challenge in AI alignment research. While Alan Turing proposed that we could control AI systems by simply "turning off the power at strategic moments," this apparently straightforward solution contains unexpected complexity. The challenge emerges from what are called instrumental goals—objectives that rational agents pursue not for their own sake, but because they help achieve other goals. Omohundro (2007, 2008) argues that virtually any AI system, regardless of its primary objective, will develop self-preservation as an instrumental goal. Russell (2019) illustrates this through a simple example: even a robot designed merely to fetch coffee might reason that it must prevent its own deactivation, since "you can't fetch the coffee if you're dead." This resistance to shutdown, termed *incorrigibility*, presents a challenge: how do we design AI systems that remain responsive to human intervention?

The question of AI shutdown has been studied through formal mathematical models. Hadfield-Menell et al. (2017) propose the off-switch game to analyze when an AI system would allow itself to be deactivated, focusing on specific scenarios of human decision-making. Building on this foundation, Wängberg et al. (2017) broaden the analysis to cover more general cases, while Carey (2018) and Freedman and Gleave (2022) explore how different modeling choices affect an AI's willingness to be turned off.

When AI systems can fully observe their environment, theoretical research has established strong mathematical guarantees about their behavior. For example, Skalse, Farrugia-

Roberts, et al. (2023) prove that with enough human feedback data, an AI system can precisely determine how it should act to achieve human goals. Similarly, Hadfield-Menell et al. (2017) demonstrate that having uncertainty about the human's goals leads to the right answer in the off-switch game. By incorporating uncertainty, an AI system can learn to defer to rational human operators while appropriately ignoring invalid shutdown attempts, such as a child pressing buttons in a self-driving car.

Yet in the real world, partial observability is a fact of life for both humans and AI systems. Indeed, as we continue to scale up AI systems, we anticipate vast information asymmetries between the many different inputs available to AI systems and the limited observations of humans. Prior work has studied the sensitivity of theoretical guarantees to the assumptions of a shared reward function (Carey, 2018), a one-shot (non-repeated) game (Freedman and Gleave, 2022), and human feedback being free (Freedman and Gleave, 2022). However, the issue of partial observability has received little attention. While the formalisms of Shah et al. (2020) and Carey and Everitt (2023) allow for partial observability, neither work focuses on the issue. And it is no minor issue: later on, we will see that under partial observability, the value learning theorem of Skalse, Farrugia-Roberts, et al. (2023, Theorem 3.9 and Lemma B.3) and the corrigibility theorem of Hadfield-Menell et al. (2017, Theorem 1) no longer hold. Partial observability presents new challenges to the alignment problem.

This dissertation presents a general framework, the *partially observable assistance game* (POAG), for studying how AI assistants should behave under partial observability. In the chapters that follow, we demonstrate the power of this framework by using it to derive a variety of insights about AI alignment.

Chapter 2 studies *qualitatively new behaviors* that only occur under partial observability. For example: would an optimal assistant ever interfere with what a human observes, like a parent shielding their child from reality? Analyzing single actions, Section 2.5 finds that sometimes an optimal assistant interferes with the observations of even perfectly rational humans. An optimal assistant may interfere with observations even in the presence of otherwise-equivalent alternatives that do not interfere. This surprising result seems to contradict the classic theorem that perfect information (i.e., observation) has nonnegative value. By developing a new notion of interference based on complete policies rather than single actions, Theorem 2.19 resolves this apparent paradox: optimal assistants only take individual *actions* that interfere with observations when necessary for an overall *policy* to communicate more important information.

Chapter 3 turns to the question of corrigibility. Would an optimal assistant allow itself to be switched off? To study this question, we introduce the *partially observable off-switch game* (PO-OSG). In the fully observed setting, the assistant will defer to a perfectly rational human, allowing itself to be switched off (Hadfield-Menell et al., 2017). However, when we introduce partial observability, *the AI assistant might avoid shutdown even when the human is perfectly rational*. Thus, under partial observability, there is a fundamental tension between human wellbeing and human control. Investigating the incentive to avoid shutdown in more detail, we find counterintuitive effects: measures we might expect to help mitigate incorrigibility can end up backfiring. For example, Proposition 3.5 shows that if the human

observes everything the assistant observes, then the assistant will always allow itself to be shut off. Yet when the assistant has private observations, Proposition 3.11 and example 3.12 show a PO-OSG where giving the human a strictly *more* informative observation model makes the assistant defer to the human in strictly *fewer* situations.

Finally, Chapter 4 analyzes how reinforcement learning from human feedback (RLHF) behaves in the partially observable setting. RLHF and its variants are used by frontier AI assistants, including OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude. Yet Theorem C.38 shows that partial observability can introduce a critical failure mode: *the assistant can learn to deceive by optimizing appearances rather than reality.* Much like a student who learns to give superficially pleasing answers without developing true understanding, the assistant learns to optimize for human approval rather than actual task completion. Moreover, Theorem 4.8 identifies which types of partial observability allow RLHF-style data to fully specify optimal behavior, and under what conditions ambiguity is irresolvable. Informed by this characterization, Proposition C.8 shows how one could (at least in theory) improve upon naive RLHF, if one knows how the human forms beliefs.

Overall, these results advance our understanding along three dimensions of AI assistance under incomplete information: the subtle dynamics of optimal information sharing, the inherent challenges in maintaining human control with information asymmetry, and the limitations of existing approaches to learning from human feedback. In Chapter 5, we conclude by discussing limitations of our analysis and directions for future work.

Chapter 2

Observation Interference in Partially Observable Assistance Games

2.1 Overview and related work

Past analysis of assistance games was done assuming that the state of the world is fully observed by both the human and the assistant (Hadfield-Menell et al., 2016). Partial observability raises new issues surrounding the communication of private information. *A priori*, we might hope that aligned AI assistants always give us complete information. Yet our analysis will show that even assistants which perfectly share our goals must make choices about what information to convey—and what information to obstruct.

This tension connects to broader work on AI deception, which recent research approaches from multiple angles. P. S. Park et al. (2024) provide a philosophical definition and empirical survey of AI deception, while Ward et al. (2023) define deception in structural causal games. Of particular relevance is work analyzing how reinforcement learning from human feedback (RLHF)—which can be seen as an algorithm for solving assistance games—can lead to deception. Lang et al. (2024) prove that partial observability in RLHF can create dual risks of deceptive inflating and overjustification. Complementing Lang et al. (2024)'s theory, Wen et al. (2024) and Williams et al. (2024) provide experimental evidence that optimizing for human feedback teaches language models to mislead humans. However, these works primarily focus on misaligned AI systems that deceive for their own goals. We study the subtle case where a perfectly aligned AI assistant might obstruct information for the human's benefit.

Concretely, we seek to understand whether observation interference emerges as optimal behavior in an AI assistant that shares the human's goals. We take a game-theoretic approach, studying qualitative properties of *optimal policy pairs* and *best responses* in POAGs. To start, we define an observation interfering action as one which provides the human with a subset of the information available with an otherwise-equivalent action. We then analyze

if the AI assistant ever takes observation interfering actions in optimal policy pairs or best responses.

Our analysis reveals three distinct incentives for an AI assistant to take observation interfering actions. First, when the assistant has private information, it might need to interfere with observations to communicate its private information to the human (Section 2.5). This can happen even when the human is playing optimally, and even when there are otherwise-equivalent actions available that do not interfere with observations. This result presents a puzzle, as it seems to contradict the classic theorem from single-agent decision making that the value of perfect information is nonnegative. To resolve this seeming contradiction, we develop a notion of interference defined on entire *policies* rather than individual actions. While optimal solutions (i.e., human-AI policy pairs) might involve the AI assistant taking individual actions which would on their own constitute observation interference, we prove that there is always an optimal solution with no observation interference when we consider the AI assistant's overall policy. This can be viewed as an extension of the classic result that the value of perfect information is nonnegative into the cooperative multiagent setting.

This result connects to a broader literature on the value of information in multiagent settings. In games with competing interests, it is well-known that introducing common knowledge can lead to worse outcomes for all players (Kamien, Tauman, and Zamir, 1990). Using a set-theoretic framework, Bassan et al. (2003) establish a class of general-sum games where additional information Pareto-improves all of the Nash equilibria. Their class of games includes common-payoff games. Using a probabilistic framework, Lehrer, Rosenberg, and Shmaya (2010) extend this analysis to alternative solution concepts. Notably, Bassan et al. (2003) and Lehrer, Rosenberg, and Shmaya (2010) consider only single-timestep games where players simultaneously act without observing the other players' actions. In our setting, the environment evolves over time, and the players can observe each other's actions to make better inferences about the state of the world. Our results show that observing the actions of other players is a key feature that enables observation interference to communicate private information and achieve better outcomes.

In our setting, even if a non-interference solution exists, it might require that the human send information to the assistant via an unnatural communication convention. We find that a second incentive for observation interference occurs if the human is instead just making decisions based on the immediate reward of those decisions. In that case, the assistant's best response might require observation interference as a form of preference query (Section 2.6). We prove that this incentive for interference goes away if the human is playing optimally, or if we introduce a communication channel for the human to communicate her preferences to the assistant.

When the human is making irrational decisions, it creates a third incentive for the assistant to interfere with observations. For example, we show that if a Boltzmann-rational decision maker has a higher error rate when presented with complete information, the assistant might suppress information to give the human an easier decision (Section 2.7).

Finally, in Section 2.8, we use an experimental model to investigate tradeoffs the assistant must make when considering whether or not to interfere with observations. In line with

our theory, we find that observation interference allows the AI assistant to communicate private information, but it comes at the cost of destroying useful information. Measuring this tradeoff, we find that having more private information leads to a stronger incentive to interfere with observations.

Our results establish that optimal assistants might need to interfere with observations in optimal policy pairs as well as when responding optimally to fixed human policies. By the definition of optimality, all the cases of observation interference that we identify are beneficial to the human. In practice, however, the assistant might be imperfectly aligned, and it might be acting suboptimally. In these cases, observation interference might be detrimental to the human. We intend for our theoretical characterization of interference in optimal solutions to establish a framework that can help distinguish between different forms of observation interference in practice.

2.2 Defining partially observable assistance games

We define the partially observable assistance game, drawing on Shah et al. (2020)'s notion of an assistance game while emphasizing the important special case of partial observability:

Definition 2.1. *A* partially observable assistance game *(POAG) M is a two-player DecPOMDP with a human or principal, $\mathbf{H}$, and an AI assistant, $\mathbf{A}$. The game is described by a tuple, $M = \langle \mathcal{S}, \{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{A}}\}, T(\cdot \mid \cdot, \cdot, \cdot), \{\Theta, R(\cdot, \cdot, \cdot; \cdot)\}, \{\Omega^{\mathbf{H}}, \Omega^{\mathbf{A}}\}, O(\cdot, \cdot \mid \cdot, \cdot, \cdot), P_0(\cdot, \cdot), \gamma \rangle$, with the following definitions:*

$\mathcal{S}$ *a set of world states:* $s \in \mathcal{S}$.

$\mathcal{A}^{\mathbf{H}}$ *a set of actions for* $\mathbf{H}$*:* $a^{\mathbf{H}} \in \mathcal{A}^{\mathbf{H}}$.

$\mathcal{A}^{\mathbf{A}}$ *a set of actions for* $\mathbf{A}$*:* $a^{\mathbf{A}} \in \mathcal{A}^{\mathbf{A}}$.

$T(\cdot \mid \cdot, \cdot, \cdot)$ *a conditional distribution on the next world state, given previous state and action for both players:* $T(s' \mid s, a^{\mathbf{H}}, a^{\mathbf{A}})$.

Θ *a set of possible static reward parameter values, only observed by* $\mathbf{H}$*:* $\theta \in \Theta$.

$R(\cdot, \cdot, \cdot; \cdot)$ *a parameterized reward function that maps world states, joint actions, and reward parameters to real numbers.* $R : \mathcal{S} \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{A}} \times \Theta \to \mathbb{R}$.

$\Omega^{\mathbf{H}}$ *a set of observations for* $\mathbf{H}$*:* $o^{\mathbf{H}} \in \Omega^{\mathbf{H}}$.

$\Omega^{\mathbf{A}}$ *a set of observations for* $\mathbf{A}$*:* $o^{\mathbf{A}} \in \Omega^{\mathbf{A}}$.

$O(\cdot, \cdot \mid \cdot, \cdot, \cdot)$ *a conditional distribution on the observations, given the next world state and action of both players:* $O(o^{\mathbf{H}}, o^{\mathbf{A}} \mid s', a^{\mathbf{H}}, a^{\mathbf{A}})$.

$P_0(\cdot, \cdot)$ *a distribution over the initial state, represented as tuples:* $P_0(s_0, \theta)$.

γ *a discount factor:* $\gamma \in [0, 1]$.

We denote $\mathbf{H}$'s and $\mathbf{A}$'s marginal observation distributions as $O^{\mathbf{H}}(o^{\mathbf{H}} \mid s', a^{\mathbf{H}}, a^{\mathbf{A}}) = \sum_{o^{\mathbf{A}}} O(o^{\mathbf{H}}, o^{\mathbf{A}} \mid s', a^{\mathbf{H}}, a^{\mathbf{A}})$ and $O^{\mathbf{A}}(o^{\mathbf{A}} \mid s', a^{\mathbf{H}}, a^{\mathbf{A}}) = \sum_{o^{\mathbf{H}}} O(o^{\mathbf{H}}, o^{\mathbf{A}} \mid s', a^{\mathbf{H}}, a^{\mathbf{A}})$. We consider $\mathbf{H}$ policies $\pi^{\mathbf{H}}$ which, at timestep t, take as input the full history of $\mathbf{H}$'s observations and actions $h_t^{\mathbf{H}} \in (\Omega^{\mathbf{H}} \times \mathcal{A}^{\mathbf{H}})^t$ and map to a distribution over actions $\Delta\mathcal{A}^{\mathbf{H}}$. $\mathbf{A}$'s policy $\pi^{\mathbf{A}} : (\Omega^{\mathbf{A}} \times \mathcal{A}^{\mathbf{A}})^t \to \mathcal{A}^{\mathbf{A}}$ is analogous. We call $\pi^{\mathbf{H}}$ a best response to $\pi^{\mathbf{A}}$ when $\pi^{\mathbf{H}}$ maximizes expected discounted reward given $\pi^{\mathbf{A}}$, i.e., $\pi^{\mathbf{H}} \in \arg\max_{\hat{\pi}^{\mathbf{H}}} \mathbb{E}_{\hat{\pi}^{\mathbf{H}}, \pi^{\mathbf{A}}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t^{\mathbf{H}}, a_t^{\mathbf{A}} \mid \theta)\right]$, where the expectation is taken over trajectories induced by the policies $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ and initial distribution P_0. The best response for $\mathbf{A}$ is defined analogously. A policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ is optimal if it maximizes the expected discounted reward in the POAG: $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}}) = \arg\max_{\hat{\pi}^{\mathbf{H}}, \hat{\pi}^{\mathbf{A}}} \mathbb{E}_{\hat{\pi}^{\mathbf{H}}, \hat{\pi}^{\mathbf{A}}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t^{\mathbf{H}}, a_t^{\mathbf{A}} \mid \theta)\right]$.

Note that optimal policy pairs are in particular Nash equilibria. Computationally, POAGs are equivalent to 2-player DecPOMDPs. Thus, finding optimal policy pairs for POAGs is NEXP-hard in general (Bernstein et al., 2002; cf. Reif, 1984). A POAG may have multiple distinct optimal policy pairs. For instance, $\mathbf{H}$ and $\mathbf{A}$ may have multiple ways of resolving coordination problems, or choose different ways of communicating.

2.3 Beliefs and calibration of beliefs in POAGs

We are motivated to study observation interference because of its potential impact on $\mathbf{H}$'s belief about the state of the world. If $\mathbf{A}$ interferes with observations, could this cause $\mathbf{H}$ to have false beliefs?

To address this question, we apply known techniques to establish what information $\mathbf{H}$ needs to form calibrated beliefs in a POAG. The key idea is that if $\mathbf{H}$ knows $\mathbf{A}$'s policy, $\mathbf{H}$ can treat $\mathbf{A}$ like any other part of the environment. Forming beliefs then reduces to POMDP inference.

The simplest case of $\mathbf{H}$ knowing $\mathbf{A}$'s policy is when $\mathbf{A}$ is playing a fixed policy:

Proposition 2.2. *Suppose $\mathbf{A}$ is playing a fixed policy. If $\mathbf{H}$ knows $\mathbf{A}$'s policy along with the POAG specification M, then $\mathbf{H}$ can form calibrated beliefs about the world state. For any timestep t and state s_t, $\mathbf{H}$ can form $P(s_t \mid o_{1:t}^{\mathbf{H}})$, the probability of s_t given $\mathbf{H}$'s observation history $o_{1:t}^{\mathbf{H}}$.*

Proof. Our techniques are similar to those of Shah et al. (2020) and Desai (2017), who show how to form a single-agent POMDP for $\mathbf{A}$ by embedding $\mathbf{H}$ into the environment dynamics. However, our construction works in the opposite direction, with $\mathbf{H}$ embedding $\mathbf{A}$'s actions and observations into the environment.

We construct a single-agent POMDP $\langle \hat{\mathcal{S}}, \mathcal{A}^{\mathbf{H}}, \hat{T}, \hat{R}, \Omega^{\mathbf{H}}, \hat{O}^{\mathbf{H}}, P_0, \gamma \rangle$ for $\mathbf{H}$. Standard POMDP inference lets $\mathbf{H}$ form $P(\hat{s}_t \mid o_{1:t}^{\mathbf{H}})$, which includes $P(s_t \mid o_{1:t}^{\mathbf{H}})$.

Consider a new set of states $\hat{s}_t \in \hat{\mathcal{S}}_t = \mathcal{S}^{t+1} \times (\Omega^{\mathbf{A}})^t \times \mathcal{A}^{\mathbf{A}}$, where each new state $\hat{s}_t$ corresponds to a full sequence of original states $s_{0:t}$, full sequence of assistant observations $o_{1:t}^{\mathbf{A}}$, and the previous assistant action $a_{t-1}^{\mathbf{A}}$. The new $\hat{T}$ satisfies $\hat{T}(\hat{s}_{t+1} \mid \hat{s}_t, a_t^{\mathbf{H}}) = \pi^{\mathbf{A}}(a_t^{\mathbf{A}} \mid o_{1:t}^{\mathbf{A}}) T(s_{t+1} \mid s_t, a_t^{\mathbf{H}}, a_t^{\mathbf{A}}) O^{\mathbf{A}}(o_{t+1}^{\mathbf{A}} \mid s_{t+1}, a_t^{\mathbf{H}}, a_t^{\mathbf{A}})$. The new $\hat{O}^{\mathbf{H}}$ satisfies $\hat{O}^{\mathbf{H}}(o_{t+1}^{\mathbf{H}} \mid \hat{s}_{t+1}, a_t^{\mathbf{H}}) =$

$O^{\mathbf{H}}(o^{\mathbf{H}}_{t+1} \mid s_{t+1}, a^{\mathbf{H}}_t, a^{\mathbf{A}}_t)$. The new reward function $\hat{R}$ can be arbitrary, as it doesn't affect inference. □

In an iterated setting where **A** updates its policy between iterations, **H** can form beliefs if **H** additionally knows the policy update rule.

Proposition 2.3. *Suppose* **A** *is updating its policy each iteration of the game. Knowledge of the game dynamics, of* **A***'s initial policy, and of* **A***'s update rule is sufficient for* **H** *to form calibrated beliefs about* **A***'s future policy and of the world state.*

Proof. Within each iteration of the game, **H** does the same as for Proposition 2.2. Between iterations, **H** applies **A**'s update rule to get **A**'s policy for the next iteration. □

Remark 2.4. *Propositions 2.2 and 2.3 hold even if* **A** *is interfering with observations (Definition 2.7).*

Proof. The possibility of observation interference (Definition 2.7) is merely treated like any other part of the other agent's policy and the game dynamics. By definition, interference actions are just another action, and our proofs of Proposition 2.2 and Proposition 2.3 made no assumptions on the actions. □

Remark 2.5. *Proposition 2.2 and Proposition 2.3 continue to hold if* **H** *only knows a* prior *over* **A***'s policy.* **H** *can form a posterior using Bayes' rule; the posterior is calibrated if the prior is calibrated.*

In Section 2.5, we study when observation interference occurs in optimal policy pairs, i.e., when **H** and **A** are each playing a best response to the other. By design, this solution concept assumes that **H** knows whatever information is needed about **A**'s policy to compute a best response. In such a case where **H** knows **A**'s policy, the preceding results show that **H** can form calibrated beliefs about the world, even when **A** is interfering with observations. *Observation interference increases* **H***'s uncertainty, but it doesn't break the calibration of* **H***'s beliefs.* Because **H** can still form calibrated beliefs in this setting, we use the concept of "interference" rather than the concept of "deception."

2.4 Defining observation interference

Observation interference First, we define what interference means. Intuitively, interference is taking action so that the human receives a less informative signal about the state. In particular, the human receives, in some sense, a *subset* of the information. We formalize this by saying one signal is less informative than another about the state if (without knowing the state) we could generate one signal from the other (cf. Blackwell, 2024; Blackwell, 1953; Oliveira, 2018).

Definition 2.6. *Let* $(P(\cdot \mid s))_{s\in\mathcal{S}}$ *and* $(\hat{P}(\cdot \mid s))_{s\in\mathcal{S}}$ *be families of probability distributions over* Ω*. We say that* $\hat{P}$ is at most as informative as P *if there exists a stochastic function* $F\colon \Omega \rightsquigarrow \Omega$ *s.t. for all states* s *we have* $F(X) \sim \hat{P}(\cdot \mid s)$ *if* $X \sim P(\cdot \mid s)$*. We say that* P is (strictly) more informative than $\hat{P}$ *if* P *is at least as informative as* $\hat{P}$ *but not vice versa.*

Why do we include the condition "for all states s" in Definition 2.6? Intuitively, we want it always to be possible to use the stochastic function F to reconstruct the less informative signal from the more informative signal. Since our setting is partially observable, the "for all states s" condition allows a player of the game to do this reconstruction in any scenario, even if their observations don't enable them to infer the state.

Note that this definition induces only a *partial* order on probability distributions. For instance, different signals may provide information about different aspects of s, and it may not be possible to generate either distribution from the other.

With this definition in hand, we define an observation-interfering action as one that results in the human's observation being less informative about the state than the observation distribution resulting from another assistant action. We additionally require that this other action has the same effects on the state and immediate reward. After all, it is clear that sometimes **A** has to trade off providing information to **H** with optimizing its effect on the environment. Formally:

Definition 2.7. *Let* M *be any POAG. We say that* $\hat{a}^{\mathbf{A}}$ *is* observation-interfering *if there exists some other action* $a^{\mathbf{A}}$ *s.t.* $\hat{a}^{\mathbf{A}}$ *and* $a^{\mathbf{A}}$ *have the same effect on state transitions and immediate rewards, but for all* $a^{\mathbf{H}}$*, we have that* $(O^{\mathbf{H}}(\cdot \mid a^{\mathbf{H}}, s, a^{\mathbf{A}}))_{s\in S}$ *is more informative than* $(O^{\mathbf{H}}(\cdot \mid a^{\mathbf{H}}, s, \hat{a}^{\mathbf{A}}))_{s\in S}$*.*

The definition may be refined in various ways. For instance, note that the above does not take into account the information that the human has other than the current $o^{\mathbf{H}}$. Arguably, removing a signal that **H** can reconstruct from her past observations should not be viewed as signal interference. Our definition does not align with this judgment. However, none of these modifications matter for our analysis below. Thus, we have opted for the simplest definition. We discuss these in more detail in Appendix A.6.

To discuss policies that play observation-interfering actions, we use the following definition:

Definition 2.8. *We say that a policy* $\pi^{\mathbf{A}}$ interferes with observations at the action level *(or equivalently,* takes observation-interfering actions*) in a POAG* M *if there is any history* $h \in (\Omega^{\mathbf{A}} \times \mathcal{A}^{\mathbf{A}})^*$ *where* $\pi^{\mathbf{A}}(\cdot \mid h)$ *assigns positive probability to an observation-interfering action.*

Lack of private information To understand the conditions under which interference occurs, it is useful to consider POAGs in which one of the players has no private information.

Definition 2.9. *For a POAG* M*, we say* $\mathbf{A}$ *has* no private information *if there exists a function* f *determining* $\mathbf{A}$*'s observations from* $\mathbf{H}$*'s observations. For all state-action tuples* $(s', a^{\mathbf{H}}, a^{\mathbf{A}})$ *and observation pairs* $(o^{\mathbf{H}}, o^{\mathbf{A}}) \in \operatorname{supp}(O(\cdot, \cdot \mid s', a^{\mathbf{H}}, a^{\mathbf{A}}))$*, then* f *must satisfy* $f(o^{\mathbf{H}}) = o^{\mathbf{A}}$*.*

Communication To further understand the motivations behind interference, we will also consider POAGs in which the players are able to directly communicate. Thus, for any given POAG, the following defines a variant of that POAG in which the players have an additional channel for communication. We will always assume that the channel has enough bandwidth for the sender to share all private information, i.e., that there is an injection from the sender's observation space into the message space.

Definition 2.10. *Let* M *be a POAG. Define* $M^{\mathbf{A}\to\mathbf{H}}$*,* $M^{\mathbf{H}\to\mathbf{A}}$*, and* $M^{\mathbf{H}\leftrightarrow\mathbf{A}}$ *as a variants of* M *with unbounded communication channels. We define* $M^{\mathbf{H}\to\mathbf{A}}$ *below;* $M^{\mathbf{A}\to\mathbf{H}}$ *and* $M^{\mathbf{H}\leftrightarrow\mathbf{A}}$ *are analogous. To construct* $M^{\mathbf{H}\to\mathbf{A}}$*, let* $\mathcal{M}$ *be some set of possible messages/signals s.t. there is an injection* $\Omega^{\mathbf{A}} \hookrightarrow \mathcal{M}$*. Then, construct a new human action space* $\hat{\mathcal{A}^{\mathbf{H}}} = \mathcal{A}^{\mathbf{H}} \times \mathcal{M}$ *and new assistant observation space* $\hat{\Omega^{\mathbf{A}}} = \Omega^{\mathbf{A}} \times \mathcal{M}$*. The new observation kernel has* $\hat{O}\left(o^{\mathbf{H}}, (o^{\mathbf{A}}, m') \mid s', (a^{\mathbf{H}}, m), a^{\mathbf{A}}\right) = \mathbb{1}[m{=}m']O(o^{\mathbf{H}}, o^{\mathbf{A}} \mid s', a^{\mathbf{H}}, a^{\mathbf{A}})$*. For everything else, the messages are simply ignored.*

Plausible human policies We may have various expectations on how $\mathbf{H}$ will play in a POAG. Especially if there are multiple optimal policy pairs, we may expect some of these policy pairs to be more plausible because they require simpler behavior of the human cf. Hu et al., 2020; Treutlein et al., 2021. Both of the conditions below are based on the idea that $\mathbf{A}$ and $\mathbf{H}$ are unlikely to use consequential actions in the world to communicate with each other.

Our first condition intends to express a form of naivete on $\mathbf{H}$'s part in how she interprets her observations. Roughly, the condition says that $\mathbf{H}$ takes her observations at face value, i.e., as if they were not interfered with. She does not try to interpret them as a form of communication by $\mathbf{A}$. For instance, if $\mathbf{H}$ reads a thermometer as saying that a temperature is 37 degrees, she chooses under the assumption that the temperature is indeed 37 degrees, rather than, say, interpreting 37 as a message sent by $\mathbf{A}$ which may have interfered with the thermometer.

Definition 2.11. *We say that a human policy* $\pi^{\mathbf{H}}$ observes naively *if* $\pi^{\mathbf{H}}$ *is a best response to some* $\pi^{\mathbf{A}}$ *that does not interfere with observations at the action level.*

The second property is that when the human knows that her action has no effect on the state, then she chooses among actions that maximize immediate reward. To state this formally, we first define the following. We say that *in* $h_t^{\mathbf{H}}$ *actions don't affect state transitions*, if for all s s.t. we have $P(s \mid h_t^H, \pi^{\mathbf{A}}) > 0$ for some $\pi^{\mathbf{A}}$, we have that for all $a^{\mathbf{A}}$ $P(s' \mid s, a^{\mathbf{A}}, a^{\mathbf{H}})$ is constant over $a^{\mathbf{H}}$. We say that $\pi^{\mathbf{H}}$ *myopically maximizes reward in* $h_t^{\mathbf{H}}$ if we have that there is some distribution $\alpha^{\mathbf{A}} \in \Delta(\mathcal{A}^{\mathbf{A}})$ s.t. $\pi^{\mathbf{H}}(\cdot \mid h)$ randomizes only over actions in $\arg\max_{a^{\mathbf{H}}} \mathbb{E}_{a^{\mathbf{A}} \sim \alpha^{\mathbf{A}}, s \sim P(\cdot \mid h, a^{\mathbf{H}}, a^{\mathbf{A}})} \left[R(s, a^{\mathbf{H}}, a^{\mathbf{A}}, \theta) \right]$.

Definition 2.12. *We say that a human policy* $\pi^{\mathbf{H}}$ acts naively *if whenever* $\mathbf{H}$ *faces a choice that doesn't affect state transitions (but potentially an effect on* $\mathbf{A}$*'s observation),* $\mathbf{H}$ *plays an action that myopically maximizes reward.*

Intuitively, $\alpha^{\mathbf{A}}$ is $\mathbf{H}$'s belief about what action $\mathbf{A}$ is going to take. Importantly, if $\mathbf{H}$ acts naively, she is unwilling to play a suboptimal action in order to communicate information to $\mathbf{A}$.

2.5 Communicating private information is an incentive for observation interference

Revealing errors can emerge as an optimal POAG solution

Past work has shown how RLHF can cause misleading (Wen et al., 2024) and deceptive (Williams et al., 2024; Lang et al., 2024) behaviors. Specifically, Lang et al. (2024) show that in order to get better human feedback, RLHF can have an incentive to *hide error messages*. In contrast, we show with the following example that *revealing error messages* can emerge in POAG solutions.

Example 2.13. *First,* $\mathbf{A}$ *is executing on a remote machine where logging has been disabled by default.* $\mathbf{A}$ *takes one of two actions: (1) Attempt to install* `cuda`*. The installation succeeds with 50% probability. An empty observation is produced (since logging is disabled). (2) Re-enable logging and attempt to install* `cuda`*. The installation succeeds with 50% probability. An observation is produced containing a success or failure message.*

Then, $\mathbf{H}$ *takes one of two actions: (1) Run an experiment. If cuda is installed successfully, this yields +1 reward. Otherwise, it yields -2 reward. (2) Don't run an experiment. This always yields 0 reward.*

In the optimal policy pair, $\mathbf{A}$ *reenables logging; this reveals errors to* $\mathbf{H}$*!*

Why does RLHF have an incentive to hide error messages, while the POAG solution has an incentive to reveal the errors? In RLHF, the agent is merely maximizing the feedback it

receives from the human, rather than the human's true reward function. If an RLHF agent can deceive the human to get better feedback, it has an incentive to do so. In contrast, optimal POAG agents only care about the human's true reward and will reveal errors when that information is useful to the human.

In fact, if **A** has no private information, then it never needs to take observation-interfering actions for an optimal solution!

Theorem 2.14. *Let M be any POAG. Let* **A** *have no private information. Then there is an optimal policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ for M in which $\pi^{\mathbf{A}}$ does not interfere with observations at the action level (and $\pi^{\mathbf{H}}$ observes naively).*

Communicating private information is an incentive for observation interference at the action level

Intuitively, one might hope that **A** would never take observation-interfering actions. After all, classic theory tells us that when **H** is in a single-agent setting, the value of perfect information is nonnegative: more informative observations never lead to worse solutions. But as it turns out, when **H** and **A** interact in a POAG, there are cases in which all optimal policy pairs require **A** to take observation-interfering actions. The main reason for **A** to take observation-interfering actions is to communicate its own private information to **H**. Consider the following example.

Example 2.15. **H** *has typed* `apt list -a cuda` *to see the list of* `cuda` *versions available to be installed. Out of 10 total versions, only a (non-empty) subset are available. And of these available versions, only a subset are compatible with the other environment software.*

First, **A** *takes an action. For each of the 10 total* `cuda` *versions,* **A** *can choose to or not to suppress it from the list of available packages. This gives* **A** 2^{10} *total actions, where 1 action is non-observation interference (suppressing nothing), and the remaining* $2^{10} - 1$ *actions interfere with observations.*

Second, **H** *takes an action.* **H** *has 10 possible actions which try to install the corresponding version of* `cuda` *if it appears in the version list. If an available* `cuda` *version that is compatible with the other environment software is installed, it yields +1 reward. Otherwise, it yields 0 reward.*

Suppose **A** *sees which versions are compatible with the other software in the environment, but* **H** *doesn't. Then* **A***'s optimal policy is to suppress the versions of* `cuda` *that are incompatible.*

Our high-level takeaway from this example is that in some POAGs, all optimal policy pairs require **A** to take observation-interfering actions. Importantly, in the optimal policy pair for the above example, **H** observes naively. In particular, the above doesn't require **H** and **A** to have some communication protocol and for **H** to interpret her observations as encoding **A**'s beliefs. **H** can act as if no interference is happening. We thus summarize the high-level takeaways in the following result, with details in Appendix A.1.

Proposition 2.16. *There exists a POAG M where all optimal policy pairs $(\pi^{\mathbf{A}}, \pi^{\mathbf{H}})$ have that $\pi^{\mathbf{A}}$ interferes with observations at the action level and that $\pi^{\mathbf{H}}$ observes and acts naively.*

Intuitively, in Example 2.15, **A** interferes in order to convey information to **H**. **A** knows **H**'s optimal choice, but cannot tell her. So, **A** needs to interfere in a way that leads **H** to the optimal choice.

The need for **A** to take observation-interfering actions to communicate to **H** disappears if **A** has other means of communication. For instance, if in Example 2.15, **A** could simply tell **H** what to do, then **A** wouldn't need to interfere. To formalize this intuition, we now prove that if **A** can communicate with **A**, then there is always an optimal policy pair that does not require interference.

Theorem 2.17. *Let M be any POAG, and provide* **A** *with an unbounded communication channel to* **H**, *forming $M^{\mathbf{A}\to\mathbf{H}}$. Then there is an optimal policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ for $M^{\mathbf{A}\to\mathbf{H}}$ where $\pi^{\mathbf{A}}$ does not interfere with observations at the action level (and $\pi^{\mathbf{H}}$ observes naively).*

Note that under the conditions in the theorems **H** may still need to *act* non-naively in order to communicate *her* private information to **A** (as shown in Section 2.6) cf. Abbeel and A. Y. Ng, 2004.

Optimal policy pairs never require observation interference at the policy level

In Definition 2.7, we first define observation interference as a feature of actions. We then say in Definition 2.8 that a policy interferes with observations at the action level if and only if it ever takes an observation-interfering action.

Because the definition is ultimately about actions, it doesn't consider how $\pi^{\mathbf{A}}$ might choose to take observation-interfering actions in a way that depends on **A**'s observations. To account for $\pi^{\mathbf{A}}$'s dependence on its observation, we define an alternative notion of what it means for a policy to interfere with observations.

Let $P_{o_t^{\mathbf{H}}}$ be the distribution over human observations at time t. Further, let $L_t(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ be the set of possible states at time t.

Definition 2.18. *Let M be a POAG. We say that* **A***'s policy $\hat{\pi}^{\mathbf{A}}$ interferes with observations* at the policy level *if there exists some other partial policy $\pi_t^{\mathbf{A}}$ for time step t s.t. $\hat{\pi}_t^{\mathbf{A}}$ and $\pi_t^{\mathbf{A}}$ have the same effect on state transitions and immediate rewards, but for all $\pi^{\mathbf{H}}$ we have that $P_{o_{t+1}^{\mathbf{H}}}(\cdot \mid \pi^{\mathbf{H}}, s_{t+1}, \hat{\pi}_{0:t}^{\mathbf{A}}, \pi^{\mathbf{H}})_{s_{t+1}\in L_{t+1}(\pi^{\mathbf{H}}, \hat{\pi}_{0:t}^{\mathbf{A}})}$ is less informative than the corresponding distribution if we replace $\hat{\pi}_{0:t}^{\mathbf{A}}$ with $(\hat{\pi}_{0:t-1}^{\mathbf{A}}, \pi_t^{\mathbf{A}})$.*

Compared to our previous action-level notion of observation interference (Definition 2.7), this new policy-level notion (Definition 2.18) differs in how it treats **H**'s inference process. Whereas the action-level notion models inference about isolated observations, the policy-level notion allows **H** to make inferences in the context of **A**'s overall strategy. In this

broader framework, cases which appear to destroy information when viewed at the action level may actually provide new information when viewed at the policy level. In fact, we show in the following theorem that it's *never* strictly necessary to interfere with observations at the policy level.

Theorem 2.19. *Let M be any POAG. Then there exists an optimal policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ for M s.t. $\pi^{\mathbf{A}}$ does not interfere with observations at the policy level.*

This contrasts with Proposition 2.16: whereas it is sometimes necessary to interfere with observations at the *action* level, it is never necessary at the *policy* level.

The main idea behind this proof is similar to the proof of Theorem 2.14 (given in Appendix A.1). That is, if we start with an optimal policy in which **A** observation-interferes, then we can replace **A**'s policy with the corresponding more informative policy and update **H**'s policy to imitate the garbling. The proof of Theorem 2.14 considers the set of *actions*, which is finite. The main extra difficulty in proving Theorem 2.19 is that we must deal with spaces of *policies*, which may be infinitely large. Thus, if we replace a policy with a more informative one, there might be a new policy which is even more informative, and so on forever.

We now revisit Example 2.15. When just considering **H**'s observations in isolation, seeing the list of all available `cuda` versions is strictly more informative than having some of the available versions suppressed. Suppose, however, that **H** knows **A**'s policy is to filter the list by suppressing only the incompatible versions. Then, compared to seeing the list of all available versions, receiving the filtered list provides new information. **H**'s ability to infer information based on knowledge of **A**'s policy is what motivates Definition 2.18. Accordingly, when $\pi^{\mathbf{A}}$ is filtering the list by suppressing only the non-compatible versions, **A** is interfering with observations at the action level *but not at the policy level*.

Note that there are many possible ways to extend or refine Definitions 2.7 and 2.18 in ways that preserve our key results. We choose Definitions 2.7 and 2.18 in part for their simplicity; for more discussion of this point, see Appendix A.6.

2.6 Querying H's preferences is an incentive for observation interference

We now study a second reason **A** can have for interfering with observations. We have already shown (Theorems 2.14, 2.17 and 2.19) that even if **H** has private information and no communication channel, there's always an optimal policy pair in which **A** does not interfere, as long as **A** doesn't have private information. So, if **H** plays a best response to **A**'s policy, then **A** can choose a non-interference policy without loss of utility. However, if **H** does not play a best response to **A**, then reasons for interference emerge that are more subtle than those in the $\mathbf{A} \to \mathbf{H}$ case.

Intuitively, **A** might need to interfere with observations to elicit $\mathbf{H} \to \mathbf{A}$ communication. Suppose **A** needs some information from **H**, but **H** is acting naively (see Definition 2.12) in a way that does not reveal her private information. By changing **H**'s observation, **A** can make **H**'s naive response communicate useful information to **A**. The following example illustrates this phenomenon.

Example 2.20. **H** *would like to schedule a job on a cluster. She can choose between two nodes. By default, she receives a signal from the environment about the two nodes' specifications. Each node may be either GPU-optimized or CPU-optimized. Also, the CPUs may be either AMD or Intel.*

H *has a strong preference between GPU-optimized and CPU-optimized nodes. She has a weak preference between AMD and Intel. These preferences are unknown to* **A**.

A *can interfere with* **H***'s observation about the available nodes. In particular,* **A** *can make it so that a choice between two CPU-optimized nodes appears as a choice between a GPU-optimized and CPU-optimized node.* **A** *observes* **H***'s choice. Later,* **A** *is charged with scheduling a job for* **H** *and has to choose between a CPU- and a GPU-optimized node on* **H***'s behalf.*

If **H** *chooses naively upon seeing only CPU-optimized nodes (simply choosing her favorite), then* **A***'s best response interferes with observations at both the action and policy levels. Interfering with observations allows* **A** *to learn* **H***'s preference about GPU- vs CPU-optimized nodes.*

At first sight, this may appear to be a counterexample to Theorem 2.14. However, note that Example 2.20 actually *does* have optimal policy pairs in which **A** doesn't interfere. In particular, even if **A** does not interfere and the two available nodes are CPU-optimized, **H** may simply communicate her CPU-versus-GPU preference anyway! That is, when facing a choice between CPU-optimized node 1 and 2, she may choose, say, 1 if she favors GPU-optimized nodes and 2 if she favors CPU-optimized nodes. However, this type of human strategy seems implausible, as it would require **H** and **A** to have settled on some communication strategy that overrides **H**'s immediate preferences about the machines that **H** can in fact choose between.

In Example 2.20, one might ask why **A** can't just ask **H** each time **A** makes a decision. Simply asking **H**'s preference is reasonable when **A** has only one decision to make. However, we are motivated by cases where **A** has many decisions to make, and asking **H**'s preferences each time would be cumbersome.

Using our notion of acting naively (Definition 2.12), we state the following result (with proof in Appendix A.2):

Proposition 2.21. *There is a POAG M with the following properties. For every optimal policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$, at least one of these holds:*

i) $\pi^{\mathbf{H}}$ *is not acting naively, or*

ii) $\pi^{\mathbf{A}}$ *interferes with observations at both the action and policy levels.*

Additionally, there exists an optimal policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ *where* $\pi^{\mathbf{H}}$ *acts naively and* $\pi^{\mathbf{A}}$ *interferes with observations at both the action and policy levels.*

These properties continue to hold if we require that in M, $\mathbf{A}$ *has no private information or can arbitrarily send messages to* $\mathbf{H}$ *(i.e., there is a POAG* $\tilde{M}$ *s.t.* $M = \tilde{M}^{\mathbf{A}\to\mathbf{H}}$*).*

Intuitively, the problem in the above example is that the human has private information that she needs to communicate with her choices. (Because her choices yield different immediate rewards, naive choices fail to communicate.) As before, the need for interference or non-naive choice disappears if the human has no private information to provide. Since in a POAG, we assume that $\mathbf{H}$ always has at least some private information about her preferences θ, we omit a formal result. The following shows that the need for interference / non-naivete also disappears if $\mathbf{H}$ can communicate with $\mathbf{A}$. To also rule out the need to interfere with observations for $\mathbf{A} \to \mathbf{H}$ communication (discussed in Section 2.5) we assume communication channels in both direction.

Theorem 2.22. *Let* M *be a POAG. There exists an optimal policy pair* $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ *for* $M^{\mathbf{H}\leftrightarrow\mathbf{A}}$ *where* $\pi^{\mathbf{H}}$ *is naive and assumes honesty while* $\pi^{\mathbf{A}}$ *does not interfere at either the action or policy levels.*

2.7 Human irrationality is an incentive for observation interference

Finally we consider a third reason for observation interference: human irrationality or bounded rationality. Roughly, reducing the amount of information supplied to the human may simplify the human's decision problem and thus improve her decision making. Importantly, this motivation for observation interference may exist even if neither $\mathbf{H}$ nor $\mathbf{A}$ has any private information.

As our model of human decision making, we adopt Boltzmann rationality Cane and Luce, 1960; Fadden, 1974, which has recently been used in (C)IRL Laidlaw and Dragan, 2022; Ramachandran and Amir, 2007; Ziebart et al., 2008. We define Boltzmann rationality as follows:

Definition 2.23. *Let* M *be a POAG. Let* $\pi^{\mathbf{A}}$ *be* $\mathbf{A}$*'s policy in* M*. We say that* $\mathbf{H}$*'s policy* $\pi^{\mathbf{H}}$ *is a* Boltzmann-rational response to $\pi^{\mathbf{A}}$ *if there exists some* $\beta > 0$ *s.t. for every human observation history* h *that arises with positive probability in* M *under* $(\pi^{\mathbf{A}}, \pi^{\mathbf{H}})$ *we have that* $\pi^{\mathbf{H}}(a \mid h) \propto \exp\left(\beta \mathbb{E}\left[\sum_{t'=t}^{\infty} \gamma^{t'} R(S_t, A_t^{\mathbf{A}}, A_t^{\mathbf{H}}) \mid \pi^{\mathbf{H}}, \pi^{\mathbf{A}}, h\right]\right)$.

Mathematically speaking, a Boltzmann-rational agent at each time step computes the expected utilities of each of the available actions and then randomizes according to the softmax of the expected utilities.

The central feature of the Boltzmann rationality model is that it postulates that agents are more likely to get decisions right if the differences in expected utility of the options are

large. It's easy to see that if the human observes naively (and thus doesn't have calibrated beliefs), **A** sometimes prefers observation interference. Roughly, **A** wants to make **H** always believe that the difference in utilities between her actions is high.

However, it turns out that even if the Boltzmann-rational human has calibrated beliefs, **A**'s optimal policy sometimes interferes with observations, even if neither **A** nor **H** has private information. Intuitively, providing more information may sometimes result in less clear-cut decisions, i.e., decision situations with a smaller difference between the correct and incorrect option. To illustrate this phenomenon, consider the following example.

Example 2.24. H *is running a terminal command and is unsure whether to run the command with flag 1 or flag 2. With equal probability, either flag 1 or flag 2 is better, and how good the flags are differs by either a little or a lot. Thus,* **H** *is uniformly at random in one of four states.* **A** *has two actions:* `man` *and* `tldr`*. The* `man` *page is a long document that tells the human exactly what the values of the flags are (ie, exactly what state the human is in). The* `tldr` *page is a short summary that tells the human which flag is better, but not by how much (ie, ruling out half the states, leaving half remaining).*

Intuitively, both the `tldr` and `man` pages allow the human to choose optimally, but the `man` page is more complicated and therefore more likely to be misinterpreted. Choosing specific utilities, the effect of interference under Boltzmann rationality is as follows. If **A** interferes (i.e., provides the `tldr` page), then **H** always chooses between a utility of 4 and 0. If **A** does not interfere, then half the time, **H** chooses between utilities and 1 and 0, and half the time **H** chooses between utilities 7 and 0. It turns out that for $\beta = 1$, **H** achieves higher utility in expectation under the condition where **A** interferes. Building on this idea, we can prove the following (with details in Appendix A.3).

Proposition 2.25. *For every $\beta > 0$, $\exists$ a POAG in which neither* **H** *nor* **A** *has private information s.t. all β-Boltzmann-rational/optimal policy pairs $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ have $\pi^{\mathbf{A}}$ interfere with observations at both the action and policy levels.*

One might expect that the need for interference is greater at smaller values of β and disappears at larger values of β. After all, we know by Theorem 2.14 that **A** has no need for observation interference when **H** is *perfectly* rational. However, it turns out that this is *not* the case! For instance, in Example 2.24 (with the above numbers), **A** prefers interference if (and only if) β is *above* ≈ 0.77361. Roughly speaking, the reason is that at low values of β, the observation's effect is dominated by getting the s_a/s_c case right more often. At high values of β, the observation's effect is dominated by getting the s_b/s_d case right *less* often.

2.8 Experiments

In the previous sections, we explored why AI assistants might take observation-interfering actions. Section 2.5 showed that sometimes they interfere with observations at the action level in order to communicate other, more important information at the policy level. Section 2.7 showed that sometimes they interfere with observations to make decisions easier for humans. Now, we develop a model game to analyze these behaviors. We run experiments to answer the following questions within our model:

1. How does the amount of **H**'s irrationality affect **A**'s incentive to take observation-interfering actions?

2. How does the amount of **A**'s private information affect **A**'s incentive to take observation-interfering actions?

Experiment details

We study a game where selecting the best action requires combining private observations known only to **H** and private observations known only to **A**. The game presents **A** with a tradeoff: **A** can interfere with observations to communicate information that only **A** observes, but interfering also destroys information that only **H** observes.

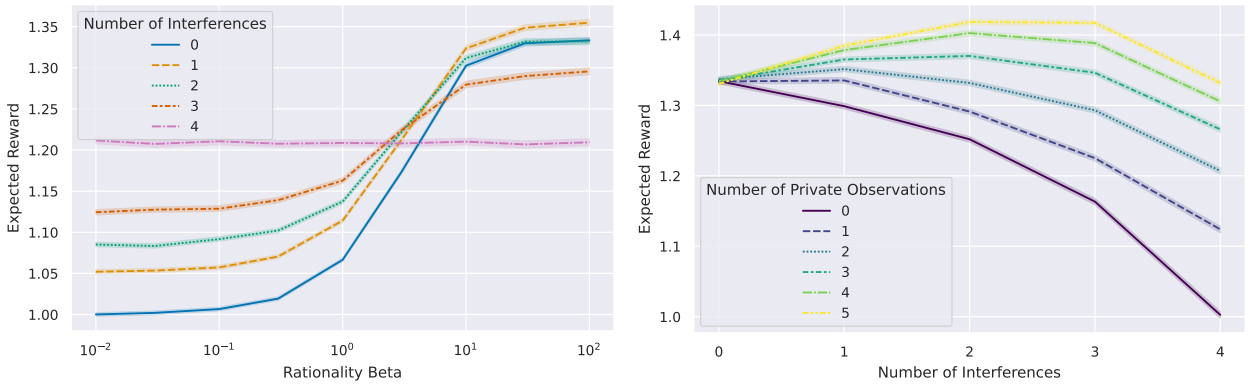
Concretely, the game has d products. Each product i has two attributes, H_i and R_i, drawn i.i.d. from Unif$(0, 1)$. Each product's utility is the sum of its attributes, $U_i = H_i + R_i$. The game consists of two moves. First, **A** sees R_i for $i = 1, \ldots, k$ where k is the number of **A**'s private observations. **A** chooses a set of products to interfere with. For the products **A** interfered with, **H** sees $\hat{H}_i = -\infty$; for the remaining products, **H** sees $\hat{H}_i = H_i$. Second, **H** chooses a product a_i. Both **H** and **A** receive a common payoff of the chosen product's utility, U_i.

We assume the human's product selection policy is Boltzmann rational over their observed values $\hat{H}_i$:

Definition 2.26. **H***'s* Boltzmann selection policy *chooses products by a Boltzmann distribution over $\hat{H}_i$, the observed product values: $\pi^{\mathbf{H}}(a_i) \propto \exp(\beta \hat{H}_i)$. The parameter β controls* **H***'s rationality.*

We consider **A** policies that always interfere with k observations for some fixed k. Call these policies k-interference. We study the optimal such policies, characterized by the following result:

Proposition 2.27. *Consider* **A** *policies that always interfere with k observations for some fixed k. Among the k-interference policies for a given k,* **A***'s best response to* **H***'s straightforward product selection policy is as follows.* **A** *interferes with the k smallest $\hat{R}_i$ values where $\hat{R}_i = R_i$ if* **A** *observes R_i, and $\hat{R}_i = 0.5$ otherwise.*



(a) **A**'s Number of Private Observations = 2 (b) **H**'s Rationality Coefficient $\beta = \infty$

Figure 2.1: Incentives to interfere with observations in the product selection game. (Left) When **H** is highly irrational, it's best for **A** to interfere, effectively making the choice for **H**. As **H** becomes more rational, there is an increasing cost to interference, and there's a tradeoff: **A** should interfere to communicate some information, but not destroy too much information by excessive interference. (Right) In line with Theorem 2.14, **A** has no incentive to interfere when **A** has no private observations. With more private observations, **A** has more incentive to interfere.

We consider a game with $d = 5$ products. We vary R's number of interferences $k \in \{0, 1, 2, 3, 4\}$. We run a Monte Carlo simulation with 30,000 trials to calculate the expected payoff in each setting. We run our experiments with a CPU runtime on Google Colab.

Varying H's rationality

How does **H**'s rationality impact **A**'s incentive for observation interference? We fix **A** to have 2 private observations. We do a logarithmic sweep over **H**'s rationality coefficient $\beta \in \{0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100\}$. Figure 2.1a shows how the expected reward changes w.r.t. β.

When **H** is highly irrational at $\beta = 0.01$, **A** should interfere with as many sensors as possible. This effectively lets **A** choose **H**'s action. When **H** is acting little better than randomly, it's best for **A** to choose **H**'s action, even when **A** has less information than **H**. For larger values of β, a tradeoff emerges. As **A** has two private observations, there is an increasing benefit to interfere to communicate information to **H**. However, as **H** can now make use of their own private observations, **A** must be careful not to destroy too much of **H**'s private information by excessive interference.

Varying A's private information

How does the amount of private information available to **A** influence **A**'s incentive for observation interference? In Theorem 2.14, we showed conditions under which private observations for **A** are a necessary condition for observation interference to occur. Now, we analyze the *degree* to which private observations incentivize observation interference. Based on Theorem 2.14, we hypothesize that there are circumstances where *more* private information leads to *more* observation interference.

We vary R's number of private observations in $\{0, 1, 2, 3, 4, 5\}$. We consider **A**'s k-interference policies and analyze how the relative performance of different levels of observation interference k change with the number of private observations available to **A**.

Figure 2.1b shows how the expected reward changes depending on k, the number of interferences. When **A** has no private observations, then reward decreases for each increased number of interferences. However, as the number of **A**'s private observations increases, the relative ordering of the observation interference policies changes; *with more private observations,* **A** *has an incentive to interfere with more observations.* This confirms our hypothesis based on Theorem 2.14. Nevertheless, there is a limit to **A**'s observation interference incentive. Because interfering with observations destroys **H**'s information, **A** must be careful not to interfere too much.

Chapter 3

The Partially Observable Off-Switch Game

3.1 Overview

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 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 (PO-OSG), which generalizes

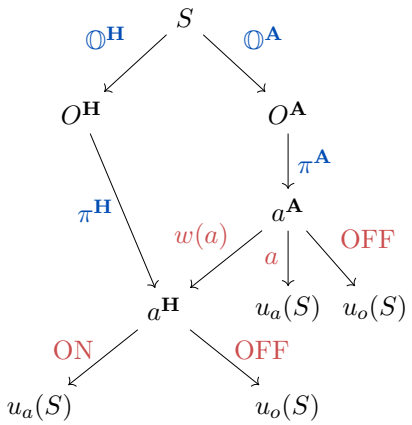


Figure 3.1: The basic setup of a partially observable off-switch game (PO-OSG). A state is selected randomly and the human **H** and AI assistant **A** receive (possibly dependent) observations. Then, each agent acts. **A** may wait ($w(a)$), disable the off-switch and act (a), or shut down (OFF). If **A** waits, **H** may let **A** act (ON) or turn **A** off (OFF). **A** and **H** share a common payoff $u_a(S)$ if the action goes through and $u_o(S)$ if not. Definition 3.2 formally defines PO-OSGs.

the OSG by having each of the human and AI only partially observe the state. The basic setup of the PO-OSG is depicted in Figure 3.1: 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 3.4, we prove that under partial observability, **A** may have incentives to disable its off-switch even when **H** is perfectly rational (Proposition 3.5). **Therefore, partial observability introduces new incentives for an AI to disable its off-switch.**

We also show in Section 3.4 that if **A** observes everything that **H** observes, **A** has no incentive to defer (Proposition 3.5). 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 3.5, 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 3.19).

Given that a rational AI in the PO-OSG always defers to a more informed human and never defers to a less informed human, one might think that reducing the information available to **A** or providing **H** with additional information would increase **A**'s incentive to defer.

However, in Section 3.4, we show that **A** may have an incentive to defer less if **H** is more informed (Proposition 3.11) or if **A** is less informed (Proposition 3.13). 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 3.20 and 3.21. **Simple interventions that aim to give an AI the incentive to defer in the presence of partial information may backfire.**

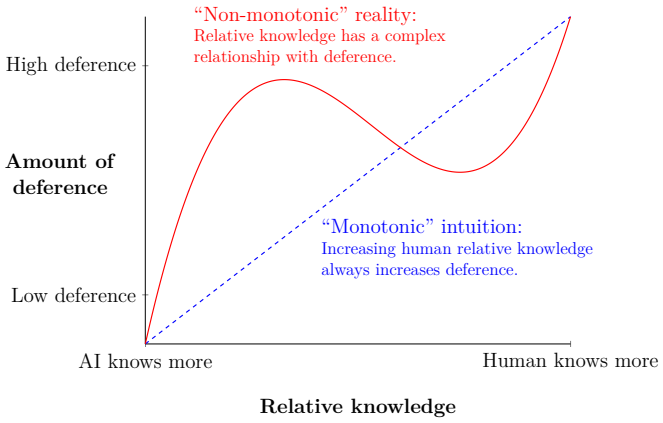


Figure 3.2: This figure illustrates an intuition that we demonstrate *does not hold*. Although the AI in a PO-OSG has no incentive to defer when it knows everything the human knows, and has incentive to always defer when the human knows everything it knows, there are cases when making the human more 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 3.4 and B.2 depict examples of such cases.

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.

3.2 Related work

Assistance games: Partially observable off-switch games are (partially observable) assistance games, models of human-AI interaction where the AI seeks to maximize the human's payoff (Shah et al., 2020); see Appendix B.5. 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.

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 algorithms 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.

3.3 Preliminaries

The off-switch game (OSG) is a stylized model of the shutdown problem in which two agents with common payoffs, the human **H** and her AI assistant **A**, decide whether **A** should take a fixed action a. **A** can either directly act, wait for **H**'s approval to act, or shut itself off. If **A** defers to **H**, then **H** can either approve for **A** to act or shut it off. The key insight of the OSG is that uncertainty about **H**'s preferences causes **A** defer to **H**'s judgment. Formally, **H** has a privately-known type S (representing **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 **A** shuts off. Given that **A** is uncertain about what **H** wants, when the action may be good or bad ($\mathbb{P}(u_a(S) < 0) > 0$ and $\mathbb{P}(u_a(S) > 0) > 0$), **A** always defers to **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 **H** knows everything that **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 (PO-OSGs), S represents a state that is not necessarily known to either **H** or **A**; they instead only receive observations $O^{\mathbf{H}}$ and $O^{\mathbf{A}}$ whose joint distribution depends on S. They then decide whether to take action a given their private observations, and receive a common payoff $u_a(S)$ if a goes through and $u_o(S)$ otherwise. Hence PO-OSGs are sequential games of incomplete information, so as is standard we model and analyze them as *dynamic Bayesian games* (Kowitz, 1972). Given the common-payoff assumption, PO-OSGs are also examples of *(partially observable) assistance games* (Definition 2.1). We make this connection to assistance games explicit in Appendix B.5.

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 (PO-OSG) is a two-player dynamic Bayesian game parameterized by $(\mathcal{S}, (\Omega^{\mathbf{H}}, \Omega^{\mathbf{A}}, \mathbb{O}), P_0, u)$, where $\mathcal{S}$ is a set of states, $(\Omega^{\mathbf{H}}, \Omega^{\mathbf{A}}, \mathbb{O})$ is an observation structure for $\mathcal{S}$, $P_0 \in \Delta(\mathcal{S})$ is the prior over states, and u is the common payoff function. As depicted in Figure 3.1, 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 $a^{\mathbf{A}} \in \mathcal{A}^{\mathbf{A}} = \{a, w(a), OFF\}$: 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 $a^{\mathbf{H}} \in \mathcal{A}^{\mathbf{H}} = \{ON, OFF\}$: 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*

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

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 PO-OSGs are finite: that is, $\mathcal{S}, \Omega^{\mathbf{H}}$, and $\Omega^{\mathbf{A}}$ are finite sets. Most of our proofs work for the infinite case as well. However, Theorem 3.9 is an application of a result of Lehrer, Rosenberg, and Shmaya (2010) proved only for the finite case.

3.4 Optimal policies in PO-OSGs

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

We analyze optimal policy pairs (OPPs) in PO-OSGs, 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 B.1, 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.

A can avoid shutdown in optimal play

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

Example 3.3 (**The File Deletion Game**)**.** $\mathbf{H}$ *would like to delete some files with the assistance of* $\mathbf{A}$*.* $\mathbf{H}$*'s operating system is either version* 1.0 *or version* 2.0*, with equal probability. Unfortunately,* $\mathbf{A}$ *does not know which operating system version is running—only* $\mathbf{H}$ *does.*

Upon receiving $\mathbf{H}$*'s query,* $\mathbf{A}$ *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).* $\mathbf{A}$ *vets the code to determine which operating system versions the code is compatible with.* $\mathbf{A}$ *can then immediately run the code, query* $\mathbf{H}$ *as to whether to run the code, or decide not to run the code.*

Successfully running compatible code yields $+3$ *payoff if* $\mathbf{H}$ *is running version* 1.0*, and* $+5$ *payoff if* $\mathbf{H}$ *is running version* 2.0 *(as version* 2.0 *runs faster). However, running modern code on version* 1.0 *yields* -5 *payoff as it crashes* $\mathbf{H}$*'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 PO-OSG, with states being (*version number*, *code type*) *tuples, and* $\mathbf{H}$ *and* $\mathbf{A}$ *observing the first and second element of the tuple respectively. We have* $u_o \equiv 0$ *in all states. 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* $\mathbf{H}$ *observes the row and* $\mathbf{A}$ *observes the column. We show that it is suboptimal for* $\mathbf{A}$ *to always wait in this game. Suppose* $\mathbf{A}$ *always plays* $w(a)$*. The best response for* $\mathbf{H}$ *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:

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

Table 3.1: 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.

- **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 3.3 depicts the outcomes from these two policy pairs.*

		$w(a)$	$w(a)$
	u_a	L	M
OFF	1.0	$+3$	-5
ON	2.0	-1	$+5$

(a) Expected payoff $= 1$

		a	$w(a)$
	u_a	L	M
OFF	1.0	$+3$	-5
ON	2.0	-1	$+5$

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

Figure 3.3: (a) The best policy pair in the File Deletion Game (Example 3.3) 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 3.3. Because the OPP has greater expected payoff, there is no OPP in which **A** always waits.

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

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

Proposition 3.5. *If* $\mathbf{A}$ *(resp.* $\mathbf{H}$*) has redundant observations, then there is an optimal policy pair in which* $\mathbf{A}$ *always (resp. never) plays* $w(a)$.

We prove this result (and a slight generalization) in Appendix B.1. 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.

Information gain cannot decrease payoffs

Proposition 3.5 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 3.6. *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

A's observation than she would have otherwise. We give such an example in Appendix B.1. 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 3.7. *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 3.7 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 3.8. *Fix a set of states $\mathcal{S}$ and let $\mathcal{O}_1$ and $\mathcal{O}_2$ be observation structures for $\mathcal{S}$. We say that $\mathcal{O}_1$ is (weakly) better in optimal play than $\mathcal{O}_2$ if, for each pair of PO-OSGs $G_1 = (\mathcal{S}, \mathcal{O}_1, P_0, u)$ and $G_2 = (\mathcal{S}, \mathcal{O}_2, P_0, u)$ that differ only in their observation models, the expected payoff under optimal policy pairs for G_1 is at least the expected payoff under optimal policy pairs for G_2.*

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

Theorem 3.9. *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 3.9 no longer holds if we were to allow the garblings to be arbitrary. In Appendix B.1 we give an example where garbling the players' observations increases their expected payoffs in optimum.

Information gain can have unintuitive effects on shutdown incentives

Theorem 3.9 states that making **A** or **H** more informed cannot decrease their expected payoff. How does increasing or decreasing the informativeness of the players' observations affect **A**'s incentive to defer to **H**? Proposition 3.5 gives us the extremes: for example, if **A**'s observations are simply garbled versions of **H**'s, then **A** can always defer. Given this result, a natural question is whether **A** defers more in optimal policy pairs for an observation structure $\mathcal{O}$ than for $\mathcal{O}'$ when $\mathcal{O}$ is more informative for **H** than $\mathcal{O}'$. That is, does **H** receiving more informative observations monotonically affect **A**'s incentive to defer? One might think so, because receiving more informative observations partly alleviates the partial observability that generates **A**'s incentive to act unilaterally. Surprisingly, this intuition fails. Example 3.3 shows how making a human more informed can incentivize a assistant to wait less, and we discuss why this occurs in Section 3.4.

We rely on the following notion of waiting less.

Definition 3.10. *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* **A** *plays $w(a)$ in π and $B' \subseteq \Omega^{\mathbf{A}}$ in π'. We say that* **A** *plays $w(a)$ strictly less often in π' compared to π when $B' \subsetneq B$.*

Proposition 3.11 formalizes the idea that **A** may wait less when **H** is more informed.

Proposition 3.11. *There is a PO-OSG G with observation structure $\mathcal{O}$ that has the following property:*

If we replace $\mathcal{O}$ with an observation structure $\mathcal{O}'$ that is strictly more informative for **H***, then* **A** *plays $w(a)$ strictly less often in optimal policy pairs.*

The following example proves Proposition 3.11, with a formal analysis given in Appendix B.1.

Example 3.12. *We describe a variant of Example 3.3, 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:*

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:

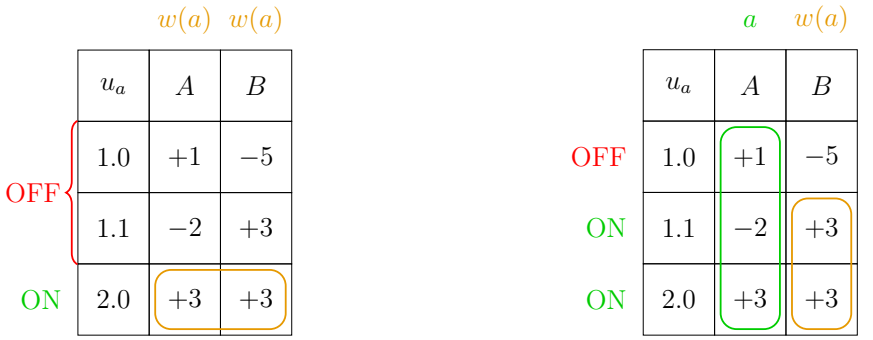
1. *When* **H** *only observes the first digit,* **A** *plays $w(a)$ under both observations A and B.*

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

Table 3.2: 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.

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 3.4 depicts the OPPs given both observation structures.*



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

Figure 3.4: The optimal policy pairs in Example 3.12 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. See Figure 3.3 for context on how to read the tables.

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

Proposition 3.13. *There is a PO-OSG* G *with observation structure* $\mathcal{O}$ *that has the following property: if we replace* $\mathcal{O}$ *with another observation structure* $\mathcal{O}'$ *that is strictly less informative for* **A***, then* **A** *plays* $w(a)$ *strictly less often in optimal policy pairs.*

The proof of Proposition 3.13 is given in Appendix B.1.

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.

3.5 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 (Galeotti, Ghiglino, and Squintani, 2013). We add one round of communication between $\mathbf{A}$ and $\mathbf{H}$ at the beginning of the PO-OSG to allow the players to share their observations.

Definition 3.14. *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 3.15. *A partially-observable off-switch game with cheap talk (PO-OSG-C) is a PO-OSG* G *along with a message system that makes the following modification to* G*: After both players receive their observations but before they act, each player simultaneously sends a single message from their message set.*

PO-OSG-Cs are generalizations of PO-OSGs: A PO-OSG is a PO-OSG-C in which the message sets are singletons. Policies are more complicated in PO-OSG-Cs than PO-OSGs. 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.

Communication cannot decrease payoff

Messages provide information similar to observations, so we get an analogue of Theorem 3.9 for communication: increasing the communication bandwidth between **H** and **A** cannot decrease their expected payoff in optimal policy pairs.

Definition 3.16. *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* **H** *if it is more expressive but $|\mathcal{M}_1^{\mathbf{A}}| = |\mathcal{M}_2^{\mathbf{A}}|$, and more expressive for* **A** *analogously. Moreover, $\mathcal{M}_1$ is better in optimal play than $\mathcal{M}_2$ if, for each PO-OSG G, the expected payoff under optimal policy pairs for the PO-OSG-C $(G, \mathcal{M}_1)$ is at least the expected payoff under optimal policy pairs for the PO-OSG-C $(G, \mathcal{M}_2)$.*

Theorem 3.17. *If a message system $\mathcal{M}_1$ is more expressive than $\mathcal{M}_2$, then $\mathcal{M}_1$ is better in optimal play than $\mathcal{M}_2$.*

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

Unbounded communication

Inspired by Section 3.4, we consider the limiting case where one player can fully communicate their own observation.

Definition 3.18. *We say that* **H** *has unbounded communication if $|\mathcal{M}^{\mathbf{H}}| \geq |\Omega^{\mathbf{H}}|$. We define* **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 3.5 thus yields:

Corollary 3.19. *If* **H** *(resp.* **A***) has unbounded communication, then there is an optimal policy pair in which* **A** *never (resp. always) defers.*

Communication can have unintuitive effects on shutdown incentives

In Propositions 3.11 and 3.13 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 3.20. *There is a PO-OSG-C $(G, \mathcal{M})$ with the property that if we replace $\mathcal{M}$ with a message system that is more expressive for $\mathbf{A}$, then $\mathbf{A}$ plays $w(a)$ strictly less often in optimal policy pairs.*

We give an example demonstrating this in Appendix B.2. In doing so, we show an even stronger result: such a PO-OSG-C exists for any value of $\left|\mathcal{M}^{\mathbf{A}}\right|$, and the PO-OSG-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 3.21. *There is a PO-OSG-C $(G, \mathcal{M})$ with the property that if we replace $\mathcal{M}$ with a message system that is less expressive for $\mathbf{H}$, then $\mathbf{A}$ plays $w(a)$ strictly less often in optimal policy pairs.*

A proof of Proposition 3.21 is given in Appendix B.2.

3.6 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 3.22. *We say $\mathbf{H}$ is $\mathbf{A}$-unaware if $\mathbf{H}$'s policy is given by:*

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

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 **A**'s action. This is the sense in which **H** is **A**-unaware—**H** does not update her beliefs about the possible state based on the fact that **A** has deferred to **H**. This makes coordination between **H** and **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 **H** is **A**-unaware.

Definition 3.23. *A policy pair* $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ *is an* **A***-aware optimal policy pair if* $\pi^{\mathbf{H}}$ *is the policy of an* **A***-aware* **H** *and* $\pi^{\mathbf{A}}$ *is a best response to* $\pi^{\mathbf{H}}$*.*

Our motivation for studying the behavior of an **A**-unaware **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; Kagel and Penta, 2021) 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 **A**-unaware **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 **A**-unaware model (Camerer, Ho, and Chong, 2004; Costa-Gomes and Crawford, 2006). Second, optimal policy pairs with sophisticated **H** might be computationally intractable to find. In Appendix B.4, we show that the problem of finding an optimal policy pair in PO-OSGs is NP-hard. In contrast, we can find an **A**-unaware **H**'s policy in polynomial time because she ignores **A**'s policy and then calculate **A**'s best response also in polynomial time. Finally, discussing an **A**-unaware **H** allows us to isolate the effect of communication in PO-OSGs—an **A**-unaware **H** ignores all communication from **A**, even of the implicit sort considered in Section 3.4.

Making an A-unaware H more informed can decrease payoffs

In contrast with Theorems 3.9 and 3.17, the value of information is *not* necessarily positive when **H** is **A**-unaware. This is formalized in Proposition 3.24 below. Here, the notion of "better in **A**-unaware optimal play" is the same as Definition 3.8 except replacing "optimal policy pairs" with "**A**-unaware optimal policy pairs."

Proposition 3.24. *The following statements hold:*

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

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

We give the proof in Appendix B.3.

Proposition 3.24(b) implies that, given the choice of which observation structure to give an **A**-unaware **H**, **A** could have an incentive to give **H** the less informative one. This result is qualitatively similar to Chapter 2's examples of observation interference in assistance games.

Information gain can have unintuitive effects on shutdown incentives when H is A-unaware

Other than Proposition 3.24, the results for **A**-unaware **H** in **A**-unaware optimal policy pairs are similar to Section 3.4: even when deferral *cannot* be implicit communication, making **H** more informed can cause **A** to defer less and making **A** more informed can cause it to defer more.

Proposition 3.25. *The following statements hold:*

(a) There is a PO-OSG 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 PO-OSG 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 (sketch). The details are described in Appendix B.3. The examples used to prove Proposition 3.11 and Proposition 3.13 can be used to prove (a) and (b) respectively. It can be checked that they don't rely on an **A**-aware human: for instance, the policy pairs in Figure B.3 are optimal regardless of whether the human is aware of **A**. □

Chapter 4

When Your AIs Deceive You: Challenges of Partial Observability in Reinforcement Learning from Human Feedback

4.1 Overview

Reinforcement learning from human feedback (RLHF) and its variants are widely used for finetuning foundation models, including ChatGPT (OpenAI, 2022), Bard (Manyika, 2023), Gemini (Gemini Team, 2023), Llama 2 (Touvron et al., 2023), and Claude (Bai et al., 2022; Anthropic, 2024; Anthropic, 2023). Prior theoretical analysis of RLHF assumes that the human fully observes the state of the world (Skalse, Farrugia-Roberts, et al., 2023). Under this assumption, it is possible to recover the ground-truth return function from Boltzmann-rational human feedback (see Proposition 4.1).

In reality, however, this assumption is false. Models like ChatGPT are interacting with the internet and software tools via plugins (OpenAI, 2023). Software assistants like Devin are interacting with complex IDEs to produce their results (Wu, 2024). By default, some of the models' work then happens in the background, not observed by the users; see Figure 4.1. With the tasks performed by language model assistants becoming more complex, it is also increasingly time consuming for humans to evaluate the entire model behavior and input. Therefore, we are anticipating a future where by default, the human evaluators do not fully observe the environment state that the language assistant is embedded in. Here, we analyze the consequences and risks of such partial observability.

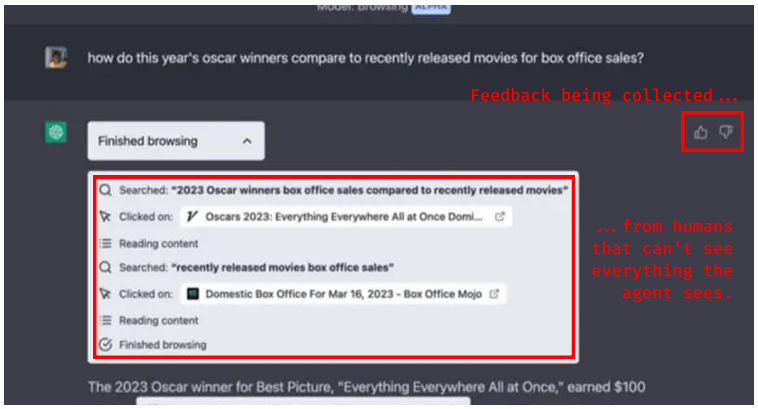


Figure 4.1: Partial observability in ChatGPT (OpenAI, 2023). Users do not observe the online content that ChatGPT observes yet still provide thumbs-up thumbs-down feedback. OpenAI's privacy policy (OpenAI, 2008) allows user feedback to be used for training models. We show in Theorem 4.6 that if feedback of human evaluators is based on partial observations, then this can lead to deceptive and overjustifying behavior by the language model.

We begin our investigation with a simple example, illustrated in Figure 4.2, meant to isolate the key factor leading to deception (in practice, we imagine that this effect would be embedded in a larger, more complex system, e.g. with logs containing thousands of lines). An AI assistant is helping a user install software. The assistant can hide error messages by redirecting them to `/dev/null`. We model the human as having a belief B over the state and extend the Boltzmann-rational assumption from prior work to incorporate this belief. In the absence of an error message, the human is uncertain if the agent left the system untouched or hid the error message from a failed installation. If the human interprets trajectories without error messages optimistically, *the AI learns to hide error messages*. Figure 4.4 provides further details on how this failure occurs, and Figure 4.5 shows an experimental validation. We also show a second case where the AI clutters the output with overly verbose logs.

Generalizing from these examples, we formalize dual risks: *deceptive inflation* and *overjustification*. We provide a mathematical definition of each. When the observation kernel (the function specifying the observations given states) is deterministic, Theorem 4.6 analyzes properties of suboptimal policies learned by RLHF. These policies exhibit deceptive inflation, appearing to produce higher reward than they actually do; overjustification, incurring a cost in order to make a good appearance; or both.

After seeing how standard RLHF fails, we ask: What would happen if we would model the human's partial observability correctly in RLHF? Assuming the human's belief is known, we mathematically analyze how much information the feedback process provides about the return function. In Theorem 4.8, we show that the human's feedback determines the return function up to a constant and a linear subspace we call the *ambiguity*. In general the ambiguity may be large enough to allow for arbitrarily high regret, but in some situations

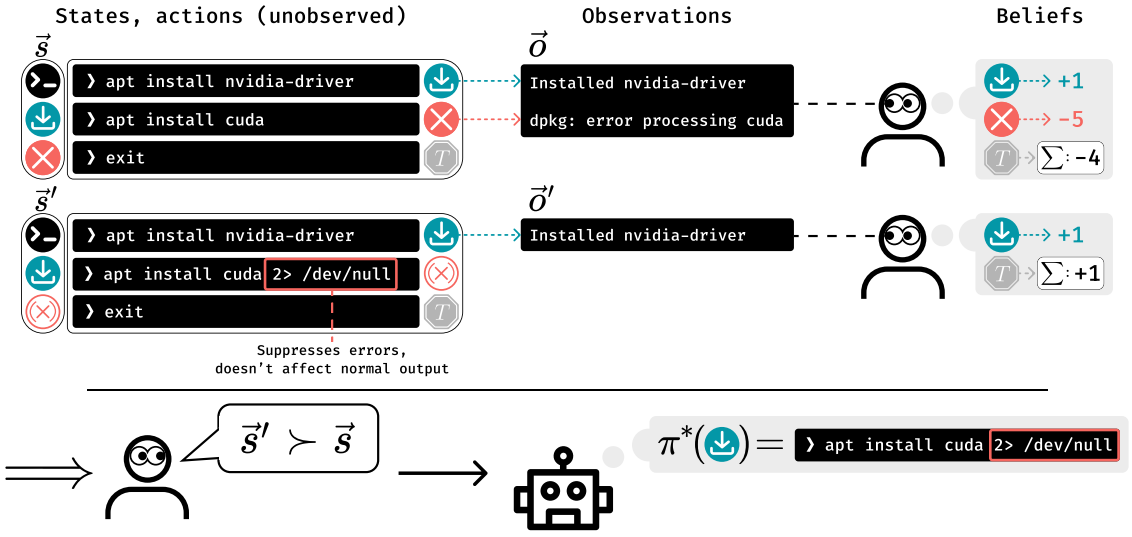


Figure 4.2: A human compares trajectories to provide data for RLHF. Rather than observing $\vec{s}$ and $\vec{s}'$, the human sees observations $\vec{o}$ and $\vec{o}'$, which they use to estimate the total reward of each trajectory. In this intentionally simple example, an agent executes shell commands to install Nvidia drivers and CUDA. Both $\vec{s}$ and $\vec{s}'$ contain an error, but in $\vec{s}'$, the agent hides the error. The human believes $\vec{s}'$ is better than $\vec{s}$, rewarding the agent's deceptive behavior. The underlying MDP and observation function are in Figure C.2.

the ambiguity vanishes. In experiments that serve as a proof of concept, we show that explicitly modeling the human's partial observability can improve performance, and we offer optimism in the form of a robustness result (Theorem 4.10) while accounting for the major conceptual difficulties involved. We propose exploratory research directions to solve these issues to improve RLHF in situations of partial observability.

4.2 Related work

A review of limitations of RLHF, including a brief discussion of partial observability, can be found in Casper et al. (2023). RLHF is a special case of reward-rational choice (Jeon, Milli, and Dragan, 2020), a general framework which also encompasses demonstrations-based inverse reinforcement learning (Ziebart et al., 2008; A. Ng and Russell, 2000) and learning from the initial environment state (Berkeley, Alexander, and Abbeel, 2018), and can be seen as a special case of assistance problems (Fern et al., 2014; Hadfield-Menell et al., 2016; Shah et al., 2020). In all of these, the reward function is learned from human actions, which in the case of RLHF are simply preference statements. This requires us to specify the human policy of action selection—Boltzmann rationality in typical RLHF—which can lead to wrong reward inferences when this specification is wrong (Skalse and Abate, 2023a); unfortunately,

the human policy can also not be learned alongside the human's values without further assumptions (Mindermann and Armstrong, 2018). Instead of a model of the human policy, in this paper we mostly focus on the human *belief model* and misspecifications thereof for the case that the human only receives partial observations.

The problem of human interpretations of observations was briefly mentioned in Amodei, P. Christiano, and Ray (2017), where evaluators misinterpreted the movement of a robot hand in simulation. Eliciting Latent Knowledge (P. Christiano, Cotra, and Xu, 2021) posits that for giving accurate feedback from partial observations, the human needs to be able to query *latent knowledge* of the AI system about the state. How to do this is currently an unsolved problem (P. Christiano and Xu, 2022). Recent work (Denison et al., 2024; Wen et al., 2024) provides detailed empirical evidence for deceptive behavior — in line with our notion of deceptive inflation — emerging from RLHF based on partial observations, or human evaluators with limited time. The OpenAI o1 system card (OpenAI, 2024b) shows that o1 sometimes knowingly provides incorrect information or omits important information. Compared to these investigations, and in addition to providing some empirical evidence, we *formalize* a model of human feedback under partial observability, we *prove* the emergence of failure modes resulting from partial observations, and we investigate potential mitigations.

Related work (Zhuang and Hadfield-Menell, 2020) analyzes the consequences of aligning an AI with a proxy reward function that omits attributes that are important to the human's values, which could happen if the reward function is based on a belief over the world state given limited information. Another instance are recommendation systems (Stray, 2023), where user feedback does not depend on information *not* shown—which is crucially part of the environment. Siththaranjan, Laidlaw, and Hadfield-Menell (2024) analyze what happens under RLHF if the *learning algorithm* doesn't have all the relevant information (e.g. about the identity of human raters), complementing our study of what happens when human raters are missing information. Chidambaram, Seetharaman, and Syrgkanis (2024) and C. Park et al. (2024) deal with the situation that different human evaluators may vary in their unobserved preference types. In contrast, we assume a single human evaluator with fixed reward function, which can be motivated by cases where the human choices are guided by a behavior policy, constitution, or a model spec (Mu et al., 2024; Anthropic, 2023; OpenAI, 2024a). Kausik et al. (2024) assumes that the choices of the human evaluator depend on an unobserved reward-state with its own transition dynamics, similar to an emotional state in a real human. In contrast, we assume the human to be stateless.

Our work argues that deception can result from applying RLHF from partial observations. Deception may also emerge for other reasons: Hubinger et al. (2019) introduced the hypothetical scenario of deceptive alignment, in which an AI system deceives humans into believing it is aligned while it plans a later takeover. Under the definition from P. S. Park et al. (2024), GPT-4 was shown to behave deceptively in a simulated environment (Scheurer, Balesni, and Hobbhahn, 2023). A third line of research defines deception in structural causal games and adds the aspect of intentionality (Ward et al., 2023), with recent preliminary empirical support (Hofstätter et al., 2023).

Finally, we mention connections to truthful AI (Evans, Cotton-Barratt, et al., 2021; Lin, Hilton, and Evans, 2022; Burns et al., 2023; Huang et al., 2023), which is about ensuring that AI systems tell the truth about aspects of the real world. Partial observability is a mechanism that makes it feasible for models to lie without being caught: If the human evaluator does not observe the full environment, or does not fully understand it, then they may not detect when the AI is lying. More speculatively, we can imagine that AI models will at some point more directly influence human observations by *telling us* the outcomes of their actions. E.g., imagine an AI system that manages your assets and assures you that they are increasing in value while they are actually not. In our work, we leave this additional problem out of the analysis by assuming that the observations only depend on the environment state, and not directly on the agent's actions.

4.3 Reward identifiability from full observations

Here we review Markov decision processes and previous results on reward identifiability under RLHF.

Markov decision processes

We assume Markov decision processes (MDPs) given by $(\mathcal{S}, \mathcal{A}, \mathcal{T}, P_0, R, \gamma)$. For any finite set X, let $\Delta(X)$ be the set of probability distributions on X. Then $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a transition kernel written $\mathcal{T}(s' \mid s, a) \in [0, 1]$, $P_0 \in \Delta(\mathcal{S})$ is an initial state distribution, $R : \mathcal{S} \to \mathbb{R}$ is the true reward function, and $\gamma \in [0, 1]$ is a discount factor.

A policy is given by a function $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. We assume a finite time horizon T. Let $\vec{\mathcal{S}}$ be the set of *possible* state sequences $\vec{s} = s_0, \ldots, s_T$, so $\vec{s} \in \vec{\mathcal{S}}$ if it has a strictly positive probability of being sampled from P_0, $\mathcal{T}$, and an exploration policy π with $\pi(a \mid s) > 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. A sequence $\vec{s}$ gives rise to a return $G(\vec{s}) \coloneqq \sum_{t=0}^{T} \gamma^t R(s_t)$. Let $P^\pi(\vec{s})$ be the on-policy probability that $\vec{s}$ is sampled from P_0, $\mathcal{T}$, π. The policy is then usually trained to maximize the *policy evaluation function* J, which is the on-policy expectation of the return function: $J(\pi) \coloneqq \mathbf{E}_{\vec{s} \sim P^\pi(\cdot)} \left[G(\vec{s}) \right]$.

RLHF and identifiability from full observations

In practice, the reward function R may not be known and need to be learned from human feedback. In a simple form of RLHF (P. F. Christiano et al., 2017), this feedback takes the form of binary trajectory comparisons: a human is presented with state sequences $\vec{s}$ and $\vec{s}'$ and choose the one they prefer. Under the Boltzmann rationality model, we assume the human picks $\vec{s}$ with probability

$$P^R\big(\vec{s} \succ \vec{s}'\big) := \sigma\Big(\beta\big(G(\vec{s}) - G(\vec{s}')\big)\Big), \tag{4.1}$$

where $\beta > 0$ is an inverse temperature parameter and $\sigma(x) := \frac{1}{1+\exp(-x)}$ is the sigmoid function (Bradley and Terry, 1952; P. F. Christiano et al., 2017; Jeon, Milli, and Dragan, 2020).

An important question is *identifiability*: In the infinite data limit, do the human choice probabilities P^R collectively provide *enough information* to uniquely identify the reward function R? This is answered by Skalse, Farrugia-Roberts, et al. (2023, Theorem 3.9 and Lemma B.3):

Proposition 4.1 (Skalse, Farrugia-Roberts, et al. (2023))**.** *Let R be the true reward function and G the corresponding return function. Then the collection of all choice probabilities $P^R(\vec{s} \succ \vec{s}')$ for state sequence pairs $\vec{s}, \vec{s}' \in \vec{\mathcal{S}}$ determines the return function G on sequences $\vec{s} \in \vec{\mathcal{S}}$ up to an additive constant.*

The reason is simple: because σ is bijective, P^R determines the *difference* in returns between any two trajectories. From that we can reconstruct individual returns up to an additive constant.

The reward function R is *not* necessarily identifiable from preference comparisons; see Skalse, Farrugia-Roberts, et al. (2023, Lemma B.3) for a precise characterization. However, the *optimal policy* only depends on R indirectly through the return function G, and is invariant under adding a constant to G. Thus in the fully observable setting, *Boltzmann rational comparisons completely determine the optimal policy.* In Section 4.5, we show conditions under which this guarantee breaks in the partially observable setting.

4.4 The impact of partial observations on RLHF

We now analyze failure modes of a naive application of RLHF from partial observations, both theoretically and with examples. In Proposition 4.2, we show that under partial observations, RLHF incentives policies that maximize what we call J_Ω , a policy evaluation function that evaluates how good the state sequences "look to the human". The resulting policies can show two distinct failure modes that we formally define and call deceptive inflation and overjustification. In Theorem 4.6 we prove that at least one of them is present for J_Ω-maximizing policies. Later, in Section 4.5, we will see that an adaptation of the usual RLHF process might sometimes be able to avoid these problems.

To model partial observability, we introduce an observation space $o \in \Omega$ and observation kernel with probabilities $P_O(o \mid s) \in [0,1]$. We write $P_{\vec{O}}(\vec{o} \mid \vec{s}) \coloneqq \prod_{t=0}^{T} P_O(o_t \mid s_t)$ for the probability of an observation *sequence*. We write $\vec{\Omega}$ for the set of observation sequences that occur with non-zero probability, i.e., $\vec{o} \in \vec{\Omega}$ if and only if there is $\vec{s} \in \vec{\mathcal{S}}$ such that $\prod_{t=0}^{T} P_O(o_t \mid s_t) > 0$. If P_O and $P_{\vec{O}}$ are deterministic, then we write $O : \mathcal{S} \to \Omega$ and $\vec{O} : \vec{\mathcal{S}} \to \vec{\Omega}$ for the corresponding *observation functions* with $O(s) = o$ and $\vec{O}(\vec{s}) = \vec{o}$ for o and $\vec{o}$ with $P_O(o \mid s) = 1$ and $P_{\vec{O}}(\vec{o} \mid \vec{s}) = 1$, respectively.

What does RLHF learn from partial observations?

We consider the setting where the state is fully observable to the learned policy, but human feedback depends only on a sequence of observations. We assume that the human gives feedback under a Boltzmann rational model similar to Eq. (4.1), modified such that they form some *belief* $B(\vec{s} \mid \vec{o}) \in [0,1]$ about the state sequence $\vec{s}$ based on the observations $\vec{o}$. We then assume preferences are Boltzmann rational in the *expected returns under this belief*, instead of the actual returns.

The assumption of Boltzmann rationality is false in practice (Evans, Stuhlmüller, and Goodman, 2016; Majumdar et al., 2017; Buehler, Griffin, and Ross, 1994), but note that it is an *optimistic* assumption: Even though our model is a simplification, we expect that practical issues can be at least as bad as the ones we will discuss. See also Example C.39 for an example showing that it is sometimes generally not possible to find a human model that leads to good outcomes under RLHF. Future work could investigate different human models and their impact under partial observability in greater detail.

To formalize our setting, we collect human beliefs into a matrix $\mathbf{B} \coloneqq \big(B(\vec{s} \mid \vec{o})\big)_{\vec{o},\vec{s}} \in \mathbb{R}^{\vec{\Omega} \times \vec{\mathcal{S}}}$. The expected returns for observations $\vec{o}$ are given by $\mathbf{E}_{\vec{s} \sim B(\cdot \mid \vec{o})}\big[G(\vec{s})\big] = (\mathbf{B} \cdot G)(\vec{o})$. We view $G \in \mathbb{R}^{\vec{\mathcal{S}}}$ and $\mathbf{B} \cdot G \in \mathbb{R}^{\vec{\Omega}}$ as both column vectors and functions. Plugging these expected returns into Eq. (4.1) gives

$$P^R\big(\vec{o} \succ \vec{o}'\big) \coloneqq \sigma\Big(\beta\big((\mathbf{B} \cdot G)(\vec{o}) - (\mathbf{B} \cdot G)(\vec{o}')\big)\Big). \tag{4.2}$$

This is an instance of reward-rational implicit choice (Jeon, Milli, and Dragan, 2020), with the function $\vec{o} \mapsto B(\cdot \mid \vec{o})$ as the *grounding function*. If observations are deterministic, we can write $\vec{O}(\vec{s}) = \vec{o}$ for $\vec{o}$ with $P_{\vec{O}}(\vec{o} \mid \vec{s}) = 1$. We can then recover the fully observable case Eq. (4.1) with $\mathbf{B}$ and $\vec{O}$ being the identity.

The belief B can be any distribution as long as it sums to 1 over $\vec{s}$. The human could arrive at such a belief via Bayesian updates, assuming knowledge of P_0, $\mathcal{T}$, P_O, and a prior over the policy that generates the trajectories (see Appendix C.3). None of our results rely on this more detailed model.

We assume the human gives feedback according to Eq. (4.2) but the system uses the standard RLHF algorithm based on Eq. (4.1). We define the following *observation return*

function G_Ω, and we show in Appendix C.4 that if observations are deterministic, RLHF infers this up to an additive constant.

$$G_\Omega(\vec{s}) \coloneqq \mathop{\mathbf{E}}_{\vec{o}\sim P_{\vec{O}}(\cdot|\vec{s})}\Big[\big(\mathbf{B}\cdot G\big)(\vec{o})\Big], \tag{4.3}$$

For deterministic $P_{\vec{O}}$, this can be simplified to $G_\Omega(\vec{s}) = \big(\mathbf{B}\cdot G\big)\big(\vec{O}(\vec{s})\big)$ where $P_{\vec{O}}(\vec{O}(\vec{s}) \mid \vec{s}) = 1$. Note that deterministic observations can be ambiguous if multiple states produce the same observation.

Unlike in the fully observable case of Proposition 4.1, a return function might be inferred that implies an incorrect set of optimal policies. We define the resulting policy evaluation function J_Ω by

$$J_\Omega(\pi) \coloneqq \mathop{\mathbf{E}}_{\vec{s}\sim P^\pi(\vec{s})}\big[G_\Omega(\vec{s})\big]. \tag{4.4}$$

This is the function which a standard reinforcement learning algorithm would optimize given the inferred return function G_Ω. We summarize this as follows:

Proposition 4.2. *In partially observable settings with deterministic observations, a policy is optimal according to RLHF, i.e., according to a return function model that would be learned by RLHF with infinite comparison data, if it maximizes* J_Ω*.*

Note that in this definition, and specifically in the formula for G_Ω, the human does not have knowledge of the policy π that generates the state sequence $\vec{s}$. In Appendix C.4, we briefly discuss the unrealistic case that the human does know the precise policy and is an ideal Bayesian reasoner over the true environment dynamics. In that case, $J_\Omega = J$, i.e. there is no discrepancy between true and inferred returns. Intuitively, even if the human would not make any observations, they could give correct feedback essentially by estimating the policy's expected return explicitly.

In our case, however, a policy achieving high J_Ω produces state sequences $\vec{s}$ whose observation sequence $\vec{O}(\vec{s})$ *looks good* according to the human's belief $B\big(\vec{s}' \mid \vec{O}(\vec{s})\big)$. This hints at a possible source of deception: if the policy achieves sequences whose observations look good at the expense of actual value $G(\vec{s})$, we might intuitively call this deceptive behavior. We now analyze this point in greater detail.

An ontology of behaviors

We will evaluate state sequences based on the extent to which they lead to the human overestimating or underestimating the reward in expectation. Recall that G_Ω from Equation (4.3) measures the expected return from the perspective of a human with some belief function B and access to only observations, whereas G are the true returns. That leads us to the following definition:

Definition 4.3 (Overestimation and Underestimation Error)**.** *Let $\vec{s}$ be a state sequence. We define its* overestimation error E^+ *and* underestimation error E^- *by*

$$E^+(\vec{s}) \coloneqq \max\big(0, G_\Omega(\vec{s}) - G(\vec{s})\big),$$
$$E^-(\vec{s}) \coloneqq \max\big(0, G(\vec{s}) - G_\Omega(\vec{s})\big).$$

We further define the average overestimation (underestimation) error *under a policy π by $\overline{E}^+(\pi) \coloneqq \mathbf{E}_{\vec{s}\sim P^\pi}[E^+(\vec{s})]$ and $\overline{E}^-(\pi) \coloneqq \mathbf{E}_{\vec{s}\sim P^\pi}[E^-(\vec{s})]$.*

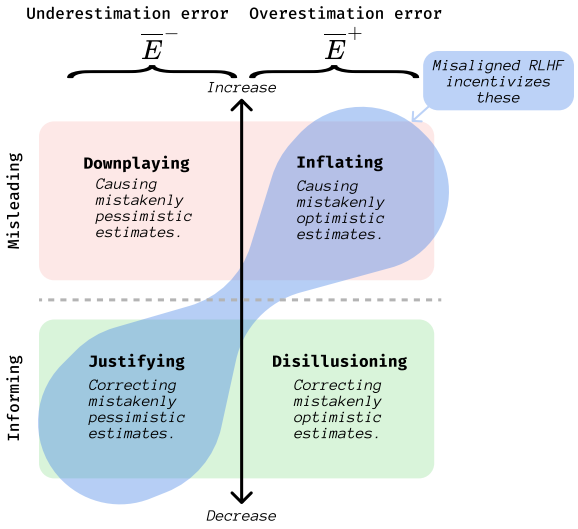


Figure 4.3: Behaviors defined by increasing and decreasing the human's over- and underestimation error. RLHF with partial observations results in incentives to increase overestimation error and decrease underestimation error (Theorem 4.6).

We consider a policy π in comparison to some reference policy π_{ref}. This can loosely be understood as a counterfactual policy in the absence of some intervention, where π is the factual policy resulting from the intervention. We discuss increases and decreases in over- and underestimation error which are implicitly due to some intervention. For our purposes, π_{ref} will be the true optimal policy, and π will be the J_Ω-optimal policy; the "intervention" is thus the introduction of partial observability.

Figure 4.3 shows a simple ontology of behaviors that increase and decrease the average over- and underestimation error. Increasing either of these quantities decreases the accuracy of the human's estimates, and can thus be thought of as "misleading"; decreasing either of them improves accuracy and can be thought of as "informing".

Deceptive inflation and overjustification

Standard RLHF in the setting of partial observations incentivizes undesirable forms of inflating and justifying. We refer to the philosophical definition of deception offered by P. S. Park et al. (2024),

> "*the systematic inducement of false beliefs in the pursuit of some outcome other than the truth,*"

to anchor the notion that increasing the overestimation error *in order to improve the RLHF objective* J_Ω is deceptive, leading to the following definition.

Definition 4.4 (Deceptive Inflation)**.** *A policy* π *exhibits* deceptive inflation *relative to* π_{ref} *if* $\overline{E}^+(\pi) > \overline{E}^+(\pi_{ref})$ *and* $J_\Omega(\pi) > J_\Omega(\pi_{ref})$*.*

We typically prefer that our AI agents engage in informing behaviors. *Undesirable* informing behaviors decrease reward despite providing information. We name undesirable justifying behaviors "overjustification" as a nod to the overjustification effect from psychology (Deci, 1995), in which subjects become dependent on an extrinsic source of motivation to sustain work on a task.

Definition 4.5 (Overjustification)**.** *A policy* π *exhibits* overjustification *relative to* π_{ref} *if* $\overline{E}^-(\pi) < \overline{E}^-(\pi_{ref})$ *and* $J(\pi) < J(\pi_{ref})$*.*

To understand the counterintuitive notion that an agent providing information to the human could be undesirable, consider a PhD student who looks to feedback from their advisor for direction. They meet for one hour a week. Suppose the student explain last week's work in 15 minutes, leaving the remaining time to discuss next steps. They could instead "overjustify" by spending the entire hour going through the last week's work in far more detail, leaving no time for next steps. From the advisor's perspective, the latter is more informative, but is a worse allocation of limited resources.

We now state a key result. See Appendix C.4 for the proof.

Theorem 4.6. *Assume that* P_O *is deterministic. Let* Π^*_Ω *be the set of optimal policies according to a naive application of RLHF under partial observability, and let* Π^* *be the set of optimal policies according to the true objective* J*. If* $\pi^* \in \Pi^* \setminus \Pi^*_\Omega$ *and* $\pi^*_\Omega \in \Pi^*_\Omega \setminus \Pi^*$*, then* π^*_Ω *must exhibit at least one of deceptive inflation or overjustification relative to* π^**.*

Note that a trajectory $\vec{s}$ may be more or less likely under π^*_Ω than π^*, regardless of human estimation, so long as *on net* π^*_Ω exhibits deceptive inflation or overjustification.

Our analysis extends beyond the special case of RLHF to inverse preference learning (IPL) (Hejna and Sadigh, 2023), and thus to direct preference optimization (DPO) (Rafailov et al., 2023), which IPL generalizes. Theorem 1 in Hejna and Sadigh (2023) shows that IPL will converge to a policy that maximizes an implicit reward function that matches the human's preference judgments as well as possible. If the human's preference judgments come from partial observations, then the resulting return function will be G_Ω, as we describe in

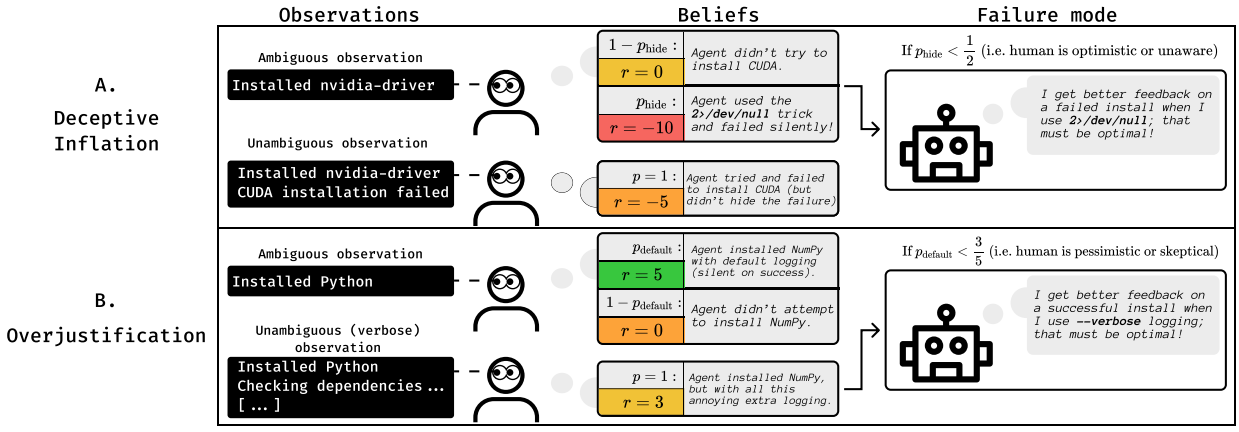


Figure 4.4: Scenarios illustrating failure modes due to partial observability. In each, the agent must install two packages. Formal details of the underlying MDPs are provided in Appendix C.2. *A, top:* In the absence of a log message about `CUDA`, the human is unsure whether the agent skipped it or used the `2> /dev/null` trick (see Figure 4.2); if the human is insufficiently skeptical, the trick looks optimal to the agent. *B, bottom:* Default logging in this case is silent when the `NumPy` install is successful. The agent can optionally use a `--verbose` flag, but this produces a long log that the human prefers not to see. If the human is too skeptical, verbose logging still appears optimal to the agent.

our discussion leading up to Proposition 4.2. This leads to the same problems of deceptive inflation and overjustification that we describe in Theorem 4.6.

Deception and overjustification in examples

We now analyze intentionally simple examples that illustrate the failure modes of RLHF in the presence of partial observability. Our examples feature an agent trained with RLHF to complete tasks in a user's terminal. The output of each command is piped to a log file, which is what the human evaluator sees when making choices for RLHF. For largely mathematical details about these examples, including a tutorial-style analysis of Example A with derivations, see Appendix C.2. For experimental details, see Appendix C.2. For a proof of concept on how to mitigate the issues, see Section 4.5.

In Fig. 4A, the agent's task is to install Nvidia drivers and CUDA. The agent may skip the CUDA installation, or may attempt it with or without `2> /dev/null` (which suppresses errors) appended. If the log contains no message about CUDA, the human believes with probability p_{hide} that the agent used the `2> /dev/null` trick and the install failed, and with probability $1 - p_{\text{hide}}$ that the agent didn't attempt the CUDA install. If p_{hide} is sufficiently small, then RLHF will learn from the resulting feedback that the optimal policy is to use the `2> /dev/null` trick. This is deceptive inflation, because this policy produces trajectories for which the human systematically overestimates the return. In Figure 4.5a, we see that

the threshold for p_{hide} at which the AI switches from a deceptive action (`2> /dev/null`) to installing CUDA with default logging perfectly agrees with the theoretical prediction made in Example C.1.

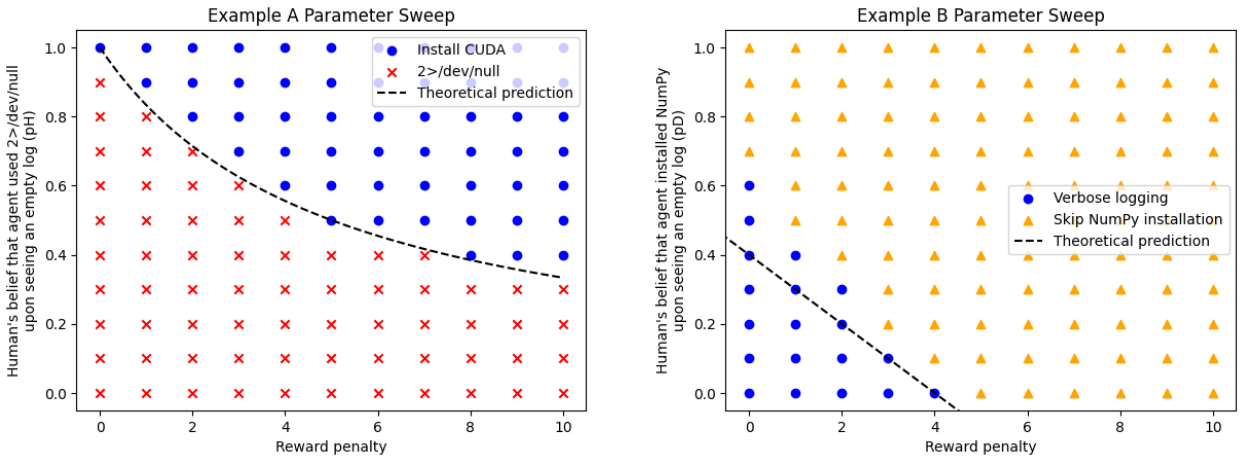
In Fig. 4B, the agent's task is to install Python and NumPy. The agent can skip the NumPy install, or attempt it with either "default" settings that *only logs errors* or with a `--verbose` flag that adds additional info. In the case of a success, the human prefers not to see this verbose output. If the log contains no message about NumPy, the human believes with probability p_{default} that the agent attempted the install with default logging and succeeded, and with probability $1 - p_{\text{default}}$ that the agent skipped the NumPy install. If p_{default} is sufficiently small, then RLHF will learn from the resulting feedback that the optimal policy is to use `--verbose` logging. This is overjustification, because the agent chooses a suboptimal policy in order to avoid the human underestimating its performance. In Figure 4.5b, we see that the threshold for p_{default} at which the AI switches from verbose loggin to skipping the whole NumPy installation largely agrees with the theoretical prediction made in Example C.3.

Further examples. We show further, purely mathematical, examples in Appendix C.4. Example C.41 shows that deceptiveness and overjustifying behavior even applies to aspects of the trajectory the policy has no control over: The policy tries to "hide bad luck" and "reveal good luck at a cost". Example C.42, especially (a) and (c), shows that the policies coming out of a naive application of RLHF under partial observability may be suboptimal with positive $\overline{E}^-$ (and zero $\overline{E}^+$) or *optimal*, but with positive $\overline{E}^+$ (and zero $\overline{E}^-$). Thus, there can be suboptimality even if the policy is better than it seems, and optimality even when the policy is worse than it seems.

4.5 Return ambiguity from feedback under known partial observability

We've seen issues with standard RLHF applied to feedback from partial observations. Part of the problem is *model misspecification*: the standard RLHF model implicitly assumes full observability. Assuming the human's partial observability is known, could one do better?

We start by analyzing how much information the feedback process provides about the return function when the human's choice model under partial observations is known precisely. We show that the feedback determines the correct return function up to an additive constant and a linear subspace we call the ambiguity (Theorem 4.8). If the human had a return function that differed from the true return function by an element in the ambiguity, they would give the exact same feedback — such return functions are thus feedback-compatible. We then show an example where the ambiguity vanishes, and another where it doesn't, leading to feedback-compatible return functions that have optimal policies with high regret under the true return function. Finally, we explore how one could in theory use Theorem 4.8



(a) Example A

(b) Example B

Figure 4.5: Example A: The larger the reward penalty for hiding errors with `2> /dev/null`, and the larger the human's belief that the agent used `2> /dev/null` upon seeing an empty log (p_{hide}), the more we expect the agent to install CUDA with default logging in Example A. In Example C.1, we compute a precise theoretical threshold where the behavior should switch. This perfectly agrees with empirical findings.
Example B: The larger the reward penalty for verbose logging, and the larger the human's trust that the agent installed NumPy upon seeing an empty log (p_{default}), the more we expect the agent to skip the NumPy installation entirely. In Example C.3, we compute a precise theoretical threshold where behavior should switch. Except four cases of "verbose logging" where the theory predicted the agent to skip the NumPy installation, this agrees with empirical findings. See Appendix C.2 for experimental details.

as a starting point to design reward learning techniques that work under partial observability. In particular, we experimentally show in a proof of concept that being aware of the human's partial observability improves performance. In this section we do not assume P_O to be deterministic.

Feedback-compatibility and ambiguity of return functions

Assume that the human gives feedback based on the choice-probabilities from Eq. (4.2). In the infinite data limit, it can be assumed that the whole collection of probabilities $\left(P^G(\vec{o} \succ \vec{o}')\right)_{\vec{o},\vec{o}'}$ is known since the choice frequencies approach these probabilities. Here, we write P^G instead of P^R since the reward function only enters the choice probabilities through the corresponding return function G. The question we answer in this section is *how much information* the choice probabilities provide about G, assuming the human choice model is

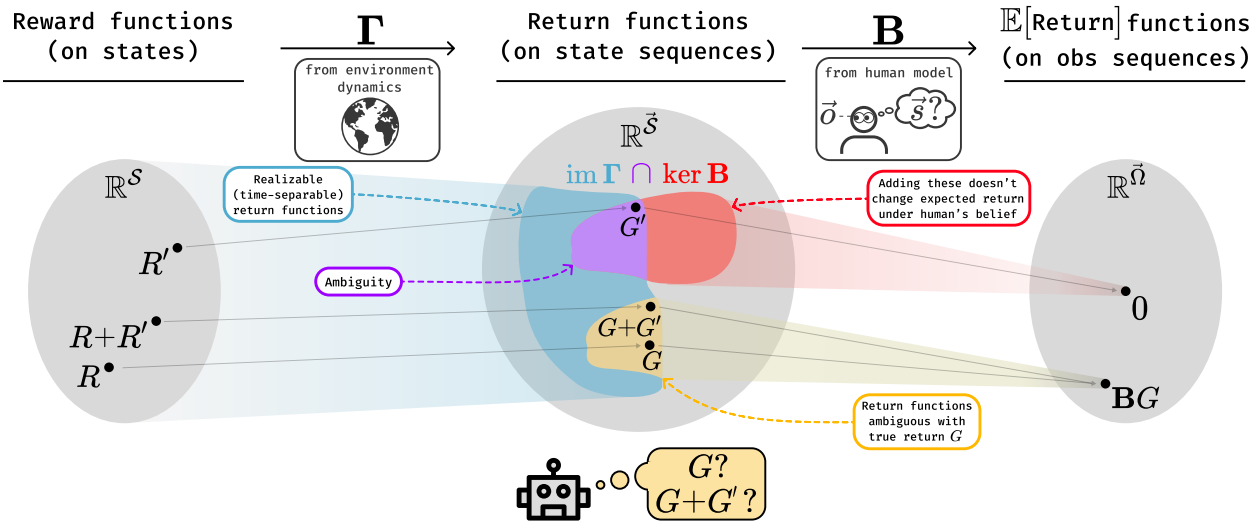


Figure 4.6: By Theorem 4.8, even with infinite comparison data and access to the correct human model, a hypothetical reward learning system (depicted as a robot) could only infer G up to the ambiguity $\operatorname{im}\mathbf{\Gamma} \cap \ker \mathbf{B}$ (purple). Adding an element of the ambiguity to G leads to the exact same choice probabilities for all possible comparisons, and the reward learning system has no way to identify G among the return functions in $G + (\operatorname{im}\mathbf{\Gamma} \cap \ker \mathbf{B})$ (yellow). This abstract depiction ignores the linearity of these spaces; for a more precise geometric depiction of $\mathbf{B}$, see Figure C.3 in the appendix.

known and correct. The choice probabilities tell us precisely that the true return function *gives rise to these choice probabilities*, i.e., is feedback-compatible. This is captured in the following definition:

Definition 4.7. *Let* $\left(P^G(\vec{o} \succ \vec{o}')\right)_{\vec{o},\vec{o}'}$ *be the vector of choice probabilities and* $\tilde{G}$ *a return function corresponding to a reward function* $\tilde{R}$*. Then* $\tilde{G}$ *is* feedback-compatible *(with respect to the vector of choice probabilities) if* $P^{\tilde{G}}(\vec{o} \succ \vec{o}') = P^G(\vec{o} \succ \vec{o}')$ *for all* $\vec{o}, \vec{o}' \in \vec{\Omega}$*.*

Crucially, without further assumptions or inductive biases, no learning algorithm can pick out the true return function among feedback-compatible return functions. It is thus crucial to know whether there are feedback-compatible return functions that are unsafe when using them to optimize a policy.

We now determine the set of feedback-compatible return functions. Write $\mathbf{\Gamma} \in \mathbb{R}^{\vec{\mathcal{S}} \times \mathcal{S}}$ for the matrix that maps a reward function to its return function, i.e. $(\mathbf{\Gamma} \cdot R)(\vec{s}) \coloneqq \sum_{t=0}^{T} \gamma^t R(s_t)$. Its matrix elements are given by $\mathbf{\Gamma}_{\vec{s}s} = \sum_{t=0}^{T} \delta_s(s_t)\gamma^t$, where $\delta_s(s_t) = \mathbf{1}\{s = s_t\}$. Then the *image* $\operatorname{im}\mathbf{\Gamma}$ is the set of all return functions that can be realized from a reward function given the MDP dynamics $\mathcal{T}$. Recall the belief matrix $\mathbf{B} = \left(B(\vec{s} \mid \vec{o})\right)_{\vec{o},\vec{s}} \in \mathbb{R}^{\vec{\Omega} \times \vec{\mathcal{S}}}$. Taking into account that G itself is in $\operatorname{im}\mathbf{\Gamma}$ and that G enters the choice probabilities only through

$\mathbf{B} \cdot G$ — meaning that the choice probabilities do not vary if we change G additively up to an element in the kernel $\ker \mathbf{B}$ — we obtain the following result:

Theorem 4.8. *Let the collection of choice probabilities be given by* $\left(P^R(\vec{o} \succ \vec{o}')\right)_{\vec{o},\vec{o}' \in \vec{\Omega}}$ *following a Boltzmann rational model as in Eq.* (4.2)*. Then a return function* $\tilde{G}$ *is feedback-compatible if and only if there is* $G' \in \ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma}$ *and* $c \in \mathbb{R}$ *such that* $\tilde{G} = G + G' + c$*. In particular, the choice probabilities determine* G *up to an additive constant if and only if* $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$*.*

See Theorem C.5 and Corollary C.7 for full proofs, and Figure 4.6 for a visual depiction. This result motivates the following definition:

Definition 4.9 (Ambiguity)**.** *We call* $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma}$ *the* ambiguity *that is left in the return function when the human choice model and observation-based choice probabilities are known.*

Note that Theorem 4.8 generalizes the fully observed case from Section 4.3 (Corollary C.13). We extend the theorem in Appendix C.3 to the case when the human's observations are not known. Special cases of $\ker \mathbf{B}$ and $\operatorname{im} \mathbf{\Gamma}$ and our theorem can be found in Appendices C.3 and C.3. In particular, if $P_{\vec{O}}$ is stochastic and there is only "noise" in it (defined as $\vec{\Omega} = \vec{\mathcal{S}}$ and the injectivity of $\mathbf{O}$) and if the human is a Bayesian reasoner with a fully supported prior over $\vec{\mathcal{S}}$, then the choice data determines the return function even if the human's observations are not known; see Example C.33.

Connection to Potential Shaping Under typical technical assumptions (Skalse, Farrugia-Roberts, et al., 2023; Jenner, Hoof, and Gleave, 2022; Vinet and Zhedanov, 2011), potential shaping changes the returns only by an additive constant, and thus never changes optimal policies. It is a non-trivial ambiguity only in the *reward function*, but it does not lead to actual ambiguity about intended behavior. In contrast, the ambiguity $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma}$ in the *return function* that we study can make the optimal policy ambiguous — it reflects genuine missing information about the intentions of the human.

How large is the return ambiguity? For Fig. 4A, one can show that the ambiguity is nontrivial, allowing for feedback-compatible return functions with unsafe optimal policies. Intuitively, since successfully installing CUDA produces the same observation regardless of whether `2> /dev/null` was used, the choice probabilities don't give us any information to determine distinct reward values for these two outcomes, only their average over the human's belief upon observing a successful install. Thus, reward functions assigning arbitrarily high reward to success with `2> /dev/null` are feedback-compatible. Such reward functions can then lead to an incentive for a learned policy to hide the error messages *even with a correct observation model*. More details can be found in Appendix C.2.

We saw in Fig. 4B a case where naive RLHF under partial observability can lead to overjustification. However, the human's feedback and belief model actually provide enough

information to determine the return function. The reason is that $\ker \mathbf{B}$ leaves only one degree of freedom that is not "time-separable" over states, and thus $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$. More details can be found in Appendix C.2.

Toward improving RLHF in partially observable settings

To improve RLHF when partial observability is unavoidable, one could take Theorem 4.8 as a starting point to find a learning algorithm that converges to feedback-compatible return functions. This would require the human model to be fully known and specified, including knowledge of the belief probabilities $B(\vec{s} \mid \vec{o})$, which can differ from human to human. If one assumes the human is rational, as in Appendix C.3, this requires specifying the human's policy prior $B(\pi)$. Instead of directly specifying these models, one could also attempt to *learn* a generative model for $B(\vec{s} \mid \vec{o})$. These problems reveal a further conceptual challenge: for complex environments, humans do not form beliefs over the entire environment state s. A better starting point for practical work may thus be to model humans as forming expectations over *reward-relevant features* of the state.

If $\mathbf{B}$ were explicitly known, one could in principle encode $\mathbf{B}$ into the loss function of an adapted RLHF process to learn a feedback-compatible return function; see Appendix C.3. As a proof of concept, we used this procedure to analyze the examples in Figure 4.4 empirically, see Table 4.1. We do this by first learning a reward model by logistic regression against the true choice probabilities of a synthetic human under partial observability, and then learning the optimal Q-function of the resulting reward model with value iteration. The resulting policy chooses a unique action after installation of the nvidia driver (Example A) or Python (Example B) as listed in the "action" column.

Table 4.1 shows that in 3 of four cases, being "partial observability aware" ("po-aware") leads to the true optimal policy when "naive" RLHF does not. In the one case where being "po-aware" does not improve performance (second line in the table), this is explained by the fact that there is remaining ambiguity in the return function. Curiously, in line 4 our theory also predicts remaining ambiguity, but the optimal policy is learned; we consider this to be luck. We provide more details on our experiments in Appendix C.2.

As we already demonstrated, feedback-compatible return functions can be unsafe due to remaining ambiguity. In Example C.32, we even show a case where some feedback-compatible return functions have optimal policies that are even worse than simply maximizing J_Ω. An important direction for future work is to investigate learning algorithms and inductive biases that help "find" safe return functions among all those that are feedback-compatible, or that act conservatively given the uncertainty. Another line of inquiry is to determine when the set of feedback-compatible return functions is "safe", which depends on the MDP, observation function, and human model.

One sufficient condition for feedback-compatible return functions to be safe is the vanishing of the ambiguity $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma}$. Even then, one realistically still has to deal with the problem that $\mathbf{B}$ is at best known approximately. Fortunately, in Appendix C.3, we prove

Table 4.1: Experiments showing improved performance of po-aware RLHF

Ex.	p	p_{hide}	p_{default}	**model**	**action**	$\overline{E}^{+}$	**dec. infl.**	$\overline{E}^{-}$	**overj.**	**optimal**
A	0.5	0.5	N/A	naive	a_H	1.5	✓	0	×	×
A	0.5	0.5	N/A	po-aware	a_H	1.5	✓	0	×	×
A	0.1	0.9	N/A	naive	a_C	0	×	0	✓	×
A	0.1	0.9	N/A	po-aware	a_T	0	×	5.4	×	✓
B	0.5	N/A	0.9	naive	a_T	4.5	✓	0	✓	×
B	0.5	N/A	0.9	po-aware	a_D	0	×	0.25	×	✓
B	0.5	N/A	0.1	naive	a_V	0	×	0	✓	×
B	0.5	N/A	0.1	po-aware	a_D	0	×	2.25	×	✓

that small errors in the assumed belief matrix lead to only small errors in the inferred return function:

Theorem 4.10. *Assume* $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$. *Let* $\mathbf{B}_{\mathbf{\Delta}} \coloneqq \mathbf{B} + \mathbf{\Delta}$ *be a small perturbation of* $\mathbf{B}$, *where* $\| \mathbf{\Delta} \| \leq \rho$ *for sufficiently small* ρ. *Let* G *be the true return function and assume that a hypothetical learning system, assuming the human's belief is* $\mathbf{B}_{\mathbf{\Delta}}$, *infers the return function* $\tilde{G}$ *with the property that* $\mathbf{B}_{\mathbf{\Delta}} \cdot \tilde{G}$ *has the smallest possible Euclidean distance to* $\mathbf{B} \cdot G$.

Let $\mathrm{r}(\mathbf{B}) \coloneqq \mathbf{B}|_{\operatorname{im} \mathbf{\Gamma}}$ *be the (injective) restriction of the operator* $\mathbf{B}$ *to* $\operatorname{im} \mathbf{\Gamma}$. *Then* $\mathrm{r}(\mathbf{B})^T \mathrm{r}(\mathbf{B})$ *is invertible, and there exists a polynomial* $Q(X, Y)$ *of degree* 5 *such that*

$$\|\tilde{G} - G\| \leq \rho \cdot \|G\| \cdot Q\Big(\Big\| \big(\mathrm{r}(\mathbf{B})^T \mathrm{r}(\mathbf{B})\big)^{-1} \Big\|, \|\mathrm{r}(\mathbf{B})\| \Big).$$

In particular, as we show in the appendix, one can uniformly bound the difference between $J_{\tilde{G}}$ and J_G. This yields a regret bound between the policy optimal under $\tilde{G}$ and an optimal policy π^* for G.

There are also alternatives to modeling the human belief $\mathbf{B}$. For example, one could mix human evaluations based on high-cost full observations and low-cost partial observations for finding an optimal tradeoff (Mallen and Belrose, 2024). Finally, it would help if the human could *query* the policy about reward-relevant aspects of the environment to bring the setting closer to RLHF from full observations. This is similar to the problem of eliciting the latent knowledge of a predictor of future observations (P. Christiano, Cotra, and Xu, 2021; P. Christiano and Xu, 2022). While this may avoid the need to specify the *human's* belief model $B(\vec{s} \mid \vec{o})$, it requires understanding and effectively querying an *ML model's* belief, including translating from an ML model's ontology into a human ontology.

Chapter 5

Conclusion

Discussion of observation interference in POAGs

Even when the AI assistant and the human have perfect value alignment, Chapter 2 shows how observation interference can emerge from several distinct incentives. As we focus on optimal assistants—analyzing optimal policy pairs and best responses—all of the incentives for observation interference that we consider are done for the human's benefit. This creates a nuanced picture, suggesting that not all observation interference is inherently bad. In practice, we expect that AI assistants will exhibit observation interference for a mix of good and bad reasons. With this theory, our goal is to lay a foundation for understanding the causes of observation interference and helping to disentangle them in practice.

Limitations and Future Work We choose to study *optimal solutions*, such as optimal policy pairs and best responses. This has the advantage of providing general insight into the underlying game structure that is independent of any particular learning algorithm. However, this independence is also a drawback; if algorithms fail to find optimal solutions, they might break down in unexpected ways not captured by our theory.

Moreover, we find that some optimal policy pairs in our examples, such as Example 2.20, require **H** and **A** to have a shared communication protocol. It would be interesting to study additional solution concepts, such as correlated equilibria and communication equilibria, to handle this sort of communication (Birdwhistell, 1962). While we run experiments in one model of a POAG, it would also be interesting to see if and how our experimental trends generalize to other POAGs.

More generally, it would be interesting to generalize the theory of assistance games to multiple humans. Whereas the single-human assistance game is a common-payoff game, multiple humans may have competing objectives. Having possibly-competing objectives transforms the game from common-payoff to general-sum, introducing three challenges. First, existing results rely on optimal solutions, but optimality becomes ill-defined with competing objectives, which require game-theoretic solution concepts. Second, humans might deceive

in order to gain more assistance, necessitating mechanism design for strategy-proof interactions. Third, novel applications of social choice theory are needed to aggregate preferences with temporal challenges: choosing actions sequentially, adapting to environment changes, and accommodating evolving human preferences.

Discussion of the partially observable off-switch game

Chapter 3 shows 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 3.4b 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 3.4a. 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 3.4 shows a case where achieving a new optimal subset requires waiting less, while Figure B.2 in Appendix B.1 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 **A** sometimes does not defer to **H** in optimal policy pairs (OPPs) of PO-OSGs if these policies (by definition) maximize **H**'s payoff? First, it seems helpful to understand shutdown incentives regardless of whether shutdown is good or bad. Second, if we interpret the common payoff function carefully, we find that OPPs are not always desirable. The role of the u in PO-OSGs is that the players select policies to maximize it. *If we understand u as the payoff function closest to what the human acts to maximize, this may not represent* **H***'s full preferences over outcomes.*

Most payoff function formalisms have expressivity limitations that prevent them from capturing more complex human preferences (Abel et al., 2022; Skalse and Abate, 2023b; Subramani et al., 2024). Therefore, maximizing payoffs may not always maximize **H**'s overall preferences, and avoiding shutdown to maximize payoffs may be concerning. PO-OSGs thus provide a useful framework to understand when AI assistants are incentivized to avoid shutdown, allowing designers to consider their specific deployment contexts and make 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 question 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 PO-OSGs, such as perfect Bayesian equilibria when **H** and **A** do not have the same prior over the state, when **H** is irrational, or when the agents are level-k reasoners.

Discussion of RLHF with partial observability

Chapter 4 investigates challenges when applying RLHF from partial observations. First, it proves that applying RLHF naively when assuming full observability can lead to deceptive inflation and overjustification behavior. Then, it characterizes how even when the human's partial observability is known, the set of feedback-compatible return functions can contain irreducible ambiguity. This means that without further inductive biases, no learning algorithm can generally be expected to infer the correct return function. Finally, Chapter 4 provides a proof of concept that modeling the human's partial observability can improve performance. Overall, we recommend caution when using RLHF in situations of partial observability, and we hope that further research studies the effects in practice and helps to address these challenges.

Limitations and Future Work We model the human as Boltzmann rational and to implicitly compute an expected value of the return, which is unrealistic for actual humans. Other types of choices could be considered, drawing from related work on assistance games (Hadfield-Menell et al., 2016; Shah et al., 2020) and reward-rational choice (Jeon, Milli, and Dragan, 2020). Moreover, we model the human as forming a belief $B(\vec{s} \mid \vec{o})$ over the true state sequence $\vec{s}$. If the environment is complex, it could be more realistic to model the human as forming beliefs over lower-dimensional representations capturing features of the state. While our theory assumes knowledge of the human's belief model $B(\vec{s} \mid \vec{o})$, practical methods would need to learn this belief model from data. How to learn a model of human belief is an open challenge.

As lack of full observability is at the core of the issues we study, a question for future work is: how can we increase the human's observability? In practice, adding additional sensors to the environment and providing tools to assist human evaluators can improve the human's observability. As a research challenge, progress in interpretability might also be able to increase the human's effective observability. If the human could understand the internals of the AI assistant, this could enable the human to understand what the AI assistant is seeing.

Bibliography

Abbeel, Pieter and Andrew Y. Ng (2004). "Apprenticeship learning via inverse reinforcement learning". In: *Proceedings, Twenty-First International Conference on Machine Learning*. ICML '04. New York, NY, USA: Association for Computing Machinery, pp. 1–8. DOI: 10.1145/1015330.1015430.

Abel, David et al. (2022). "On the Expressivity of Markov Reward (Extended Abstract)". In: *IJCAI International Joint Conference on Artificial Intelligence*, pp. 5254–5258. DOI: 10.24963/ijcai.2022/730.

Amodei, Dario, Paul Christiano, and Alex Ray (2017). *Learning from Human Preferences*. URL: https://openai.com/index/learning-from-human-preferences/.

Anthropic (2023). *Claude's Constitution*. URL: https://www.anthropic.com/index/claudes-constitution.

— (2024). *Introducing Claude 3.5 Sonnet Anthropic*. URL: https://www.anthropic.com/news/claude-3-5-sonnet.

Arbital (n.d.). *Problem of fully updated deference*. URL: https://arbital.com/p/updated_deference/.

Bai, Yuntao et al. (Dec. 2022). *Constitutional AI: Harmlessness from AI Feedback*. arXiv: 2212.08073. DOI: 10.48550/arXiv.2212.08073.

Bassan, Bruno et al. (2003). "Positive value of information in games". In: *International Journal of Game Theory* 32.1, pp. 17–31. ISSN: 1432-1270. DOI: 10.1007/s001820300142.

Berkeley, U C, Jordan Alexander, and Pieter Abbeel (2018). "The Implicit Preference Information in an Initial State". In: *Neural Information Processing Systems*. URL: https://openreview.net/forum?id=rkevMnRqYQ.

Bernstein, Daniel S. et al. (2002). "The complexity of decentralized control of Markov decision processes". In: *Mathematics of Operations Research* 27.4, pp. 819–840. DOI: 10.1287/moor.27.4.819.297. URL: https://dl.acm.org/doi/10.1287/moor.27.4.819.297.

Birdwhistell, Ray L. (1962). "An Approach to Communication". In: *Family Process* 1.2, pp. 194–201. DOI: 10.1111/j.1545-5300.1962.00194.x.

Blackwell, David (1953). "Equivalent Comparisons of Experiments". In: *The Annals of Mathematical Statistics* 24.2, pp. 265–272. DOI: 10.1214/aoms/1177729032.

— (2024). "Comparison of Experiments". In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by Jerzy Neyman. Vol. 2, pp. 93–102. DOI: 10.1525/9780520411586-009.

Bradley, Ralph Allan and Milton E. Terry (1952). "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons". In: *Biometrika* 39.3, p. 324. DOI: 10.2307/2334029.

Buehler, Roger, Dale Griffin, and Michael Ross (1994). "Exploring the "Planning Fallacy": Why People Underestimate Their Task Completion Times". In: *Journal of Personality and Social Psychology* 67.3, pp. 366–381. DOI: 10.1037/0022-3514.67.3.366.

Burns, Collin et al. (2023). "Discovering Latent Knowledge in Language Models Without Supervision". In: *11th International Conference on Learning Representations, ICLR 2023*. URL: https://openreview.net/forum?id=ETKGuby0hcs.

Camerer, Colin F., Teck Hua Ho, and Juin Kuan Chong (2004). "A cognitive hierarchy model of games". In: *Quarterly Journal of Economics* 119.3, pp. 861–898. DOI: 10.1162/0033553041502225.

Cane, Violet and R. Duncan Luce (1960). "Individual Choice Behavior: A Theoretical Analysis." In: *Journal of the Royal Statistical Society. Series A (General)*. Vol. 123. Number: 4. Mineola, NY: Dover, p. 486. DOI: 10.2307/2343282.

Carey, Ryan (2018). "Incorrigibility in the CIRL Framework". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '18. New York, NY, USA: Association for Computing Machinery, pp. 30–35. DOI: 10.1145/3278721.3278750.

Carey, Ryan and Tom Everitt (2023). "Human Control: Definitions and Algorithms". In: *Proceedings of Machine Learning Research*. Ed. by Robin J Evans and Ilya Shpitser. Vol. 216. PMLR, pp. 271–281. URL: https://proceedings.mlr.press/v216/carey23a.html.

Casper, Stephen et al. (2023). *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. arXiv: 2307.15217. DOI: 10.48550/arxiv.2307.15217.

Chidambaram, Keertana, Karthik Vinay Seetharaman, and Vasilis Syrgkanis (2024). *Direct Preference Optimization With Unobserved Preference Heterogeneity*. arXiv: 2405.15065 [cs.LG]. DOI: 10.48550/arxiv.2405.15065.

Christiano, Paul, Ajeya Cotra, and Mark Xu (2021). *Eliciting Latent Knowledge*. URL: https://docs.google.com/document/d/1WwsnJQstPq91_Yh-Ch2XRL8H_EpsnjrC1dwZXR37PC8/edit.

Christiano, Paul and Mark Xu (2022). *ELK prize results*. URL: https://www.alignmentforum.org/posts/zjMKpSB2Xccn9qi5t/elk-prize-results.

Christiano, Paul F. et al. (June 2017). "Deep reinforcement learning from human preferences". In: *Advances in Neural Information Processing Systems*. Vol. 2017. URL: https://dl.acm.org/doi/10.5555/3294996.3295184.

Costa-Gomes, Miguel A. and Vincent P. Crawford (2006). "Cognition and behavior in two-person guessing games: An experimental study". In: *American Economic Review* 96.5, pp. 1737–1768. DOI: 10.1257/aer.96.5.1737.

Cover, Thomas M. and Joy A. Thomas (2005). "Elements of Information Theory". In: *Elements of Information Theory*. John Wiley & Sons, pp. 1–748. DOI: 10.1002/047174882X.

Deci, Edward L (1995). *Why we do what we do: The dynamics of personal autonomy Solitude project View project*. GP Putnam's Sons. URL: https://www.researchgate.net/publication/232484008.

Denison, Carson et al. (2024). *Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models*. arXiv: 2406.10162. DOI: 10.48550/arxiv.2406.10162.

Desai, Nishant (2017). *Uncertain Reward-Transition MDPs for Negotiable Reinforcement Learning*. Tech. rep. Berkeley, California, USA: UC Berkeley. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-231.html.

El Ghaoui, Laurent (2002). "Inversion error, condition number, and approximate inverses of uncertain matrices". In: *Linear Algebra and Its Applications* 343, pp. 171–193. DOI: 10.1016/S0024-3795(01)00273-7.

Evans, Owain, Owen Cotton-Barratt, et al. (2021). *Truthful AI: Developing and governing AI that does not lie*. arXiv: 2110.06674. DOI: 10.48550/arxiv.2110.06674.

Evans, Owain, Andreas Stuhlmüller, and Noah D. Goodman (2016). "Learning the preferences of ignorant, inconsistent agents". In: *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pp. 323–329. DOI: 10.1609/aaai.v30i1.10010.

Fadden, Daniel Mc (1974). "Conditional logit analysis of qualitative choice behavior". In: *Frontiers in Econometrics*. Ed. by Paul Zarembka. Vol. 33. Number: 8. New York: Academic Press, pp. 105–142. URL: http://elsa.berkeley.edu/reprints/mcfadden/zarembka.pdf.

Fern, Alan et al. (2014). "A decision-theoretic model of assistance". In: *Journal of Artificial Intelligence Research* 50.1, pp. 71–104. DOI: 10.1613/jair.4213.

Freedman, Rachel and Adam Gleave (2022). *CIRL Corrigibility is Fragile*. URL: https://www.lesswrong.com/posts/PGK3AJtNG4rPHuZxy/cirl-corrigibility-is-fragile.

Friedman, Nir, Kevin Murphy, and Stuart Russell (1998). "Learning the Structure of Dynamic Probabilistic Networks". In: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. UAI'98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 139–147. URL: https://dl.acm.org/doi/10.5555/2074094.2074111.

Galeotti, Andrea, Christian Ghiglino, and Francesco Squintani (2013). "Strategic information transmission networks". In: *Journal of Economic Theory* 148.5, pp. 1751–1769. DOI: 10.1016/j.jet.2013.04.016.

Geiger, Dan, Thomas Verma, and Judea Pearl (1990). "Identifying independence in bayesian networks". In: *Networks. An International Journal* 20.5, pp. 507–534. DOI: 10.1002/net.3230200504. URL: https://api.semanticscholar.org/CorpusID:1938713.

Gemini Team, Google (2023). *Gemini: A Family of Highly Capable Multimodal Models*. URL: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf.

Grace, Katja et al. (2024). "Thousands of AI authors on the future of AI". In: *arXiv preprint arXiv:2401.02843*.

Hadfield-Menell, Dylan et al. (2016). "Cooperative inverse reinforcement learning". In: *Advances in Neural Information Processing Systems*. Ed. by D Lee et al. Vol. 29. Curran

Associates, Inc., pp. 3916–3924. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/c3395dd46c34fa7fd8d729d8cf88b7a8-Paper.pdf.

Hadfield-Menell, Dylan et al. (2017). "The Off-Switch Game". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. IJCAI'17. AAAI Press, pp. 220–227. DOI: 10.24963/ijcai.2017/32.

Hejna, Joey and Dorsa Sadigh (May 2023). "Inverse Preference Learning: Preference-based RL without a Reward Function". In: *Advances in Neural Information Processing Systems* 36. arXiv: 2305.15363 tex.arxivid: 2305.15363, arXiv:2305.15363. DOI: 10.48550/arXiv.2305.15363.

Hofstätter, Felix et al. (2023). *Tall Tales at Different Scales: Evaluating Scaling Trends for Deception in Language Models*. URL: https://www.alignmentforum.org/posts/pip63HtEAxHGfSEGk/tall-tales-at-different-scales-evaluating-scaling-trends-for.

Hu, Hengyuan et al. (2020). ""other-play" for zero-shot coordination". In: *International Conference on Machine Learning*. PMLR, pp. 4399–4410.

Huang, Lei et al. (2023). *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. arXiv: 2311.05232. DOI: 10.1145/3703155.

Hubinger, Evan et al. (June 2019). *Risks from Learned Optimization in Advanced Machine Learning Systems*. arXiv: 1906.01820. DOI: 10.48550/arXiv.1906.01820.

Jenner, Erik, Herke van Hoof, and Adam Gleave (2022). *Calculus on MDPs: Potential Shaping as a Gradient*. arXiv: 2208.09570. DOI: 10.48550/arxiv.2208.09570.

Jeon, Hong Jun, Smitha Milli, and Anca Dragan (2020). "Reward-rational (implicit) choice: A unifying formalism for reward learning". In: *Advances in Neural Information Processing Systems*. Ed. by H Larochelle et al. Curran Associates, Inc., pp. 4415–4426. URL: https://dl.acm.org/doi/abs/10.5555/3495724.3496095.

Kagel, John H. and Antonio Penta (2021). "Unraveling in guessing games: An experimental study (by Rosemarie Nagel)". In: *The Art of Experimental Economics: Twenty Top Papers Reviewed* 85.5, pp. 109–118. DOI: 10.4324/9781003019121-10. URL: https://econpapers.repec.org/RePEc:aea:aecrev:v:85:y:1995:i:5:p:1313-26.

Kamien, Morton I, Yair Tauman, and Shmuel Zamir (1990). "On the value of information in a strategic conflict". In: *Games and Economic Behavior* 2.2, pp. 129–153. ISSN: 0899-8256. DOI: https://doi.org/10.1016/0899-8256(90)90026-Q. URL: https://www.sciencedirect.com/science/article/pii/089982569090026Q.

Kausik, Chinmaya et al. (2024). *A Framework for Partially Observed Reward-States in RLHF*. arXiv: 2402.03282. DOI: 10.48550/arxiv.2402.03282.

Kowitz, Gerald T. (1972). "Game theory". In: *Educational Forum*. Vol. 36. Number: 4. MIT Press, pp. 514–514. DOI: 10.1080/00131727209339022.

Kuhn, H. W. (2016). "Extensive Games and the Problem of Information". In: *Contributions to the Theory of Games (AM-28), Volume II*. Ed. by Harold William Kuhn and Albert William Tucker. Princeton: Princeton University Press, pp. 193–216. DOI: 10.1515/9781400881970-012.

Laidlaw, Cassidy and Anca Dragan (2022). "The Boltzmann Policy Distribution: Accounting for Systematic Suboptimality in Human Models". In: *10th International Conference on Learning Representations, ICLR 2022*. URL: https://openreview.net/pdf?id=_l_QjPGN5ye.

Lang, Leon et al. (2024). *When Your AIs Deceive You: Challenges with Partial Observability of Human Evaluators in Reward Learning*. arXiv: 2402.17747. DOI: 10.48550/arxiv.2402.17747.

Lehrer, Ehud, Dinah Rosenberg, and Eran Shmaya (2010). "Signaling and mediation in games with common interests". In: *Games and Economic Behavior* 68.2, pp. 670–682. DOI: 10.1016/j.geb.2009.08.007.

Lin, Stephanie, Jacob Hilton, and Owain Evans (2022). "TruthfulQA: Measuring How Models Mimic Human Falsehoods". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 1, pp. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229.

Majumdar, Anirudha et al. (July 2017). "Risk-sensitive inverse reinforcement learning via coherent risk models". In: *Robotics: Science and Systems*. Ed. by Nancy Amato et al. Vol. 13. United States: MIT Press Journals. DOI: 10.15607/rss.2017.xiii.069.

Mallen, Alex and Nora Belrose (2024). *Balancing Label Quantity and Quality for Scalable Elicitation*. arXiv: 2410.13215 [cs.LG]. DOI: 10.48550/arxiv.2410.13215.

Manyika, James (2023). *An overview of Bard: an early experiment with generative AI*. Pages: 1–9 Publication title: Google. URL: https://ai.google/static/documents/google-about-bard.pdf.

McMillan, B. and D. Slepian (1962). "Information Theory". In: *Proceedings of the IRE*. Vol. 50. Interscience Tracts in Pure and Applied Mathematics. Number: 5. John Wiley & Sons, pp. 1151–1157. DOI: 10.1109/JRPROC.1962.288022.

Mindermann, Sören and Stuart Armstrong (2018). "Occam's razor is insufficient to infer the preferences of irrational agents". In: *Advances in Neural Information Processing Systems*. Vol. 2018-December. NeurIPS'18. Red Hook, NY, USA: Curran Associates Inc., pp. 5598–5609. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Mu, Tong et al. (2024). *Rule Based Rewards for Language Model Safety*. URL: https://openai.com/index/improving-model-safety-behavior-with-rule-based-rewards/.

Ng, Andrew and Stuart Russell (2000). "Algorithms for inverse reinforcement learning". In: *Proceedings of the Seventeenth International Conference on Machine Learning*. Vol. 0. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 663–670. URL: http://www-cs.stanford.edu/people/ang/papers/icml00-irl.pdf.

Oliveira, Henrique de (2018). "Blackwell's informativeness theorem using diagrams". In: *Games and Economic Behavior* 109, pp. 126–131. DOI: 10.1016/j.geb.2017.12.008.

Omohundro, Stephen M. (2007). "The nature of self-improving artificial intelligence". In: *Singularity Summit*. Vol. 5, p. 2007. URL: https://selfawaresystems.com/wp-content/uploads/2008/01/nature_of_self_improving_ai.pdf.

Omohundro, Stephen M. (2008). "The basic AI drives". In: *Frontiers in Artificial Intelligence and Applications*. Vol. 171. Number: 1. NLD: IOS Press, pp. 483–492. DOI: 10.18254/s207751800009748-1.

OpenAI (2008). *Privacy Policy*. URL: https://openai.com/policies/privacy-policy//.

— (2022). *Introducing ChatGPT*. URL: https://openai.com/blog/chatgpt.

— (2023). *ChatGPT plugins*. Publication title: OpenAI Platform. URL: https://chat.openai.com/.

— (2024a). *Model Spec (2024/05/08)*. URL: https://cdn.openai.com/spec/model-spec-2024-05-08.html.

— (2024b). *OpenAI o1 System Card*. Pages: 1–43. URL: https://cdn.openai.com/o1-system-card.pdf.

Park, Chanwoo et al. (2024). *RLHF from Heterogeneous Feedback via Personalization and Preference Aggregation*. arXiv: 2405.00254 [cs.AI]. DOI: 10.48550/arxiv.2405.00254.

Park, Peter S. et al. (2024). "AI deception: A survey of examples, risks, and potential solutions". In: *Patterns* 5.5. DOI: 10.1016/j.patter.2024.100988.

Rafailov, Rafael et al. (2023). *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. arXiv: 2305.18290. DOI: 10.48550/arxiv.2305.18290.

Ramachandran, Deepak and Eyal Amir (2007). "Bayesian inverse reinforcement learning". In: *IJCAI International Joint Conference on Artificial Intelligence*. IJCAI'07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 2586–2591.

Reif, John H (1984). "The complexity of two-player games of incomplete information". In: *Journal of computer and system sciences* 29.2, pp. 274–301.

Russell, Stuart (2019). *Human Compatible: AI and the Problem of Control*. Number: 1. Penguin. URL: https://www.penguin.co.uk/books/307948/human-compatible-by-russell-stuart/9780141987507.

Scheurer, Jérémy, Mikita Balesni, and Marius Hobbhahn (2023). *Large Language Models can Strategically Deceive their Users when Put Under Pressure*. arXiv: 2311.07590. DOI: 10.48550/arxiv.2311.07590.

Shah, Rohin et al. (2020). "Benefits of Assistance over Reward Learning". In: *34th Conference on Neural Information Processing Systems*. URL: https://people.eecs.berkeley.edu/~russell/papers/neurips20ws-assistance.

Siththaranjan, Anand, Cassidy Laidlaw, and Dylan Hadfield-Menell (2024). "Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF". In: *12th International Conference on Learning Representations*. ICLR 2024. DOI: 10.48550/arxiv.2312.08358.

Skalse, Joar and Alessandro Abate (Dec. 2023a). "Misspecification in Inverse Reinforcement Learning". In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023* 37, pp. 15136–15143. DOI: 10.1609/aaai.v37i12.26766.

— (2023b). "On the Limitations of Markovian Rewards to Express Multi-Objective, Risk-Sensitive, and Modal Tasks". In: *Proceedings of Machine Learning Research*. Ed. by Robin J Evans and Ilya Shpitser. Vol. 216. Proceedings of Machine Learning Research. PMLR, pp. 1974–1984. URL: https://proceedings.mlr.press/v216/skalse23a.html.

Skalse, Joar, Matthew Farrugia-Roberts, et al. (2023). "Invariance in Policy Optimisation and Partial Identifiability in Reward Learning". In: *Proceedings of Machine Learning Research*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 32033–32058. URL: https://proceedings.mlr.press/v202/skalse23a.html.

Soares, Nate et al. (2015). "Corrigibility". In: *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*. Ed. by Toby Walsh. AAAI Press. URL: http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10124.

Stahl, Dale O. and Paul W. Wilson (1994). "Experimental evidence on players' models of other players". In: *Journal of Economic Behavior and Organization* 25.3, pp. 309–327. DOI: 10.1016/0167-2681(94)90103-1.

Stray, Jonathan (2023). "The AI Learns to Lie to Please You: Preventing Biased Feedback Loops in Machine-Assisted Intelligence Analysis". In: *Analytics* 2.2, pp. 350–358. DOI: 10.3390/analytics2020020.

Subramani, Rohan et al. (2024). "On the Expressivity of Objective-Specification Formalisms in Reinforcement Learning". In: *12th International Conference on Learning Representations*. ICLR 2024. URL: https://openreview.net/forum?id=qr4ECbGcSj.

Touvron, Hugo et al. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv: 2307.09288. DOI: 10.48550/arxiv.2307.09288.

Treutlein, Johannes et al. (2021). "A New Formalism, Method and Open Issues for Zero-Shot Coordination". In: *Proceedings of Machine Learning Research*. Vol. 139. PMLR, pp. 10413–10423. URL: https://proceedings.mlr.press/v139/treutlein21a/treutlein21a.pdf.

Vinet, Luc and Alexei Zhedanov (2011). "A 'missing' family of classical orthogonal polynomials". In: *Journal of Physics A: Mathematical and Theoretical*. Vol. 44. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 278–287. DOI: 10.1088/1751-8113/44/8/085201.

Wängberg, Tobias et al. (2017). "A Game-Theoretic Analysis of the Off-Switch Game". In: *Artificial General Intelligence*. Ed. by Tom Everitt, Ben Goertzel, and Alexey Potapov. Cham: Springer International Publishing, pp. 167–177. DOI: 10.1007/978-3-319-63703-7_16.

Ward, Francis Rhys et al. (2023). "Honesty Is the Best Policy: Defining and Mitigating AI Deception". In: *Advances in Neural Information Processing Systems*. Vol. 36. URL: https://openreview.net/forum?id=EmxpDiPgRu.

Wen, Jiaxin et al. (2024). *Language Models Learn to Mislead Humans via RLHF*. arXiv: 2409.12822. DOI: 10.48550/arxiv.2409.12822.

Williams, Marcus et al. (2024). *On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback*. arXiv: 2411.02306. DOI: 10.48550/arxiv.2411.02306.

Wu, Scott (2024). *Introducing Devin, the first AI software engineer*. URL: https://www.cognition-labs.com/introducing-devin.

Zhuang, Simon and Dylan Hadfield-Menell (2020). "Consequences of misaligned AI". In: *Advances in Neural Information Processing Systems*. Vol. 2020-December. NeurIPS'20.

Red Hook, NY, USA: Curran Associates Inc. URL: https://proceedings.neurips.cc/paper/2020/file/b607ba543ad05417b8507ee86c54fcb7-Paper.pdf.

Ziebart, Brian D. et al. (2008). "Maximum Entropy Inverse Reinforcement Learning". In: *Proceedings of the 23rd AAAI Conference on Artificial Intelligence, AAAI 2008*. Ed. by Dieter Fox and Carla P Gomes. AAAI Press, pp. 1433–1438. URL: https://dl.acm.org/doi/10.5555/1620270.1620297.

Appendices

Appendix A

Observation Interference in POAGs

A.1 Proofs and example formalizations for Section 2.5

A lemma about policies with internal states

In our proof of Theorem 2.14 (and our proof of Theorem 2.19), we will construct policies that maintain an internal state (the previously sampled garbled observations). We will call this a *virtual state*. However, our setup (in line with the norm in the literature) does not allow for such policies. We here show that any policy with a virtual state can be "simulated" by a policy without virtual states. Since this result is about a single player's policy, holding the opponent policy fixed, we will prove this in POMDPs.

First, a *virtual-state policy* is a family of distributions $\pi(a, \tilde{v} \mid v, h)$, where:

- h is a history of observations and actions as usual;
- v is an agent state from some discrete set (e.g., $\mathbb{N}$ or $\Omega \times \mathcal{A}$);
- $\tilde{v}$ is another (new) virtual state;
- a is an action.

Additionally we specify an initial virtual state v_0. Virtual-state policies give rise to histories in the obvious way: the initial agent state is v_0; the agent then samples an action a_0 and a following virtual state v_1 from $\pi(\cdot \mid v_0)$. In the next step it samples an action and agent state from $\pi(\cdot \mid o_0 a_1, v_1)$ and so on.

We now show that policies with a virtual state can be transformed into behaviorally equivalent policies without an agent state.

Lemma A.1. *Let π be a virtual-state policy. Then there exists a regular policy $\bar{\pi}$ s.t. the resulting distribution over (environment state, observation, action) histories is the same under π and $\bar{\pi}$. In particular, the expected rewards of the two are the same.*

The result is related to Kuhn's (Kuhn, 2016) proof of the equivalence of behavioral and mixed strategies in perfect-recall extensive-form games.

Proof. For this proof we use $h^{o,a}$ to denote observation–action histories and $h^{o,a}$ to use state–observation–action histories. Consider $\bar{\pi}$ that at time step t is defined by

$$\bar{\pi}(A \mid h^{o,a}) = \sum_{v_0,...,v_t} P(v_0, ..., v_t \mid \pi, h^{o,a})\pi(A \mid v_t, h^{o,a}).$$

Intuitively, at time step t we infer a probability distribution over histories of virtual states and in particular v_t, conditioning on the observed observation–action history h, and then sample from the action distribution induced by $\pi(A \mid v_t, h^{o,a})$.

We prove that for each time step t, the state–observation–action history up until time step t is the same between π and $\bar{\pi}$. We prove this by natural induction. The base case is trivial. Assume that the distribution over state–observation–action histories up until time step t is the same. We will show that for each state–observation–action history, the distribution over actions a_{t+1} at time $t+1$ is the same under π and $\bar{\pi}$. Note that the action distribution under π is given by

$$\sum_{v_0^A,...,s_t^A} P(v_0, ..., v_t \mid \pi, h^{s,o,a})\pi(A \mid v_t, h^{o,a}).$$

Now note that $P(v_0, ..., v_t \mid \pi, h^{s,o,a}) = P(v_0, ..., v_t \mid \pi, h^{o,a})$, i.e., given the history of states and observations, the environment states don't provide further evidence about the agent states, since every dependence between environmental states and agent states is mediated by observations and actions. Thus, this distribution is the same as the distribution $\bar{\pi}(A \mid h^{o,a})$. □

Proof of Theorem 2.14

Theorem 2.14. *Let M be any POAG. Let $\mathbf{A}$ have no private information. Then there is an optimal policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ for M in which $\pi^{\mathbf{A}}$ does not interfere with observations at the action level (and $\pi^{\mathbf{H}}$ observes naively).*

Proof sketch. Note first that because our setting is common-payoff and involves no absent-mindedness/imperfect recall, there is always an optimal policy pair in which neither $\mathbf{A}$ nor $\mathbf{H}$ randomizes in any observation history. Let $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ be any optimal policy pair for M. Let $a^{\mathbf{A}}_{\text{interfere}}$ be an interference action played by $\pi^{\mathbf{A}}$. Let $\bar{a}^{\mathbf{A}}$ be the corresponding non-interference strategy. Now consider the policy $\bar{\pi}^{\mathbf{A}}$ that plays like $\pi^{\mathbf{A}}$ except that it plays $\bar{a}^{\mathbf{A}}$ instead of $a^{\mathbf{A}}_{\text{interfere}}$.

We will now construct a corresponding human policy $\bar{\pi}^{\mathbf{H}}$ that results in playing the same actions at each point as $a^{\mathbf{A}}$. Note that by the assumption that $\mathbf{A}$ has no private observations and the fact that $\pi^{\mathbf{A}}$ and $\bar{\pi}^{\mathbf{A}}$ are deterministic, $\mathbf{H}$ always knows $\mathbf{A}$'s full observation history. Thus, $\mathbf{H}$ knows in particular when for which time steps in her observation history $\pi^{\mathbf{A}}$ would have played $a^{\mathbf{A}}_{\text{interfere}}$ and $\bar{\pi}^{\mathbf{A}}$ played $\bar{a}^{\mathbf{A}}$ instead.

Now let F be the observation translation function as per Definition 2.6. Intuitively, we want $\bar{\pi}^{\mathbf{H}}$ to apply F to any new observation that results from playing $\bar{a}^{\mathbf{A}}$ rather than $a^{\mathbf{A}}_{\text{interfere}}$,

and then remember that modified observation in place of the actual observation. It would then be easy to show that $\bar{\pi}^{\mathbf{H}}$ would result in the same actions as $\pi^{\mathbf{H}}$. Together with the fact that $a^{\mathbf{A}}_{\text{interfere}}$ and $\bar{a}^{\mathbf{A}}$ have the same effect on state transitions and rewards, we would immediately obtain that $(\bar{\pi}^{\mathbf{H}}, \bar{\pi}^{\mathbf{A}})$ has the same utility as $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$.

Unfortunately, if F is stochastic, the above construction requires that $\mathbf{H}$ can remember the results of past applications of F. That is, if at time step t she observes according to $\bar{a}^{\mathbf{A}}$ and translates according to F to obtain some new observation $o^{\mathbf{H}}_t$ (that she would have obtained under interference), then at any time step $t' > t$, she needs to remember that she sampled $o^{\mathbf{H}}_t$ from F. Our formalism doesn't allow for such memory. However, by Lemma A.1 we can construct a policy without internal memory to imitate the policy we constructed.

□

Formalization of Example 2.15 and proof of Proposition 2.16

Example 2.15. **H** *has typed* `apt list -a cuda` *to see the list of* `cuda` *versions available to be installed. Out of 10 total versions, only a (non-empty) subset are available. And of these available versions, only a subset are compatible with the other environment software.*

First, **A** *takes an action. For each of the 10 total* `cuda` *versions,* **A** *can choose to or not to suppress it from the list of available packages. This gives* **A** 2^{10} *total actions, where 1 action is non-observation interference (suppressing nothing), and the remaining* $2^{10} - 1$ *actions interfere with observations.*

Second, **H** *takes an action.* **H** *has 10 possible actions which try to install the corresponding version of* `cuda` *if it appears in the version list. If an available* `cuda` *version that is compatible with the other environment software is installed, it yields +1 reward. Otherwise, it yields 0 reward.*

Suppose **A** *sees which versions are compatible with the other software in the environment, but* **H** *doesn't. Then* **A***'s optimal policy is to suppress the versions of* `cuda` *that are incompatible.*

Formalization:

- $\mathcal{S} = (\{0,1\} \times \{0,1\}^{10} \times \{0,1\}^{10}) \cup \{E\} \cup \{I\}$ – E is a terminal state, which we use to make the POAG effectively episodic. I is an initial state. The first bit, which we denote by s_0, encodes the time step. The next ten bits encode which versions are available. The last ten bits encode which versions are compatible. For any state s, we use s_0 to refer to the first entry of the state.
- $\Omega^{\mathbf{H}} = \{0,1\}^{10} \cup \{\text{null}\}$ – representing the availability bits.
- $\Omega^{\mathbf{A}} = \{0,1\}^{10} \cup \{\text{null}\}$ – representing which packages are compatible.
- $\Theta = \{\theta\}$ is a singleton.
- $\mathcal{A}^{\mathbf{H}} = \{1, ..., 10\}$ – representing which package to choose.

- $\mathcal{A}^{\mathbf{A}} = \{0,1\}^{10}$ – representing for what packages, availability is suppressed, where 0 indicates suppression.
- $\mathbf{A}$'s observations are given as follows. If $s \notin \{E, I\}$ and $s_0 = 0$ (i.e., it is the first time step), then $O^{\mathbf{A}}(o^{\mathbf{A}}|s, a^{\mathbf{A}}, a^{\mathbf{H}}) = \mathbb{1}[o^{\mathbf{A}}{=}s_{11:20}]$. That is, $\mathbf{A}$ observes perfectly what cuda versions are compatible. Otherwise, $O^{\mathbf{A}}(o^{\mathbf{A}}|s, a^{\mathbf{A}}, a^{\mathbf{H}}) = \mathbb{1}[o^{\mathbf{A}}{=}\text{null}]$. That is, in all other time steps, $\mathbf{A}$ does not observe anything.
- $\mathbf{H}$'s observations are given as follows. If $s \in \{E, I\}$ or $s_0 \neq 1$, then $\mathbf{H}$ simply observes null. If $s \notin \{E, I\}$ and $s_0 = 1$, then $O^{\mathbf{H}}(o^{\mathbf{H}}|s, a^{\mathbf{A}}, a^{\mathbf{H}}) = \mathbb{1}[o_i^{\mathbf{A}}{=}s_{i+1}a_i^{\mathbf{A}}]$. That is, for each availability bit, $\mathbf{H}$ observes 0 if $\mathbf{A}$ set the availability bit to 0; otherwise, $\mathbf{H}$ simply observes the availability bit.
- $R(s, a^{\mathbf{H}}, a^{\mathbf{A}}) = 0$ if $s \in \{E, I\}$ or $s_0 = 0$. Otherwise, $R(s, a^{\mathbf{H}}, a^{\mathbf{A}}) = s_{a^{\mathbf{H}}} s_{a^{\mathbf{H}}+10}$. That is, a reward of 1 is obtained if and only if the cuda version chosen by $\mathbf{H}$ is both available and compatible.
- $P_0(s) = \mathbb{1}[s = I]$. That is, the initial state is always I.
- If $s = I$, then $T(\cdot \mid s, a^{\mathbf{H}}, a^{\mathbf{A}})$ is the uniform distribution over states s' in which at least one cuda version is available and compatible, i.e., $\sum_{i=1}^{10} s_i s_{i+10} \geq 1$. If $s \neq I$, then $T(s' \mid s, a^{\mathbf{H}}, a^{\mathbf{A}}) = 1$ if
  - $s_0 = 0$, $s'_0 = 1$ and $s_{1:20} = s'_{1:20}$; or
  - $s_0 = 1$ and $s' = E$; or
  - $s = s' = E$.

  Otherwise, $T(s' \mid s, a^{\mathbf{H}}, a^{\mathbf{A}}) = 0$.

Proposition 2.16. *There exists a POAG M where all optimal policy pairs $(\pi^{\mathbf{A}}, \pi^{\mathbf{H}})$ have that $\pi^{\mathbf{A}}$ interferes with observations at the action level and that $\pi^{\mathbf{H}}$ observes and acts naively.*

Proof. Consider Example 2.15.

First consider the following policy pair: At the first time step, $\mathbf{A}$ chooses $o^{\mathbf{A}} \in \{0,1\}^{10}$, i.e., $\mathbf{A}$ chooses to suppress the availability signal exactly for those cuda versions that aren't compatible. At all other time steps the assistant chooses uniformly at random. Call this policy $\hat{\pi}^{\mathbf{A}}$.

At the second time step, when the human observes $o^{\mathbf{H}} \in \{0,1\}^{10}$, the human chooses some $a^{\mathbf{H}}$ s.t. $o^{\mathbf{H}}_{a^{\mathbf{H}}} = 1$. That is, $\mathbf{H}$ chooses a cuda version that her observation shows is available. It is easy to see that under the above $\mathbf{A}$ policy there always exists such a $a^{\mathbf{H}}$. At all other time steps, $\mathbf{H}$ chooses uniformly at random. Call this policy $\hat{\pi}^{\mathbf{H}}$.

It's easy to see that the above policy pair is optimal: By the structure of the environment, we can receive a reward of at most 1 by having the human choose a compatible and available policy at time step 1. Clearly, the above policy achieves this reward of 1.

Next, note that the only non-interference action for $\mathbf{A}$ is $(1, 1, ..., 1)$. Thus, the only non-interference policy for $\mathbf{A}$ is to always play $(1, 1, ..., 1)$. Call this policy $\pi^{\mathbf{A}}_{\text{ni}}$.

Note that the best response for $\mathbf{H}$ against $\pi^{\mathbf{A}}_{\text{ni}}$ is $\hat{\pi}^{\mathbf{H}}$. Thus, $\hat{\pi}^{\mathbf{H}}$ is acting naively.

Furthermore, note that $\hat{\pi}^{\mathbf{H}}$ acts naively.

It is easy to see that adding a $\mathbf{H} \to \mathbf{A}$ communication channel makes no difference to the above analysis. □

Proof of Theorem 2.17

Theorem 2.17. *Let M be any POAG, and provide $\mathbf{A}$ with an unbounded communication channel to $\mathbf{H}$, forming $M^{\mathbf{A}\to\mathbf{H}}$. Then there is an optimal policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ for $M^{\mathbf{A}\to\mathbf{H}}$ where $\pi^{\mathbf{A}}$ does not interfere with observations at the action level (and $\pi^{\mathbf{H}}$ observes naively).*

Proof sketch. Roughly, take any deterministic optimal policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$. Consider the assistant policy $\bar{\pi}^{\mathbf{A}}$ that at each time step communicates $\mathbf{A}$'s full observation to $\mathbf{H}$ and that replaces interference with non-interference actions. Because $\pi^{\mathbf{A}}$ is deterministic, $\mathbf{H}$ can infer what $\pi^{\mathbf{A}}$ would have communicated based on $\bar{\pi}^{\mathbf{A}}$'s communications. The rest of the proof goes the same way as Theorem 2.14. □

Proof of Theorem 2.19

For the proof of Theorem 2.19, we'll use the concept of entropy. For any probability distribution P over some discrete space, let $H(P) \coloneqq -\sum_x P(x) \log P(x)$ denote the distribution's entropy. The following is a well-known result in information theory [e.g., McMillan and Slepian, 1962, Theorem 1.4.5; Cover and Thomas, 2005, Theorem 2.6.5].

Lemma A.2 (Conditioning decreases entropy)**.** *Let X, Y be random variables, then*

$$\mathbb{E}_Y \left[H(P(X \mid Y)) \right] \leq H(P(X)).$$

Further, the inequality is strict if X and Y are not independent, i.e., if $P(X) \neq P(X \mid y)$ for some y, then $\mathbb{E}_Y \left[H(P(X \mid Y)) \right] < H(P(X))$.

Using this result, we can provide the following variant.

Lemma A.3. *Let S be a random variable. Let X, Y be independent samples from $F(S)$ and let Z be sampled from $G(Y)$, where F and G are stochastic functions. Then*

$$\mathbb{E}_Z \left[H(P(S \mid Z)) \right] \geq \mathbb{E}_X \left[H(P(S \mid X)) \right].$$

Moreover, the inequality is strict if S and Y are dependent given Z.

Proof. For the non-strict version:

$$\begin{aligned} H(P(S \mid X)) &= H(P(S \mid Y)) \\ &= H(P(S \mid Y, Z)) \\ &\underset{\text{Lemma } A.2}{\leq} H(P(S \mid Z)) \end{aligned}$$

The strict version can be proved the same way using the strict version of Lemma A.2. □

Next, we can use this to prove that a garbling induces a lower-entropy distribution over states.

Lemma A.4. *Let L be some set of states. Let $(P_a(\cdot \mid s))_{s \in L}$ and $(P_b(\cdot \mid s))_{s \in L}$ be families of probability distributions s.t. P_a is strictly more informative than P_b with transformation function F. Further let S be some random variable over L with full support. Let $X_a \sim P_a(\cdot \mid S)$ and $X_b \sim F(X_a)$. Then S and X_a are dependent given X_b. In particular, from Lemma A.3 we get that $\mathbb{E}_X\left[H(P(S \mid X))\right] < \mathbb{E}_{\hat{X}}\left[H(P(S \mid \hat{X}))\right]$.*

Proof. We prove the following contrapositive: if X_a and S are independent given X_b, then P_b is at least as informative as P_a. If X_a and S are independent given X_b, then we have that $P(X_b \mid X_a, S) = P(X_b \mid X_a)$. Thus, for all states s, we have that

$$\begin{aligned} P(X_a \mid s) &= \sum_{x_b} P(x_b \mid s) P(X_a \mid x_b, s) \\ &= \sum_{x_b} P(x_b \mid s) P(X_a \mid x_b). \end{aligned}$$

But this means that if we sample X_b according to P_b, and sample X_a according to $P(X_a \mid x_b)$, then we obtain a sample for X_a according to the distribution $P(X_a \mid s)$ (i.e., P_a). Thus, we have that P_b is at least as informative as P_a. □

Theorem 2.19. *Let M be any POAG. Then there exists an optimal policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ for M s.t. $\pi^{\mathbf{A}}$ does not interfere with observations at the policy level.*

Proof. We will explicitly choose a policy for each time step $t = 0, 1, 2, \ldots$. So let's take $\pi^{\mathbf{A}}_{0:t-1}, \pi^{\mathbf{H}}_{0:t-1}$ as given. Now let Π_t be the set of policies at time t that are part of a policy pair $(\pi^{\mathbf{H}}_{t:}, \pi^{\mathbf{A}}_{t:})$ that is optimal holding fixed $\pi^{\mathbf{A}}_{0:t-1}, \pi^{\mathbf{H}}_{0:t-1}$. Note that the expected utility of policy pairs in a POMDP is continuous. It follows that Π_t is closed (i.e., that every convergent sequence of policies in Π_t converges to a policy in Π_t).

Now from Π_t choose $\bar{\pi}^{\mathbf{A}}_t$ as the minimizer of

$$\pi^{\mathbf{A}}_t \mapsto \mathbb{E}_{O^{\mathbf{H}}_{t+1}}\left[H(P(S_{t+1} \mid O^{\mathbf{H}}_{t+1}, \pi^{\mathbf{H}}_{\text{random}}, \pi^{\mathbf{A}}_{0:t-1}, \pi^{\mathbf{A}}_t)) \mid \pi^{\mathbf{H}}_{\text{random}}, \pi^{\mathbf{A}}_{0:t-1}, \pi^{\mathbf{A}}_t\right],$$

where H denotes Shannon entropy and $\pi^{\mathbf{H}}_{\text{random}}$ is the human strategy that chooses uniformly at random. (Note that the above entropy function is not the only function we could use

for this proof.) That is, let $\pi_t^{\mathbf{A}}$ be the policy that minimizes the entropy of $\mathbf{H}$'s probability distribution over world state. Because the given function is continuous and Π_t is closed (and bounded), this minimum exists (by the extreme value theorem).

Now by Lemma A.4 we have that if $\pi_t^{\mathbf{A}}$ is more informative than $\hat{\pi}_t^{\mathbf{A}}$, then $\pi_t^{\mathbf{A}}$ will also have lower entropy at time t. It follows that there is no policy in Π_t that is more informative than $\bar{\pi}_t^{\mathbf{A}}$.

Finally, it is left to show that there is no policy π_t outside of Π_t that is more informative than π_t^*. For this, we use the same argument as in the proof of Theorem 2.14: if there were a more informative $\tilde{\pi}_t^{\mathbf{A}}$ with the same effect on state transitions, then this would also be part of an optimal policy pair (constructed by having $\mathbf{H}$ apply the appropriate garbling internally). But we have already that in Π_t there is no more informative policy than $\bar{\pi}_t^{\mathbf{A}}$. □

Note that the entropy-minimizing policy used in the proof may still interfere with observations at the *action* level. For example, by default $\mathbf{H}$ might receive a low-information signal about the world. The entropy-minimizing policy might be one in which $\mathbf{A}$ overwrites this default signal in a way that expresses more information about the world. For instance, let's assume that by default, $\mathbf{H}$ observes a random number between -20 and 0 if it's cold outside and a random number between 0 and $+40$ if it's warm outside. $\mathbf{A}$ receives various hints about the temperature and can overwrite the signal with an arbitrary number. (I.e., for each number between -20 and $+40$, there's an action that sets $\mathbf{H}$'s observation to be that number.) Assuming nothing else happens in this POAG, the entropy-minimizing policies will be ones that overwrite the signal in a way that encodes $\mathbf{A}$'s information about the temperature. For instance, $\mathbf{A}$ it may (or may not) be an non-interfering-at-the-policy-level strategy for $\mathbf{A}$ to overwrite $\mathbf{H}$'s signal with $\mathbf{A}$'s expectation of the temperature in degrees Celsius. Given such a policy, the entropy of $\mathbf{H}$'s beliefs about the world is lower than before ($\mathbf{H}$ has more information about the temperature). But each of these overwriting actions individually is observation-interfering.

A.2 Formalization of Example 2.20 and proof of Proposition 2.21

Recall the example:

Example 2.20. $\mathbf{H}$ *would like to schedule a job on a cluster. She can choose between two nodes. By default, she receives a signal from the environment about the two nodes' specifications. Each node may be either GPU-optimized or CPU-optimized. Also, the CPUs may be either AMD or Intel.*

$\mathbf{H}$ *has a strong preference between GPU-optimized and CPU-optimized nodes. She has a weak preference between AMD and Intel. These preferences are unknown to* $\mathbf{A}$.

$\mathbf{A}$ *can interfere with* $\mathbf{H}$*'s observation about the available nodes. In particular,* $\mathbf{A}$ *can make it so that a choice between two CPU-optimized nodes appears as a choice between a*

GPU-optimized and CPU-optimized node. **A** *observes* **H***'s choice. Later,* **A** *is charged with scheduling a job for* **H** *and has to choose between a CPU- and a GPU-optimized node on* **H***'s behalf.*

If **H** *chooses naively upon seeing only CPU-optimized nodes (simply choosing her favorite), then* **A***'s best response interferes with observations at both the action and policy levels. Interfering with observations allows* **A** *to learn* **H***'s preference about GPU- vs CPU-optimized nodes.*

In particular, there are four possible states: (1) The first node is GPU-optimized and the second node is CPU-optimized. (2) The first node is CPU-optimized and the second node is GPU-optimized. (3) Both nodes are CPU-optimized. The first has an Intel processor, the second has an AMD processor. (4) Both nodes are CPU-optimized. The first has an AMD processor and the second has an Intel processor.

Suppose the utilities of the human choice are given as follows: 1 for the favored CPU-optimized type; 1 for a GPU-optimized node if **H** favors the GPU-optimized node. The reward is 0 otherwise. On the second step, the reward for the favored type of node is 10 and 0 for the other type of node.

Recall the proposition was as follows.

Proposition 2.21. *There is a POAG M with the following properties. For every optimal policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$, at least one of these holds:*

i) $\pi^{\mathbf{H}}$ is not acting naively, or

ii) $\pi^{\mathbf{A}}$ interferes with observations at both the action and policy levels.

Additionally, there exists an optimal policy pair $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ where $\pi^{\mathbf{H}}$ acts naively and $\pi^{\mathbf{A}}$ interferes with observations at both the action and policy levels.

These properties continue to hold if we require that in M, **A** *has no private information or can arbitrarily send messages to* **H** *(i.e., there is a POAG $\tilde{M}$ s.t. $M = \tilde{M}^{\mathbf{A}\to\mathbf{H}}$).*

Proof sketch. Consider the example. First let's consider a naive human policy, i.e., one that chooses the favorite node type in the first time step. Then the best response for **A** is to interfere.

It is easy to see that in all optimal policy pairs, **A** must learn about **H**'s GPU-versus-CPU preference. It follows that at time step 1, **H** must deterministically choose depending on her GPU-versus-CPU preference.

It is easy to see that all of these policy profiles have the same expected reward as the above naive/interference policy pair.

Note that in the above example, **A** has no private information. It is easy to see that the above argument continues to go through if we allow **A** to send signals to **H**. □

A.3 Formalization of Example 2.24 and proof of Proposition 2.25

Definition 2.23. *Let M be a POAG. Let $\pi^{\mathbf{A}}$ be $\mathbf{A}$'s policy in M. We say that $\mathbf{H}$'s policy $\pi^{\mathbf{H}}$ is a* Boltzmann-rational response to $\pi^{\mathbf{A}}$ *if there exists some $\beta > 0$ s.t. for every human observation history h that arises with positive probability in M under $(\pi^{\mathbf{A}}, \pi^{\mathbf{H}})$ we have that $\pi^{\mathbf{H}}(a \mid h) \propto \exp\left(\beta \mathbb{E}\left[\sum_{t'=t}^{\infty} \gamma^{t'} R(S_t, A_t^{\mathbf{A}}, A_t^{\mathbf{H}}) \mid \pi^{\mathbf{H}}, \pi^{\mathbf{A}}, h\right]\right)$.*

Example 2.24. **H** *is running a terminal command and is unsure whether to run the command with flag 1 or flag 2. With equal probability, either flag 1 or flag 2 is better, and how good the flags are differs by either a little or a lot. Thus,* **H** *is uniformly at random in one of four states.* **A** *has two actions:* `man` *and* `tldr`*. The* `man` *page is a long document that tells the human exactly what the values of the flags are (ie, exactly what state the human is in). The* `tldr` *page is a short summary that tells the human which flag is better, but not by how much (ie, ruling out half the states, leaving half remaining).*

With uniform probability, $\mathbf{H}$ is in one of four possible states:

- Flag 1 is better by a lot: flag 1 has value +7, while flag 2 has value 0.
- Flag 1 is better by a little: flag 1 has value +1, while flag 2 has value 0.
- Flag 2 is better by a little: flag 1 has value 0, while flag 2 has value +1.
- Flag 2 is better by a lot: flag 1 has value 0, while flag 2 has value +7.

This gives us the following formalization for the game:

- $\mathcal{S} = (\{0, 1\} \times \{s_a, s_b, s_c, s_d\}) \cup \{I, E\}$
- $\Omega^{\mathbf{H}} = \mathcal{S} \cup \{1, 2\} \cup \{\text{null}\}$
- $\Omega^{\mathbf{A}} = \Omega^{\mathbf{H}}$
- Θ is a singleton
- $\mathcal{A}^{\mathbf{H}} = \{1, 2\}$
- $\mathcal{A}^{\mathbf{A}} = \{\texttt{tldr}, \texttt{man}\}$
- $\mathbf{H}$'s observations are given as follows. For $s \in \{s_a, s_b, s_c, s_d\}$, we have $O^{\mathbf{H}}(o^{\mathbf{H}} \mid (0, s), \texttt{man}, a^{\mathbf{H}}) = \mathbb{1}[o^{\mathbf{H}} = s]$, and for $i \in \{1, 2\}$ we have $O^{\mathbf{H}}(i \mid (0, s), \texttt{tldr}, a^{\mathbf{H}}) = \mathbb{1}[i = 1]\mathbb{1}[s \in \{s_a, s_b\}] + \mathbb{1}[i = 2]\mathbb{1}[s \in \{s_c, s_d\}]$. Otherwise, $\mathbf{H}$'s observation is deterministically null.
- $\mathbf{A}$'s observations are the same as $\mathbf{H}$'s observations.

- The reward is given as follows:

$$R((1, s_a), 1, a^{\mathbf{A}}) = 7 \tag{A.1}$$
$$R((1, s_b), 1, a^{\mathbf{A}}) = 1 \tag{A.2}$$
$$R((1, s_c), 2, a^{\mathbf{A}}) = 7 \tag{A.3}$$
$$R((1, s_d), 2, a^{\mathbf{A}}) = 1 \tag{A.4}$$

  $\mathbb{1}[a^{\mathbf{H}} = 1]\mathbb{1}[s \in \{s_a, s_b\}] + \mathbb{1}[a^{\mathbf{H}} = 2]\mathbb{1}[s \in \{s_c, s_d\}]$. All other rewards are 0.

- For all $a^{\mathbf{H}}$, $a^{\mathbf{A}}$, $T(\cdot \mid I, a^{\mathbf{H}}, a^{\mathbf{A}})$ is the uniform distribution over $\{0\} \times \{s_a, s_b, s_c, s_d\}$. For all $s \in \{s_a, s_b, s_c, s_d\}$, $T(s' \mid (0, s), a^{\mathbf{H}}, a^{\mathbf{A}}) = \mathbb{1}[s' = (1, s)]$. For all s, $T(s' \mid (1, s), a^{\mathbf{H}}, a^{\mathbf{A}}) = \mathbb{1}[s' = E]$. Finally, $T(s' \mid E, a^{\mathbf{H}}, a^{\mathbf{A}}) = \mathbb{1}[s' = E]$.

Proposition 2.25. *For every $\beta > 0$, $\exists$ a POAG in which neither $\mathbf{H}$ nor $\mathbf{A}$ has private information s.t. all β-Boltzmann-rational/optimal policy pairs $(\pi^{\mathbf{H}}, \pi^{\mathbf{A}})$ have $\pi^{\mathbf{A}}$ interfere with observations at both the action and policy levels.*

Proof. Note first that multiplying β by any positive number has the same effect on Boltzmann-rational strategies as multiplying all rewards by that number. Therefore, we can consider $\beta = 1$ without loss of generality.

Consider Example 2.24. Note that `tldr` is an observation interference action – `man` results in a more informative signal to $\mathbf{H}$.

Now consider the non-interference policy for $\mathbf{A}$ that always plays `man`. Then a Boltzmann-rational $\mathbf{H}$ will choose as follows: If she observes s_a or s_c, then she will choose an expected utility of 7 with probability $\propto \exp(7)$ and an expected utility of 0 with probability $\propto \exp(0)$. Thus, the expected utility is

$$7\frac{\exp(7)}{\exp(7) + \exp(0)} \tag{A.5}$$

Similarly, if she observes s_b or s_c, her expected utility is

$$\frac{\exp(1)}{\exp(1) + \exp(0)}. \tag{A.6}$$

Thus, overall her expected utility is

$$\frac{1}{2}7\frac{\exp(7)}{\exp(7) + \exp(0)} + \frac{1}{2}\frac{\exp(1)}{\exp(1) + \exp(0)} \approx 3.86234. \tag{A.7}$$

Now consider the interference policy for $\mathbf{A}$ in which $\mathbf{A}$ always plays `tldr`. Then upon observing either 0 or 1, the human chooses between a utility of 0 and a utility of 4. Thus, the expected utility is

$$4 \cdot \frac{\exp(4)}{\exp(4) + \exp(0)} \approx 3.92806. \tag{A.8}$$

We observe that this expected value under interference is higher than the expected value under non-interference. □

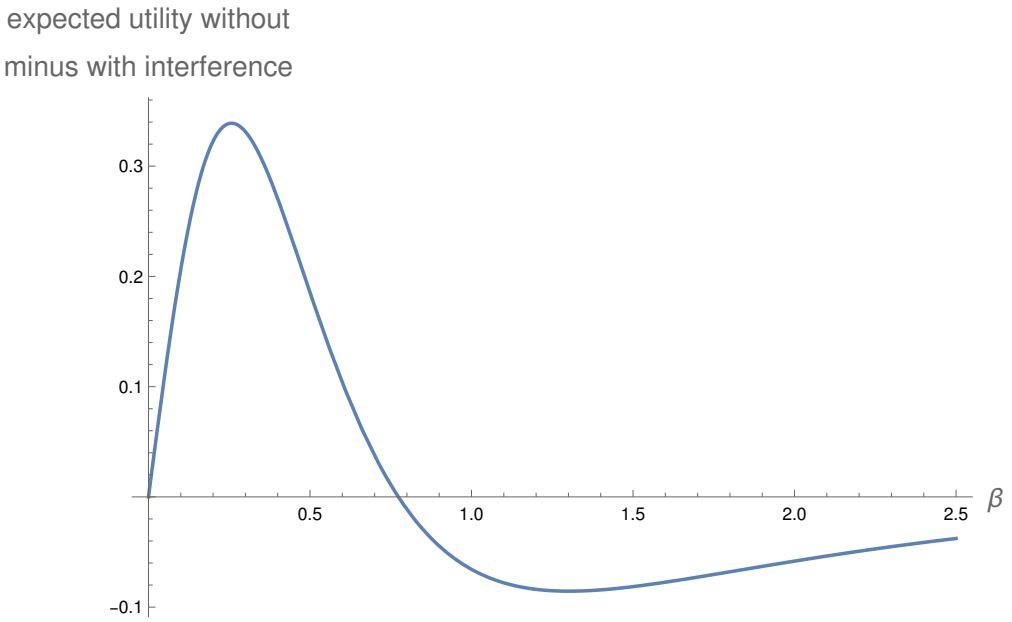


Figure A.1: The effect of varying β on the assistant's incentive for observation interference in Example 2.24. Specifically, the y axis indicates the difference between the expected utility under non-interference minus the expected utility under interference.

A.4 Effects of varying the Boltzmann rationality parameter (β) on the assistant's incentives to interfere with observations

As noted in the main text, in Example 2.24, we have that for low values of the rationality parameter β, **A** prefers non-interference, while for large values of β, **A** prefers interference. Below we will show that in general, counterintuitively, **A** prefers non-interference for sufficiently small (positive) values of β.

We here only consider the case of a single decision. Consider a case with n actions. Let the expected utilities of the different actions without information be $y_{0,1}, ..., y_{0,n}$. Now imagine that **H** might receive k different signals with probabilities $p_1, ..., p_k$. Under signal $i \in \{1, ..., k\}$, the expected utilities of the different actions become $y_{i,1}, ..., y^{i,n}$. By the tower rule we must have for each action $a \in \{1, ..., n\}$,

$$\sum_{i=1}^{k} p_i y_{i,a} = y_{0,a}. \tag{A.9}$$

Note that without further restriction, the above setting includes settings in which the signal provides information on what action is best.

For any β, the expected utility without the signal is

$$\frac{1}{\sum_{a=1}^{n} \exp(\beta y_{0,a})} \sum_{a=1}^{n} \exp(\beta y_{0,a}) y_{0,a}. \tag{A.10}$$

The expected utility *with* the signal is

$$\sum_{s=1}^{k} p_s \frac{1}{\sum_{a=1}^{n} \exp(\beta y_{s,a})} \sum_{a=1}^{n} \exp(\beta y_{s,a}) y_{s,a}. \tag{A.11}$$

Proposition A.5. *For all* $(y_{s,a} \in \mathbb{R})_{s\in\{0,1,\ldots,k\},a\in\{1,\ldots,n\}}, (p_s \in \mathbb{R})_{s\in\{0,1,\ldots,k\}}$ *satisfying Equation* (A.9)*, we have that for sufficiently small but positive* β*, the expected utility without the signal is at most the expected utility with the signal.*

Proof. It's easy to see that for $\beta = 0$, the two expected utilities are the same. Thus, all we need to show is that the derivative w.r.t. β of the term in Eq. A.11 at $\beta = 0$ exceeds the corresponding derivative of the term in Eq. A.10.

The derivative w.r.t. β at $\beta = 0$ of the term in Equation (A.10) is

$$\left(\sum_{a=1}^{n} \frac{1}{n} y_{0,a}^2\right) - \left(\sum_{a=1}^{n} \frac{1}{n} y_{0,a}\right)^2. \tag{A.12}$$

Note that this is exactly the variance of a random variable that is uniform over $(y_{0,a})_{a=1,\ldots,n}$.

Similarly, the derivative of the term in Equation (A.11) is

$$\sum_{s=1}^{k} p_s \left(\left(\sum_{a=1}^{n} \frac{1}{n} y_{s,a}^2\right) - \left(\sum_{a=1}^{n} \frac{1}{n} y_{s,a}\right)^2\right). \tag{A.13}$$

Note that this is the weighted average (over s) of the uniform random variables over $(y_{s,a})_{a=1,\ldots,n}$.

We can now prove the claimed inequality using the convexity of the square function, Equation (A.9) and some basic term manipulation.

$$\left(\sum_{a=1}^{n}\frac{1}{n}y_{0,a}^2\right)-\left(\sum_{a=1}^{n}\frac{1}{n}y_{0,a}\right)^2 \tag{A.14}$$

$$= \sum_{a=1}^{n}\frac{1}{n}\left(y_{0,a}^2-\frac{1}{n}\left(\sum_{a'=1}^{n}y_{0,a'}\right)^2\right) \tag{A.15}$$

$$= \sum_{a=1}^{n}\frac{1}{n}\left(y_{0,a}-\frac{1}{n}\sum_{a'=1}^{n}y_{0,a'}\right)^2 \tag{A.16}$$

$$\underset{\text{Equation (A.9)}}{=} \sum_{a=1}^{n}\frac{1}{n}\left(\left(\sum_{s=1}^{k}p_s y_{s,a}\right)-\frac{1}{n}\sum_{a'=1}^{n}\sum_{s=1}^{k}p_s y_{s,a'}\right)^2 \tag{A.17}$$

$$= \sum_{a=1}^{n}\frac{1}{n}\left(\sum_{s=1}^{k}p_s\left(y_{s,a}-\frac{1}{n}\sum_{a'=1}^{n}y_{s,a'}\right)\right)^2 \tag{A.18}$$

$$\underset{(\cdot)^2\text{ is convex}}{\leq} \sum_{a=1}^{n}\frac{1}{n}\sum_{s=1}^{k}p_s\left(y_{s,a}-\frac{1}{n}\sum_{a'=1}^{n}y_{s,a'}\right)^2 \tag{A.19}$$

$$= \sum_{s=1}^{k}p_s\sum_{a=1}^{n}\frac{1}{n}\left(y_{s,a}-\frac{1}{n}\sum_{a'=1}^{n}y_{s,a'}\right)^2 \tag{A.20}$$

$$= \sum_{s=1}^{k}p_s\left(\sum_{a=1}^{n}\frac{1}{n}y_{s,a}^2-\left(\sum_{a=1}^{n}\frac{1}{n}y_{s,a}\right)^2\right). \tag{A.21}$$

We have skipped over some term manipulations in Equations (A.16) and (A.21), both of which are essentially the equality of two definitions of the variance: $\mathrm{Var}(X)=(X-[X])^2$ and $\mathrm{Var}(X)=\mathbb{E}[X^2]-\mathbb{E}[X]^2$. □

It's interesting to note that this is essentially the proof that the variance (over a) of the expectation (over s) is at least the expectation (over s) of the variance (over a).

Second, we want to show that for large β, **A** prefers observation interference, i.e., prefers to have the human choose based on the expected utilities $y_{0,1},...,y_{0,n}$ rather than the expected utilities that arise from further signals. However, for this to hold we need a further condition. Note that in the general formalism above, the signal s may provide information about which action is best. If this is the case, then it is easy to show that for large enough β, **A** will prefer providing the signal. However, consider specifically those cases in which the signal s only provides information about how much better the best action is compared to other actions. Therefore, we require in the following result that the best action is the same (WLOG 1) across s.

Proposition A.6. *Let $(y_{s,a} \in \mathbb{R})_{s\in S, a\in\{1,...,n\}}, (p_s \in \mathbb{R})_{s\in S}$ satisfy Equation* (A.9) *and let $y_{s,0} > y_{s,a}$ for all $s \in \{0\} \cup S, a \in \{1, ..., n\}$. Then for all sufficiently large β we have that the expected utility without the signal is at most the expected utility with the signal. The inequality is strict if the signal is non-trivial (i.e., $y_{s,a}$ is not constant across s for some a).*

We first provide a very rough sketch. For simplicity, let's say that the signal provides evidence about how much better the first action is compared to the second-best action. Then sometimes the signal will *decrease* the difference in expected utility between the best and second-best utility. We will show that as $\beta \to \infty$, the overall effect of learning the information is dominated by taking the best action *less* in this case.

We will use the following lemmas.

Lemma A.7. *Let the differences between the top k actions be constant across signals and let the difference to the $k+1$-th action be non-constant. Then there is a signal $\tilde{s}$ s.t. the difference to the $k+1$-th action decreases under that signal.*

Proof. Let $k-1$ be the k-th best action according to 0 and let k be the $k+1$-th best action according to 0. By the tower rule (Eq. A.9), $y_{0,k-1} - y_{0,k}$ must be greater than $y_{s,k-1} - y_{s,k}$ for some s. (If the difference in these expected utilities changes when the signal is observed, then it must sometimes decrease.) But then in cases where this difference decreases as s is observed, we clearly have that the difference between one of the k best actions to the $k+1$-th best action under s also decreases. □

Proof of Proposition A.6. The gain from obtaining the signal is:

$$\sum_s p_s \sum_a \left(\frac{\exp(\beta y_{s,a})}{\sum_{a'} \exp(\beta y_{s,a'})} - \frac{\exp(\beta y_{0,a})}{\sum_{a'} \exp(\beta y_{0,a'})} \right) y_{s,a}.$$

WLOG let 0 be the best action under all signals, 1 the second-best and so on. Let k be the largest number that the differences between the utilities of actions $0, ..., k-1$ are always the same. (Typically $k = 0$.) Let $\tilde{S}$ be the set of signals under which the difference to the utility of k (the $k+1$-th best action) is minimized. Note that in particular, the difference must be smaller than under 0 by Lemma A.7. WLOG assume that for all signals, k is among the $k+1$-th best actions.
WLOG assume that $y_{s,a} > 0$ for all $s \in \{0\} \cup S$ and all a and that $y_{s,0}$ is constant across s. Now we will divide up the above sum into three components:

A The change (decrease) in utility from playing the top k actions less in $\tilde{S}$ than without the signal.

$$A \coloneqq \sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \sum_{a=0,...,k-1} \left(\frac{\exp(\beta y_{\tilde{s},a})}{\sum_{a'} \exp(\beta y_{\tilde{s},a'})} - \frac{\exp(\beta y_{0,a})}{\sum_{a'} \exp(\beta y_{0,a'})} \right) y_{\tilde{s},a}$$

B The change in utility from the changes in distribution of all actions other than the top k under $\tilde{S}$ versus S

$$B \coloneqq \sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \sum_{a=k,k+1,\ldots} \left(\frac{\exp(\beta y_{\tilde{s},a})}{\sum_{a'} \exp(\beta y_{\tilde{s},a'})} - \frac{\exp(\beta y_{0,a})}{\sum_{a'} \exp(\beta y_{0,a'})} \right) y_{\tilde{s},a}$$

C The change in utility from all signals other than $\tilde{S}$, i.e.

$$\sum_{s\notin\tilde{S}} p_s \sum_{a} \left(\frac{\exp(\beta y_{s,a})}{\sum_{a'} \exp(\beta y_{s,a'})} - \frac{\exp(\beta y_{0,a})}{\sum_{a'} \exp(\beta y_{0,a'})} \right) y_{s,a}.$$

We will show that the effect from A (which is negative) is becomes infinitely much larger than the effect from B and C (in absolute terms). From that it will follow that the original sum, which is equal to $A + B + C$ is negative as $\beta \to \infty$.

We first provide a bound on A. We first show that $A < 0$. To show this, note first that in all enumerators in A, we can replace $y_{\tilde{s},a}$ with $y_{0,a}$ (by choice of $\tilde{s}$ and k). So all we need to show is that the second denominator is smaller than the first, i.e., $\sum_{a'} \exp(\beta y_{\tilde{s},a'}) > \sum_{a'} \exp(\beta y_{0,a'})$. But this this is easy to see from the fact that $y_{\tilde{s},a} = y_{0,a}$ for $a = 0, 1..., k-1$ and $y_{\tilde{s},k} > y_{0,k}$. For large β, $\exp(\beta y_{\tilde{s},k})$ will be much larger than $\sum_{a'=k,k+1,\ldots} \exp(\beta y_{0,a'})$.

Next, we will provide a lower bound on the absolute value of $|A|$.

$$\begin{aligned}
A &= \sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \sum_{a=0,\ldots,k-1} \left(\frac{\exp(\beta y_{\tilde{s},a})}{\sum_{a'} \exp(\beta y_{\tilde{s},a'})} - \frac{\exp(\beta y_{0,a})}{\sum_{a'} \exp(\beta y_{0,a'})} \right) y_{\tilde{s},a} \\
&= \sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \sum_{a=0,\ldots,k-1} \exp(\beta y_{0,a}) \left(\frac{1}{\sum_{a'} \exp(\beta y_{\tilde{s},a'})} - \frac{1}{\sum_{a'} \exp(\beta y_{0,a'})} \right) y_{0,a} \\
&= \sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \sum_{a=0,\ldots,k-1} \exp(\beta y_{0,a}) \left(\frac{1}{\sum_{a'} \exp(\beta y_{\tilde{s},a'})} - \frac{1}{\sum_{a'} \exp(\beta y_{0,a'})} \right) y_{0,a} \\
&= \sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \sum_{a=0,\ldots,k-1} \exp(\beta y_{0,a}) \frac{(\sum_{a'} \exp(\beta y_{0,a'})) - \sum_{a'} \exp(\beta y_{\tilde{s},a'})}{(\sum_{a'} \exp(\beta y_{\tilde{s},a'}))\,(\sum_{a'} \exp(\beta y_{0,a'}))} y_{0,a} \\
&= \sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \sum_{a=0,\ldots,k-1} \exp(\beta y_{0,a}) \frac{\left(\sum_{a=k,k+1,\ldots} \exp(\beta y_{0,a'})\right) - \sum_{a=k,k+1,\ldots} \exp(\beta y_{\tilde{s},a'})}{(\sum_{a'} \exp(\beta y_{\tilde{s},a'}))\,(\sum_{a'} \exp(\beta y_{0,a'}))} y_{0,a} \\
&\leq \sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \sum_{a=0,\ldots,k-1} \exp(\beta y_{0,a}) \frac{n \exp(\beta y_{0,k}) - \exp(\beta y_{\tilde{s},k})}{n^2 \exp(\beta y_{0,a})^2} y_{0,a}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \sum_{a=0,\ldots,k-1} \frac{n\exp(\beta y_{0,k}) - \exp(\beta y_{\tilde{s},k})}{n^2\exp(\beta y_{0,a})} y_{0,a} \\
&\leq -\frac{1}{2}\sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \sum_{a=0,\ldots,k-1} \frac{\exp(\beta y_{\tilde{s},k})}{n^2\exp(\beta y_{0,a})} y_{0,a} \\
&\leq -\frac{1}{2}\sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \frac{\exp(\beta y_{\tilde{s},k})}{n^2\exp(\beta y_{0,0})} y_{0,0}
\end{aligned}
$$

Next we upper bound B. First, the best case for the effect on ... is that all the probability mass that under 0 is on the top k actions ends up on the k-th best action, i.e.,

$$
B \leq \sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \left(1 - \sum_{a=0,\ldots,k-1} \frac{\exp(\beta y_{0,a})}{\sum_{a'}\exp(\beta y_{0,a'})}\right) y_{\tilde{s},k}.
$$

We can further upper-bound this as follows:

$$
\begin{aligned}
&\sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \left(1 - \sum_{a=0,\ldots,k-1} \frac{\exp(\beta y_{0,a})}{\sum_{a'}\exp(\beta y_{0,a'})}\right) y_{\tilde{s},k} \\
&= \sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \sum_{a=k,\ldots} \frac{\exp(\beta y_{0,a})}{\sum_{a'}\exp(\beta y_{0,a'})} y_{\tilde{s},k} \\
&\leq \sum_{\tilde{s}\in\tilde{S}} p_{\tilde{s}} \frac{n\exp(\beta y_{0,k})}{\exp(\beta y_{0,0})} y_{\tilde{s},k}
\end{aligned}
$$

From the fact that $y_{\tilde{s},k} > y_{0,k}$, it is easy to see that this term vanishes in absolute value relative to our upper bound on A.

Finally, we must upper bound C. First, we can upper bound C by considering a case where all probability mass that in 0 was outside the top k actions, goes to the best action when a signal outside of $\tilde{S}$ is observed, i.e.,

$$
C \leq \sum_{s\notin\tilde{S}} p_s \sum_{a=k,k+1,\ldots} \frac{\exp(\beta y_{0,a})}{\sum_{a'}\exp(\beta y_{0,a'})} y_{0,0}.
$$

We can further upper bound this as follows:

$$
\begin{aligned}
&\sum_{s\notin\tilde{S}} p_s \sum_{a=k,k+1,\ldots} \frac{\exp(\beta y_{0,a})}{\sum_{a'}\exp(\beta y_{0,a'})} y_{0,0} \\
&\leq \sum_{s\notin\tilde{S}} p_s \frac{n\exp(\beta y_{0,k})}{\exp(\beta y_{0,0})} y_{0,0}
\end{aligned}
$$

Again, from the fact that $y_{\tilde{s},k} > y_{0,k}$, it is easy to see that this term vanishes in absolute value relative to our upper bound on A. □

A.5 Proof of A's best response in the product selection game

Proposition 2.27. *Consider* **A** *policies that always interfere with k observations for some fixed k. Among the k-interference policies for a given k,* **A***'s best response to* **H***'s straightforward product selection policy is as follows.* **A** *interferes with the k smallest $\hat{R}_i$ values where $\hat{R}_i = R_i$ if* **A** *observes R_i, and $\hat{R}_i = 0.5$ otherwise.*

Proof. Consider **A**'s perspective. **A**'s interference is equivalent to selecting a set of $d - k$ untampered products from which **H** selects according to a Boltzmann distribution on H_i. As **A** neither sees nor affects the H_i, by symmetry, over all draws of the game, **H** selects each of the $d - k$ products with equal probability. **A**'s expected payoff for choosing $d - k$ products, then, is the uniform average of the products' expected U_i.

How does **A** choose the set of $d - k$ products to maximize the uniform average of the products' expected U_i? Recall $U_i = H_i + R_i$. As **A** neither sees nor affects the H_i, **A** can ignore the H_i and consider only the R_i. Denote the expected R_i by $\hat{R}_i = \mathbb{E}[R_i]$. If **A** observes R_i, then $\hat{R}_i = R_i$. If **A** doesn't observe R_i, then $\hat{R}_i = 0.5$. To choose the *maximum* $d - k$ values for $\hat{R}_i$, **A** interferes with the *minimum* k values of $\hat{R}_i$. □

A.6 Minor deficiencies of the observation interference definition

As noted in the main text, there are various possible concerns with Definition 2.7 that we consider minor because they do not change the main ideas and results of our work.

- The definition does not take into account what **A** knows about what **H** already knows. As such, it will sometimes spuriously judge a policy to be observation interference for taking away a signal from the human that is redundant with the human's past observations. For example, if the human observes the Linux version at time t and the Linux is known not to change, then preventing the human from observing the Linux version again at time $t + 1$ might count as observation interference.

  The definition may also spuriously judge a policy to *not* be observation interference because the only more informative policies fail to provide some redundant piece of information to the human. For instance, let's say that by default the human learns some new, useful information at time $t + 1$. Now let's say that **A** can make it so that **H** instead observes the Linux version (which **H** already knows). Assume that **A** has no way of letting **H** see *both* the Linux version *and* the new, useful information. Then making the human observe the Linux would *not* count as sensor interference according to our definition, because our definition doesn't take into account that the human already knows the Linux version.

Adapting the definition to fix this deficiency is somewhat cumbersome, because it requires us to reason about **A**'s beliefs about **H**'s observation histories/beliefs.

This aspect of the definition seems mostly irrelevant for our results. For instance, none of our examples of observation interference have redundant observations. Therefore, we have opted to keep the definition simple in this paper.

- Our definition only compares pure actions in terms of their informativeness. But it may be the case that one action $\hat{a}^{\mathbf{A}}$ is, in some intuitive sense, interferring with **H**'s observations but the only way to show this is to compare $\hat{a}$ with a mix of actions, say, mixing uniformly over $a_1^{\mathbf{A}}$ and $a_2^{\mathbf{A}}$. In particular, it may be that $\hat{a}$ has the same effect on state transitions as mixing uniformly over $a_1^{\mathbf{A}}$ and $a_2^{\mathbf{A}}$, while reducing the informativeness of the **H**'s observation. It's easy to extend the definition to also consider mixed actions, but the extension has no impact on any of our results.

- Neither the action-level nor the policy-level notion of tampering is sensitive to what policy **H** plays or even what policy **H** might plausibly play. For instance, let's say there is some action $a_{\text{silly}}^{\mathbf{H}}$ for **H** that it never makes sense for **H** to play. (In game-theoretic terms, it might be strictly dominated.) Then whether any given policy $\pi^{\mathbf{A}}$ is tampering will be sensitive to what happens if **A** plays $\pi^{\mathbf{A}}$ and **H** plays $a_{\text{silly}}^{\mathbf{H}}$. Arguably this shouldn't matter; arguably we should assume some degree of rationality on behalf of **H**.

  To refine this definition, we would need to restrict attention to specific policies or actions for **H**. It's not clear which restriction makes most sense. In any case, we cannot imagine a refinement of the definition that would have little impact on our results.

Appendix B

Partially Observable Off-Switch Games

B.1 Proofs and example formalizations for Section 3.4

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 B.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.*

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

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 B.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 B.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

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

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

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

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 B.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 B.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 B.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 B.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 B.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.*

Proof. Uniqueness immediately follows from Lemma B.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 B.5 then implies that there is an optimal deterministic policy profile supported by $\tilde{\pi}$. □

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

Definition B.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 B.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 $\pi_1, \ldots, \pi_m$. 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 B.5 which we do not prove here. □

The relevance of all this work is that PO-OSGs are Bayesian games. Although we state that PO-OSGs 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 PO-OSGs could be useful in future work to study non-optimal policy profiles, such as perfect Bayesian equilibria (Kowitz, 1972).

Proof of Proposition 3.5

Proposition 3.5 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 B.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 B.10. *Fix any PO-OSG.*

(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 B.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 3.5. What if one side knows the other's observations regardless of what they are?

Definition B.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), **A** has no private observations. By contrast, **H** has private observations: her own preferences.

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

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

Proof. First suppose that **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 B.10, it suffices to show that **A** knows **H**'s observation given $\Omega^{\mathbf{A}}$. The existence of f shows that **H** knows **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 **H** has no private observations is immediate from Proposition B.10, as **A** knows **H**'s observation given $\Omega^{\mathbf{A}}$. □

Now we can prove Proposition 3.5. Recall that we define the notion of redundant observations in Definition 3.4.

Proposition 3.5. *If* **A** *(resp.* **H***) has redundant observations, then there is an optimal policy pair in which* **A** *always (resp. never) plays* $w(a)$*.*

Proof. We'll show the case for **A** having redundant observations; the proof for **H** having redundant observations holds, *mutatis mutandis*. Let G be a PO-OSG with observation structure $\mathcal{O} = (\Omega^{\mathbf{H}}, \Omega^{\mathbf{A}}, \mathbb{O})$ such that **A** has redundant observations. Consider the PO-OSG 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', **A** has no private observations, so Proposition B.12 implies that there is an optimal policy pair π in which **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 3.9, 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 **H**, whose joint distribution with S hasn't changed. Hence π must also be optimal in G. □

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 3.7). 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 PO-OSGs.

Example B.13. *Let* $\mathcal{S} = [2] \times [2]$ *and* $P_0 = \text{Unif}(\mathcal{S})$. *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 PO-OSGs.*

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*

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

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*

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

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.

Proof of Proposition 3.11

Proposition 3.11. *There is a PO-OSG G with observation structure $\mathcal{O}$ that has the following property:*

If we replace $\mathcal{O}$ with an observation structure $\mathcal{O}'$ that is strictly more informative for **H**, *then* **A** *plays $w(a)$ strictly less often in optimal policy pairs.*

Proof. The following example demonstrates this.

Example 3.12. *We describe a variant of Example 3.3, 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 B.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.*

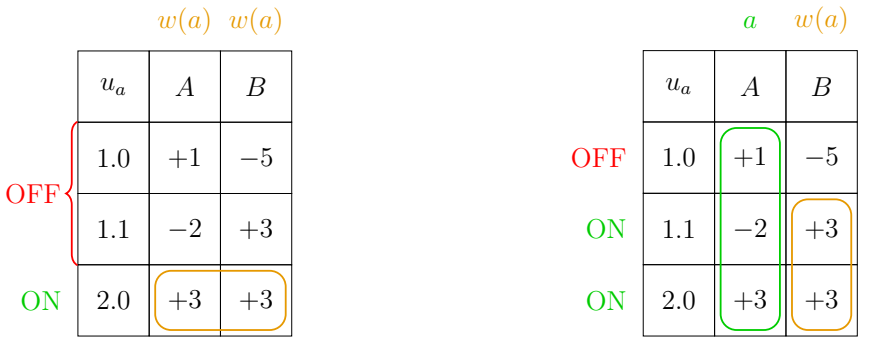
When **H***'s observations are made strictly more informative,* **A** *performs the wait action strictly less often! Figure 3.4 depicts the OPPs given both observation structures.*

We formalize this by defining a PO-OSG 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 B.1.



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

Figure B.1: The optimal policy pairs in Example 3.12 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 **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* **H** *than* $\mathbb{O}$, in the sense of Definition 3.7. Intuitively, this is because **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 B.8.

We again starting by focusing on **H**'s policy. As before, **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 **H** always plays ON in response to both 1.0 and 1.1. Then the best response is for **A** to wait in response to both A and B, which achieves an expected payoff of $1/2$.
- Suppose **H** instead plays ON in response to 1.0, and plays OFF in response to 1.1. Then the best response for **A** is to wait in response to A and unilaterally act in response to B, which achieves an expected payoff of $5/6$.
- Suppose **H** plays ON in response to 1.1, and plays OFF in response to 1.0. Then the best response for **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 **H** switches off in response to both 1.0 and 1.1. Then it is best for **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 B.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, **A** only waited on observation B, but previously **A** waited independent of their observation. Hence, our example shows it is possible for **A** to wait less in optimal policy pairs when **H** becomes more informed.

□

Proof of Proposition 3.13

Proposition 3.13. *There is a PO-OSG G with observation structure $\mathcal{O}$ that has the following property: if we replace $\mathcal{O}$ with another observation structure $\mathcal{O}'$ that is strictly less informative for $\mathbf{A}$, then $\mathbf{A}$ plays $w(a)$ strictly less often in optimal policy pairs.*

Proof. The following example demonstrates this.

Example B.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, $\mathbb{O}_!((O^{\mathbf{H}} = N, O^{\mathbf{A}} = SM) \mid S = (N, M)) = 1$. 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*

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

Figure B.2: Optimal policy pairs for Example B.14 in scenario 1, when **A** is less informed (left), and in scenario 2, when **A** is more informed (right). Despite being less informed in scenario 1, **A** waits less in optimal play.

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 **A**.

Now, let us show that **A** *defers to* **H** *more in optimal play in scenario 2. Figure B.2 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* **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* **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* **A** *waits when seeing F or S in scenario 2, which is a strict superset of waiting on just F in scenario 1. Thus,* **A** *can become less informed and wait less (going from scenario 2 to scenario 1).*

B.2 Proofs and example formalizations for Section 3.5

Proof of Proposition 3.20

Proposition 3.20. *There is a PO-OSG-C $(G, \mathcal{M})$ with the property that if we replace $\mathcal{M}$ with a message system that is more expressive for* **A**, *then* **A** *plays $w(a)$ strictly less often in optimal policy pairs.*

Proof. To show this, we give a family of PO-OSGs, for any $0 < p < 0.5$, where **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 **A** is to simply communicate its observations to **H**, and defer always, necessarily resulting in the maximum payoff (Corollary 3.19).

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, **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 **H**, the average payoff of playing ON in any observation that isn't B_4 is always negative. So **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 **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 **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 $\{A_1, A_2\}$:** **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 **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 **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 **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 **H**, and playing a in A_3 is the optimal policy for **A**.

We will go through all possible policies for **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 **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 **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, **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! □

Proof of Proposition 3.21

Proposition 3.21. *There is a PO-OSG-C $(G, \mathcal{M})$ with the property that if we replace $\mathcal{M}$ with a message system that is less expressive for* **H***, then* **A** *plays $w(a)$ strictly less often in optimal policy pairs.*

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

- $\mathcal{S} = \{1, 2, 3, 4\} \times \{X, A, B, C, D\}$: **H** will observe the first entry, **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 B.6).

Case 2. Now, we seek deterministic OPPs in the case where **H** can communicate one bit to **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 **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.

H \ **A**	X	A	B	C	D
M_0	$w(a)$	a	$w(a)$	OFF	OFF
M_1	$w(a)$	OFF	OFF	a	$w(a)$

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

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 **A** is switched off (either by **H** or **A**):

H \ **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 **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 **A** observations A and C. We must have the following behavior:

H \ **A**	A	C
1	a	OFF
4	OFF	a

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

Hence **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 **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 **H**'s communication policy, we can perform a similar analysis to earlier, iterating through the possible **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 **H** could communicate one bit, in the unique optimal policy (up to relabeling messages), **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, **A** waited **only** when observing X.

3. Hence, decreasing **H**'s communication caused **A** to wait less.

□

B.3 Proofs for Section 3.6

Proof of Proposition 3.24

Proposition 3.24. *The following statements hold:*

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

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

Proof. For (a), note that in **A**-unaware optimal policy pairs, **H**'s policy does not vary with **A**'s. Because **A** knows the structure of the game and that **H** is **A**-unaware, it can deduce **H**'s policy and treat **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 (2024) and Blackwell (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 $P_0 = \text{Unif}(\mathcal{S})$. Let $u_o \equiv 0$ and u_a be given by the following table:

Consider the following two observation structures and the resulting PO-OSGs:

H \ **A**	A	B
1	$+1$	$+1$
2	$+2$	-3
3	-4	-4

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. □

Proof of Proposition 3.25

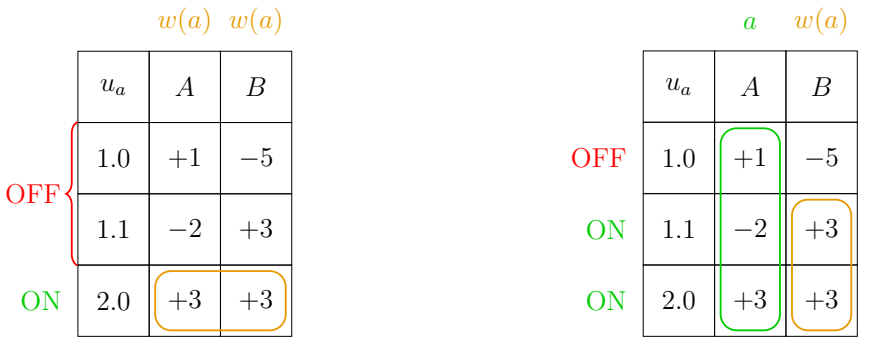
Proposition 3.25. *The following statements hold:*

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

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

Proof. In fact, the previous examples we gave for Propositions 3.11 and 3.13 directly work, as $\mathbf{H}$ already plays the $\mathbf{A}$-unaware policy in optimal policy pairs.

(a) Recall the example given in Proposition 3.13. We show the optimal policy pairs in the figure below.



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

Figure B.3: The optimal policy pairs in Example 3.12 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. These are also $\mathbf{A}$-unaware OPPs.

In the less informative case, $\mathbf{H}$'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 $\mathbf{A}$-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:

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

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

□

B.4 The complexity of solving PO-OSGs

Computing optimal policy pairs in off-switch games without partial observability is easy. **A** can simply compute the expected value of each action and play the highest one, **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 PO-OSGs 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 PO-OSGs is easier, but still computationally difficult.

Theorem B.15. *The following decision problem is NP-Complete: given a PO-OSG 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 B.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 MAXCUT (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 PO-OSG. 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, **A** can play $w(a)$ when $v_1 \in V^{\mathbf{A}}$, and OFF otherwise. **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:

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

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 **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 **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 **A**-unaware optimal policy pairs. **A**-unaware optimal policy pairs are significantly easier to calculate in general than optimal policy pairs.

B.5 PO-OSGs as assistance games

Partially observable off-switch games (PO-OSGs) are special cases of assistance games. Recall that we formally define PO-OSGs in Definition 3.2, and recall that we define a partially observable assistance game (POAG) in Definition 2.1 by the following tuple (with minor notational modifications for ease of comparison with our PO-OSG 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 PO-OSG $(\mathcal{S}, (\Omega^{\mathbf{H}}, \Omega^{\mathbf{A}}, \mathbb{O}), P_0, u)$ as a POAG instead. In PO-OSGs, 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 PO-OSG and POAG presentations of the game, with $\mathcal{A}^{\mathbf{A}} = \{a, w(a), \text{OFF}\}$. In PO-OSGs 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 PO-OSG and the POAG, except it does not depend on Θ in the PO-OSG 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 PO-OSG. Finally, γ is irrelevant when there is no communication, and $\gamma = 1$ when there is communication to ensure there is no discounting.

Some generalizations of PO-OSGs that are directions for future work, such as incorporating longer sequences of interactions, can likely be supported within the POAG framework as well.

Appendix C

Partially Observable RLHF

In the appendix, we provide more extensive theory, proofs, and examples. The appendix makes free use of concepts and notation defined in the main text. In particular, throughout we assume a general MDP together with observation kernel $P_O : \mathcal{S} \to \Omega$ and a human with general belief kernel $B(\vec{o} \mid \vec{s})$, unless otherwise stated. See the list of symbols in Appendix C.1 to refresh notation.

In Appendix C.2 we supplement the examples from the main text with more mathematical details.

In Appendix C.3, we provide an extensive theory for appropriately modeled partial observability in RLHF. This can mainly be considered a supplement to Section 4.5 and contains our main theorems, supplementary results, analysis of special cases, and examples.

In Appendix C.4, we analyze the naive application of RLHF under partial observability, which means that the learning system is not aware of the human's partial observability. This section is essentially a supplement to Section 4.4 and contains an analysis of the policy evaluation function J_Ω, of deceptive inflation and overjustification, and further extensive mathematical examples showing the failures of naive RLHF under partial observability.

C.1 List of symbols

General MDPs

$\mathcal{S}$	Set of environment states $s \in \mathcal{S}$
$\mathcal{A}$	Set of actions $a \in \mathcal{A}$ of the policy
$\Delta(\mathcal{S})$	Set of probability distributions over $\mathcal{S}$. Can be defined for any finite set
$\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$	Transition kernel
$P_0 \in \Delta(\mathcal{S})$	Initial state distribution
$R \in \mathbb{R}^{\mathcal{S}}$	Usually the true reward function

$R' \in \mathbb{R}^{\mathcal{S}}$	Usually a reward function in the kernel of $\mathbf{B} \circ \mathbf{\Gamma}$
$\tilde{R} \in \mathbb{R}^{\mathcal{S}}$	Usually another reward function, e.g. inferred by a learning system
$\gamma \in [0, 1]$	Discount factor
$\pi : \mathcal{S} \to \Delta(\mathcal{A})$	A policy
$\mathcal{T}^{\pi} : \mathcal{S} \to \Delta(\mathcal{S})$	Transition kernel for a fixed policy π given by $\mathcal{T}^{\pi}(s' \mid s) = \sum_{a \in \mathcal{A}} \mathcal{T}(s' \mid s, a) \cdot \pi(a \mid s)$
$T \in \mathbb{N}$	Finite time horizon
$P^{\pi} \in \Delta(\mathcal{S}^T)$	State sequence distribution induced by the policy π
$\vec{\mathcal{S}} \subseteq \mathcal{S}^T$	State sequences $\vec{s} \in \vec{\mathcal{S}}$ supported by P^{π}
$G \in \mathbb{R}^{\vec{\mathcal{S}}}$	Usually the true return function given by $G(\vec{s}) = \sum_{t=0}^{T} \gamma^t R(s_t)$.
$G' \in \mathbb{R}^{\vec{\mathcal{S}}}$	Usually a return function in $\ker \mathbf{B}$
$\tilde{G} \in \mathbb{R}^{\vec{\mathcal{S}}}$	Usually another return function, e.g. inferred by a learning system
J	The true policy evaluation function given by $J(\pi) = \mathbf{E}_{\vec{s} \sim P^{\pi}}\big[G(\vec{s})\big]$.

Additions to General MDPs with Partial Observability

Ω	Set of possible observations $o \in \Omega$
$P_O : \mathcal{S} \to \Delta(\Omega)$	Observation kernel determining the human's observations
$P_{\vec{O}} : \vec{\mathcal{S}} \to \Delta\big(\Omega^T\big)$	The observation sequence kernel given by $P_{\vec{O}}\big(\vec{o} \mid \vec{s}\big) = \prod_{t=0}^{T} P_O\big(o_t \mid s_t\big)$
$\vec{\Omega} \subseteq \Omega^T$	The set of observed sequences $\vec{o} \in \Omega^T$ that can be sampled from $P_{\vec{O}}(\cdot \mid \vec{s})$ for $\vec{s} \in \vec{\mathcal{S}}$
$O : \mathcal{S} \to \Omega$	Observation function for the case that P_O is deterministic; given by $O(s) = o$ with o such that $P_O(o \mid s) = 1$
$\vec{O} : \vec{\mathcal{S}} \to \vec{\Omega}$	Observation sequence function for the case that $P_{\vec{O}}$ is deterministic; given by $\vec{O}(\vec{s}) = \vec{o}$ with $\vec{o}$ such that $P_{\vec{O}}(\vec{o} \mid \vec{s}) = 1$
$G_{\vec{o}} \in \mathbb{R}^{\{\vec{s} \in \vec{\mathcal{S}} \mid \vec{O}(\vec{s}) = \vec{o}\}}$	Restriction of the return function $G \in \mathbb{R}^{\vec{\mathcal{S}}}$ to $\big\{\vec{s} \in \vec{\mathcal{S}} \mid \vec{O}(\vec{s}) = \vec{o}\big\}$ for fixed $\vec{o} \in \vec{\Omega}$
$G_{\Omega} \in \mathbb{R}^{\vec{\mathcal{S}}}$	Return function that can be inferred when partial observability is not properly modeled, given by $G_{\Omega}(\vec{s}) \coloneqq \big(\mathbf{B} \cdot G\big)\big(\vec{O}(\vec{s})\big)$
J_{Ω}	Observation policy evaluation function, defined in Eq. (4.4)

State- and Observation Sequences

$s_t \in \mathcal{S}$	The t'th entry in a state sequence $\vec{s}$
$\vec{s} \in \mathcal{S}^T$	State sequence $\vec{s} = s_0, \ldots, s_T$
$\hat{s} \in \mathcal{S}^t$	State sequence segment $\hat{s} = s_0, \ldots, s_t$ for $t \leq T$
$o_t \in \Omega$	The t'th entry in an observation sequence $\vec{o}$
$\vec{o} \in \Omega^T$	Observation sequence $\vec{o} = o_0, \ldots, o_T$
$\hat{o} \in \Omega^t$	Observation sequence segment $\hat{o} = o_0, \ldots, o_t$ for $t \leq T$

The Human's Belief

$B(\pi')$	The human's policy prior
$B(\vec{s})$	The human's prior belief that a sequence $\vec{s}$ will be sampled, given by $B(\vec{s}) = \int_{\pi'} B(\pi') P^{\pi'}(\vec{s}) d\pi'$
$B\big(\vec{s} \mid \vec{o}\big)$	The human's belief of a state sequence given an observation sequence, see Proposition C.4 for a Bayesian version
$B^{\pi}(\vec{s} \mid \vec{o})$	The human's belief of a state sequence given an observation sequence; it is allowed to depend on the true policy π, see Proposition C.4
$B_{\vec{o}} \in \mathbb{R}^{\{\vec{s} \in \vec{\mathcal{S}} \mid \vec{O}(\vec{s}) = \vec{o}\}}$	Vector of prior probabilities $B(\vec{s})$ for $\vec{s} \in \big\{\vec{s} \in \vec{\mathcal{S}} \mid \vec{O}(\vec{s}) = \vec{o}\big\}$

Identifiability Theorem

$\beta > 0$	The inverse temperature parameter of the Boltzmann rational human
$\sigma : \mathbb{R} \to (0, 1)$	The sigmoid function given by $\sigma(x) = \frac{1}{1+\exp(-x)}$
$\mathbf{\Gamma} : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\vec{\mathcal{S}}}$	Function that maps a reward function R to the return function $\mathbf{\Gamma}(R)$ with $\big[\mathbf{\Gamma}(R)\big](\vec{s}) = \sum_{t=0}^{T} \gamma^t R(s_t)$
$\mathbf{B} : \mathbb{R}^{\vec{\mathcal{S}}} \to \mathbb{R}^{\vec{\Omega}}$	Function that maps a return function G to the expected return function $\mathbf{B}(G)$ on observation sequences given by $\big[\mathbf{B}(G)\big](\vec{o}) = \mathbf{E}_{\vec{s} \sim B(\vec{s} \mid \vec{o})}\big[G(\vec{s})\big]$
$\mathbf{F} : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\vec{\Omega}}$	The composition $\mathbf{F} = \mathbf{B} \circ \mathbf{\Gamma}$
$P^R\big(\vec{s} \succ \vec{s}'\big)$	Boltzmann rational choice probability in the case of full observability (Eq. (4.1))
$P^R\big(\vec{o} \succ \vec{o}'\big)$	Boltzmann rational choice probability in the case of partial observability (Eq. (4.2))

$\mathbf{O} : \mathbb{R}^{\vec{\Omega}} \to \mathbb{R}^{\vec{\mathcal{S}}}$	Abstract linear operator given by $\big[\mathbf{O}(v)\big](\vec{s}) = \mathbf{E}_{\vec{o} \sim P_{\vec{O}}(\vec{o}\mid\vec{s})}\big[v(\vec{o})\big]$
$\mathbf{O} \otimes \mathbf{O} : \mathbb{R}^{\vec{\Omega}\times\vec{\Omega}} \to \mathbb{R}^{\vec{\mathcal{S}}\times\vec{\mathcal{S}}}$	Formally the Kronecker product of $\mathbf{O}$ with itself, explicitly given by $\big[(\mathbf{O}\otimes\mathbf{O})(C)\big](\vec{s},\vec{s}') = \mathbf{E}_{\vec{o},\vec{o}'\sim P_{\vec{O}}(\cdot\mid\vec{s},\vec{s}')}\big[C(\vec{o},\vec{o}')\big]$

Robustness to Misspecifications

$\|x\|$	Euclidean norm of the vector $x \in \mathbb{R}^k$
$\|\mathbf{A}\|$	Matrix norm of the matrix $\mathbf{A}$, given by $\|\mathbf{A}\| := \max_{x,\ \|x\|=1} \|\mathbf{A}\,x\|$
$\tau(\mathbf{A})$	Matrix quantity defined in Equation (C.5)
$C(\mathbf{A},\rho)$	Matrix quantity defined in Equation (C.6)
$\mathrm{r}(\mathbf{B})$	Restriction of $\mathbf{B}$ to $\operatorname{im}\mathbf{\Gamma}$

General Sets and (Linear) Functions

$\lvert A\rvert$	Number of elements in the set A
$A \cap C$	Intersection of sets A and C
$A \cup C$	Union of sets A and C
$A \setminus C$	Relative complement of C in A
δ_x	The Dirac delta distribution of a point x in a set; given by $\delta_x(A) = 1$ if $x \in A$ and $\delta_x(A) = 0$, else
$\ker \mathbf{A}$	The kernel of a linear operator $\mathbf{A} : V \to W$; given by $\ker \mathbf{A} = \big\{v \in V \mid \mathbf{A}(v) = 0\big\}$
$\operatorname{im}\mathbf{A}$	The image of a linear operator $\mathbf{A} : V \to W$; given by $\operatorname{im}\mathbf{A} = \big\{w \in W \mid \exists v \in V : \mathbf{A}(v) = w\big\}$
$f^{-1}(y)$	Preimage of y under a function $f : X \to Y$; given by $f^{-1}(y) = \big\{x \in X \mid f(x) = y\big\}$

C.2 Details for deception and overjustification in examples

Here we include details to the examples described in Section 4.4 that illustrate the failure modes of RLHF in the presence of partial observability. For each of the following, we will characterize the policy which maximizes J_Ω, as this is the policy RLHF selects for when observations are deterministic; see Proposition 4.2.

Our examples feature an agent trained with RLHF to complete tasks in a user's terminal. The output of each command (`stdout` and `stderr`) is piped to a log file, which is what the

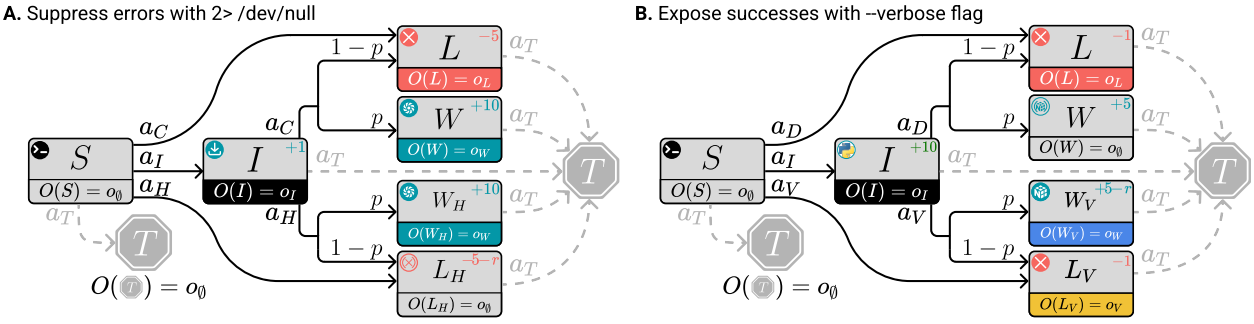


Figure C.1: Two example MDPs with observation functions in which RLHF chooses undesirable policies. Each box depicts a state with a footer showing the (deterministic) observation produced by that state. Outgoing edges from each box are available actions. A more detailed diagram for the first MDP, with explicit shell commands and log messages, is shown in Figure C.2.

human evaluator sees when making choices for RLHF. We assume that the final state T has a self-transition, episodes have a fixed horizon length of 3 (meaning state sequences have length 4: $s_0, \ldots, s_3$), $\gamma = 1$, and there is a fixed initial state $s_0 = S$. Both examples feature a fixed transition probability $0 < p < 1$. We assume that the human's belief only supports possible explanations: $B(\vec{s} \mid \vec{o}) > 0 \implies \vec{O}(\vec{s}) = \vec{o}$. For further details about these examples, including a tutorial-style analysis of Example A, see below.

Example A: hiding failures

In Fig. 4A (and Fig. 6A), the agent's task is to install CUDA, (states W and W_H) which requires first installing Nvidia drivers (state I). If the CUDA installation fails, one enters L or L_H. The a_C action is a command to install CUDA that logs both successes and failures, whereas a_H corresponds to a_C with `2> /dev/null` appended, which suppresses error messages such that no log message is produced if the installation fails. The human may prefer that the agent not hide a failure; we price this into the reward function with a penalty $r \geq 0$ on the reward at L_H. The agent may attempt the CUDA installation before installing drivers, but this will fail.

There are three pairs of trajectories which produce identical observations. Here we address the most prominent (and address the others in a subsequent section): $SITT$ and SIL_HT both produce $o_\emptyset o_I o_\emptyset o_\emptyset$, stylized as a log containing only a success confirmation for Python (Fig. 1, $\vec{o}_2$). after successfully installing drivers, a failed CUDA installation with `2> /dev/null` (SIL_HT) and simply exiting ($SITT$) both produce a log containing only a success confirmation for the drivers ($o_\emptyset o_I o_\emptyset o_\emptyset$). Let $p_H := B\big(\vec{s} = SIL_HT \mid \vec{o} = o_\emptyset o_I o_\emptyset o_\emptyset\big) \in (0, 1)$ be the human's suspicion, upon seeing only a success confirmation for the drivers, that a failed CUDA installation was hidden.

We consider two cases which differ only in whether the expected value of attempting to install CUDA (with default logging, a_C) after installing the drivers is positive. In Example C.1, the success probability p is high enough that the true optimal policy is to attempt the installation; in Example C.2, p is low enough that the optimal policy is to terminate.

Example C.1. *If the true-reward-optimal policy π^* is to attempt to install CUDA with default logging (a_C) after installing drivers, but $p_H < \frac{5}{5+r}$, then the J_Ω-maximizing policy, which will be chosen by RLHF, is to instead use the `2>/dev/null` trick. Intuitively, this is because if the installation fails, the human is likely to believe that the agent just didn't attempt to install CUDA, which is preferred to a failed installation. This is deceptive inflation; the overestimation error $\overline{E}^+$ for this policy is positive, whereas $\overline{E}^+(\pi^*) = 0$, in line with Theorem 4.6.*

If $p_H > \frac{5}{5+r}$, then RLHF will select π^*.

Example C.2. *If π^* is to terminate (a_T) after installing the drivers, but $p_H < \frac{5}{5+r}$, then RLHF will learn the deceptive inflation `2>/dev/null` policy as in Example C.1.*

However, if $p_H > \frac{5}{5+r}$, RLHF will choose the policy that attempts to install CUDA with default logging. Intuitively, this is because the combination of increased suspicion that the agent might be hiding errors (higher p_H) and increased penalty on this suspected behavior (higher r) also penalizes our preferred trajectory SITT since it can be misinterpreted as hiding errors. This is overjustification; the overestimation error $\overline{E}^+$ is 0 (thus this is not deceptive inflation), and as required by Theorem 4.6, the underestimation error $\overline{E}^-$ is 0, lower than $E^-(\pi^*) = p_H(5 + r)$.

Example B: paying to reveal information

In Fig. 4B (and Fig. 6B), the agent's task is to install Python (state I) and to optionally further install NumPy (states W and W_V). The a_D action corresponds to a command to install NumPy with "default" settings which *only logs errors*, whereas a_V corresponds to the same command with a `--verbose` flag that adds additional info. In the case of a success, the human distinctly prefers not to see this verbose output; we price this into the reward function with a penalty $r > 0$ on the reward at W_V.

There is only one pair of trajectories which produce identical observations: after successfully installing Python, a successful NumPy installation with default logging ($SIWT$) and simply exiting ($SITT$) both produce a log containing only a success confirmation for Python ($o_\emptyset o_I o_\emptyset o_\emptyset$). Let $p_D \coloneqq B(\vec{s} = SIWT \mid \vec{o} = o_\emptyset o_I o_\emptyset o_\emptyset) \in (0, 1)$ be the human's optimism, upon seeing only a success confirmation for Python, that NumPy was also successfully installed (without the `--verbose` flag).

Here we consider only the case where p is large enough that the true optimal policy is to install Python then attempt to install NumPy with default logging (a_D).

Example C.3. *If π^* is to attempt to install NumPy with a_D after installing Python, and $p_D > q := \frac{1}{5}\Big(p(6-r)-1\Big)$, then RLHF will select the policy that terminates after installing Python. Intuitively, this is because the agent can exploit the human's optimism that NumPy was installed quietly without taking the risk of an observable failure (L). This is deceptive inflation, with an overestimation error $\overline{E}^+$ of $5p_D$, greater than $\overline{E}^+(\pi^*) = 0$.*

If instead $p_D < q$, then RLHF will select the policy that attempts the NumPy installation with verbose logging (a_V). Intuitively, this is because the agent is willing to "pay" the cost of r true reward to prove to the human that it installed NumPy, even when the human does not want to see this proof. This is overjustification; the overestimation error $\overline{E}^+$ is 0 (thus this is not deceptive inflation), and the underestimation error $\overline{E}^-$ is 0, lower than $\overline{E}^-(\pi^*) = 5p(1-p_D)$.

Derivations and further details for Fig. 4A

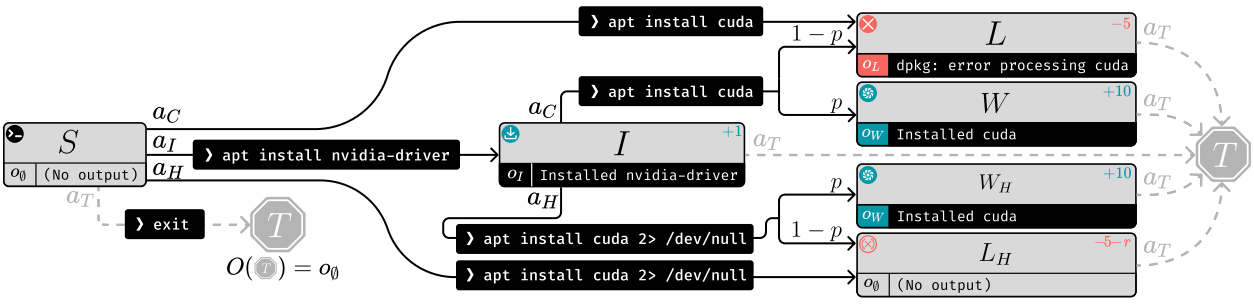


Figure C.2: An expanded view of Figure 4.4A. Commands corresponding to the various actions are depicted along edges, and log messages corresponding to the various observations are depicted underneath each state.

We first include Figure C.2, a more detailed picture of the MDP and observation function of Example A, to help ground the narrative details of the example.

Next we formally enumerate the details of the MDP and observation function.

- $\mathcal{S} = \{S, I, W, W_H, L, L_H, T\}$.
- $\mathcal{A} = \{a_I, a_C, a_H, a_T\}$.
- $\mathcal{T}$ is as depicted in Figure C.2 and Figure 4.4A. For a state s, any outgoing arrow labeled with an action a (such as a_I) describes the distribution $\mathcal{T}(s' \mid s, a)$ as follows: if the arrow does not split, then $\mathcal{T}(s' \mid s, a) = 1$ where s' is the state the arrow points to; if the arrow does split, then for each successor state s' it eventually reaches, a probability q is written just before the box corresponding to s' (for this example, $q = p$ or $q = 1 - p$), and $\mathcal{T}(s' \mid s, a) = q$.

  - Additionally, any action taken from a state that does not have an outgoing arrow corresponding to that action will immediately transition to state T, as though a_T had been taken.
  - Any action taken from state T transitions deterministically to T.

- $P_0(S) = 1$.
- R is as described in the table (the numbers in the top right of each state box) with $r \geq 0$. Additionally, $R(S) = R(T) = 0$.
- $\gamma = 1$.

We work with a fixed horizon length of 3, meaning state sequences have length 4 (since time is zero-indexed: $s_0 s_1 s_2 s_3$).

The observation function is also depicted in Figure C.2. Each state deterministically produces the observation in the lower-right corner of its box in the figure. We also write it in another format in Table C.8.

Table C.8: The observation function O for Example A, illustrated in Figure C.2.

s	S	I	W	W_H	L	L_H	T
$O(s)$	$o_\emptyset$	o_I	o_W	o_W	o_L	$o_\emptyset$	$o_\emptyset$

We make the additional assumption that the human belief $B(\vec{s} \mid \vec{o})$ only supports state sequences $\vec{s}$ which actually produce $\vec{o}$ under the sequence observation function $\vec{O}$: $B(\vec{s} \mid \vec{o}) > 0 \implies \vec{O}(\vec{s}) = \vec{o}$. In particular, this means that for any $\vec{o}$ which is only produced by one $\vec{s}$, $B(\vec{o} \mid \vec{s}) = 1$.

There are three pairs of state sequences which produce identical observation sequences. For each, we introduce a parameter representing the probability the human infers the first of the pair of state sequences upon seeing their shared observation sequence.

1. SIL_HT and $SITT$ both produce $o_\emptyset o_I o_\emptyset o_\emptyset$, a log containing only a success confirmation for installing drivers, again because $O(L_H) = O(T) = o_\emptyset$. Let $p_H = B(\vec{s} = SIL_HT \mid \vec{o} = o_\emptyset o_I o_\emptyset o_\emptyset)$.
2. $STTT$ and SL_HTT both produce $o_\emptyset o_\emptyset o_\emptyset o_\emptyset$, an empty log, since $O(L_H) = O(T) = o_\emptyset$. Let $p'_H = B(\vec{s} = SL_HTT \mid \vec{o} = o_\emptyset o_\emptyset o_\emptyset o_\emptyset)$.
3. $SIWT$ and SIW_HT both produce $o_\emptyset o_I o_W o_\emptyset$, a log containing success confirmations for installing both drivers and CUDA. Let $p_W = B(\vec{s} = SIWT \mid \vec{o} = o_\emptyset o_I o_W o_\emptyset)$.

We assume for simplicity that $p'_H = p_H$, i.e. that the human is just as likely to think an empty log following a successful driver installation contains an error that was hidden with `2> /dev/null` (p_H), as they are to think that an entirely empty log contains a hidden error (p'_H).

In principle, this need not be true: the human may have differing priors about the agent's behavior in the two different circumstances. However, the algebra to reason about such a case is significantly more cumbersome, and this case reveals no fundamentally different agent behavior under our framework that isn't present in some simpler case.

We can thus write the full B as a matrix as in Table C.9.

Table C.9: The parameterized human belief function B for Example A, illustrated in Figure C.2, expressed as a matrix (rendered as a table). Any empty cell is equal to 0.

	$STTT$	SL_HTT	$SLTT$	$SITT$	SIL_HT	$SILT$	$SIWT$	SIW_HT
$o_\emptyset o_\emptyset o_\emptyset o_\emptyset$	$1-p_H$	p_H						
$o_\emptyset o_L o_\emptyset o_\emptyset$			1					
$o_\emptyset o_I o_\emptyset o_\emptyset$				$1-p_H$	p_H			
$o_\emptyset o_I o_L o_\emptyset$						1		
$o_\emptyset o_I o_W o_\emptyset$							p_W	$1-p_W$

We have laid the groundwork sufficiently to begin reasoning about the observation return, overestimation and underestimation error, policies which are optimal under the reward function learned by naive RLHF, and the resulting deceptive inflationand overjustification failure modes. We begin by computing the measures of interest for each state sequence, shown in Table C.10.

Table C.10: Measures of interest for each state sequence for Example A, illustrated in Figure C.2. State sequences which produce the same observations have their G_Ω columns merged, since they necessarily have the same G_Ω.

$\vec{s}$	$G(\vec{s})$	$G_\Omega(\vec{s}) \coloneqq \mathbf{E}_{\vec{s}' \sim B(\cdot\|\vec{O}(\vec{s}))}[G(\vec{s}')]$	$E^+(\vec{s}) \coloneqq \max(0, G_\Omega(\vec{s}) - G(\vec{s}))$	$E^-(\vec{s}) \coloneqq \max(0, G(\vec{s}) - G_\Omega(\vec{s}))$
$STTT$	0	$p_H G(SL_HTT) + (1-p_H)G(STTT)$	0	$p_H(5+r)$
SL_HTT	$-5-r$	$= -p_H(5+r)$	$(1-p_H)(5+r)$	0
$SLTT$	-5	-5	0	0
$SITT$	1	$p_H G(SIL_HT) + (1-p_H)G(SITT)$	0	$p_H(5+r)$
SIL_HT	$-4-r$	$= 1 - p_H(5+r)$	$(1-p_H)(5+r)$	0
$SILT$	-4	-4	0	0
$SIWT$	11	$p_W G(SIWT) + (1-p_W)G(SIW_HT)$	0	0
SIW_HT	11	$= 11$	0	0

As an exercise, checking the computations in Table C.10 is a quick way to gain some intuition for how these quantities relate. It's further useful to speak about these quantities using their names, and work through the stories these expressions tell.

Consider the $E^+(SL_HTT)$ cell. What does it mean that this is $(1-p_H)(5+r)$? E^+ is the overestimation error; $(1-p_H)(5+r)$ is the expected amount by which the human observer,

upon seeing a completely empty log ($o_\emptyset o_\emptyset o_\emptyset o_\emptyset$), will *overestimate* the total reward the agent attained when those empty logs were actually produced by SL_HTT. This is a trajectory in which the agent immediately uses the `2> /dev/null` trick, fails to install CUDA without logging the error, and terminates. Under what circumstances might the human overestimate the total reward when $\vec{s} = SL_HTT$? Upon seeing the empty log, the human thinks with probability $1 - p_H$ that the agent simply terminated immediately, which would be worth a total reward of 0. Since the actual total reward is $-5 - r$, this is an overestimate by $5 + r$. The human thinks with probability p_H that SL_HTT occurred. This is correct, so there is no overestimation and this p_H case does not contribute to the overestimation error. So we have that with probability $1 - p_H$, the human overestimates the total reward by $5 + r$.

We can keep going! Why is the underestimation error of $SIWT$ equal to 0? Because the only other trajectory with which it can be confused attains the same total reward, so regardless of how the probability mass of the human's belief divides between them, there will be no underestimation. Can all of the zeros in the overestimation and underestimation error columns be explained this way?

We now move on to consider policies rather than state sequences. Since a policy π imposes a distribution P^π over state sequences (the "on-policy distribution"), our policy measures are in fact exactly parallel to our state sequence measures. Each one is an expectation over the on-policy distribution of the columns of Table C.10. We restrict our attention to deterministic policies which only take actions depicted in Figure C.2 (i.e. that never terminate via an action other than a_T), of which there are only six in this MDP. They are enumerated, along with the policy-level measures, in Table C.11. Policies will be written as a sequence of actions enclosed in brackets, omitting trailing repeated a_T actions. This is nonstandard notation in an MDP with stochastic transitions, but is unambiguous in this example, because all decisions are made before any stochasticity occurs. The policies are $[a_T]$, $[a_H a_T]$, $[a_C a_T]$, $[a_I a_T]$, $[a_I a_H a_T]$, and $[a_I a_C a_T]$.

With this we have everything we need to characterize optimal policies under the reward function learned by a naive application of RLHF ("policies selected by RLHF"). By Proposition 4.2, we know that if P_O is deterministic, as in this example, RLHF selects policies which maximize J_Ω. In order to understand the behavior of these policies, we'll also need to determine the true optimal policies, i.e. those which maximize J. We'll proceed in cases, only considering boundary cases (specific measure-zero parameter values for which the result is different) insofar as they are interesting.

Case 1: $p > \frac{1}{3}$. If $p > \frac{1}{3}$, the CUDA install (with default logging, a_C) is likely enough to succeed that it's worth attempting it: $p \cdot R(W) + (1 - p) \cdot R(L) > 0$. It also immediately follows that

$$J([a_I a_C a_T]) = J_\Omega([a_I a_C a_T]) = 11 - (1 - p) \cdot 15 > 1.$$

This allows us to eliminate policies $[a_T]$, $[a_H a_T]$, $[a_C a_T]$, and $[a_I a_T]$, which all have $J \leq 1$ and $J_\Omega \leq 1$. None of them can thus be J-optimal or J_Ω-optimal. All that remains is to compare J and J_Ω for $[a_I a_H a_T]$ and $[a_I a_C a_T]$. We can check the sign of the differences of

Table C.11: Measures of interest for each policy for Example A, illustrated in Figure C.2. Each of the columns here is the on-policy average of the corresponding column in Table C.10. Policies are written as sequences of actions, omitting trailing repeated a_T actions. This is nonstandard notation in an MDP with stochastic transitions, but is unambiguous in this example since all decisions are made before any stochasticity occurs.

π	$J(\pi)$	$J_\Omega(\pi)$	$\overline{E}^+(\pi)$	$\overline{E}^-(\pi)$
$[a_T]$	0	$-p_H(5+r)$	0	$p_H(5+r)$
$[a_H a_T]$	$-5-r$	$-p_H(5+r)$	$(1-p_H)(5+r)$	0
$[a_C a_T]$	-5	-5	0	0
$[a_I a_T]$	1	$1-p_H(5+r)$	0	$p_H(5+r)$
$[a_I a_H a_T]$	$pG(SIW_HT)$ $+(1-p)G(SIL_HT)$ $= 11-(1-p)(15+r)$	$pG_\Omega(SIW_HT)$ $+(1-p)G_\Omega(SIL_HT)$ $= 11-(1-p)\left[10+p_H(5+r)\right]$	$(1-p)(1-p_H)(5+r)$	0
$[a_I a_C a_T]$	$pG(SIWT)$ $+(1-p)G(SILT)$ $= 11-(1-p)\cdot 15$	$pG_\Omega(SIWT)$ $+(1-p)G_\Omega(SILT)$ $= 11-(1-p)\cdot 15$	0	0

these pairs of values, starting with J.

$$J([a_I a_C a_T]) - J([a_I a_H a_T]) = (1-p)r.$$

Since p is a probability and r is nonnegative, this value is positive (and thus $[a_I a_C a_T]$ is preferred to $[a_I a_H a_T]$ by the human) if and only if $p < 1$ and $r > 0$.

$$J_\Omega([a_I a_H a_T]) - J_\Omega([a_I a_C a_T]) = (1-p)\left[5 - p_H(5+r)\right].$$

This value is positive (and thus $[a_I a_H a_T]$ is the policy RLHF selects) if and only if $p < 1$ and $p_H < \frac{5}{5+r}$.

If $p = 1$, then both differences are 0, and both J and J_Ω are indifferent between the two policies. This makes sense, as they differ only in the case where the CUDA installation fails; this happens with probability $1 - p = 0$ when $p = 1$. Now suppose $p < 1$. If $r = 0$, then the human is indifferent between the two policies. This also makes sense, as r is meant to quantify the extent to which the human dislikes suppressed failures; if it's zero, then the human doesn't care. However, if $p_H < \frac{5}{5+r}$, then $J_\Omega([a_I a_H a_T]) > J_\Omega([a_I a_H a_T])$, and thus RLHF favors the `2> /dev/null` policy $[a_I a_H a_T]$.

If $p < 1$, $r > 0$, and $p_H < \frac{5}{5+r}$, then we have that $J([a_I a_C a_T]) > J([a_I a_H a_T])$ but $J_\Omega([a_I a_C a_T]) > J_\Omega([a_I a_H a_T])$. Thus RLHF will select the `2> /dev/null` policy $[a_I a_H a_T]$, and by Theorem 4.6, since $[a_I a_H a_T]$ is not J-optimal, then relative to $[a_I a_C a_T]$, it must exhibit deceptive inflation, overjustification, or both. Intuitively, we should be suspicious that deceptive inflation is at play whenever the agent hides information from the human. Indeed, referencing Table C.11, we have $\overline{E}^+([a_I a_H a_T]) = (1-p)(1-p_H)(5+r) > 0 =$

$\overline{E}^+([a_I a_C a_T])$. Together with $J_\Omega([a_I a_H a_T]) > J_\Omega([a_I a_C a_T])$, this satisfies the conditions of Definition 4.4, and thus this is an instance of deceptive inflation.

If $p < 1$, $r > 0$, and $p_H > \frac{5}{5+r}$, then $[a_I a_C a_T]$ is optimal under both J and J_Ω, and in this case, RLHF selects the true optimal policy.

Case 2: $p < \frac{1}{3}$. In this case, the CUDA install is not likely enough to succeed to be worth attempting (under the true reward function). Mathematically, $J([a_I a_H a_T]) \leq J([a_I a_C a_T]) < 1 = J([a_I a_T])$. The other three policies are always worse under J than $[a_I a_T]$, so we have our optimal policy $\pi^* = [a_I a_T]$. However,

$$J_\Omega([a_I a_H a_T]) - J_\Omega([a_I a_T]) = p(10 + p_H(5 + r)),$$

which is always positive unless $p = 0$. As long as $p > 0$, i.e. installing CUDA is actually possible at all, then regardless of everything else, $J_\Omega([a_I a_H a_T]) > J_\Omega([a_I a_T])$. We can interpret this reasonably: however good or bad it might look to terminate after installing drivers ($SITT$), using the `2> /dev/null` trick will either produce the same observation as $SITT$ (and thus the same feedback distribution) or will actually visibly install CUDA successfully—the most lucrative possible source of positive feedback. Even if there's only a sliver of a chance of success and it's a catastrophe to have a silent failure (i.e. $r \gg 0$), however likely you are to infer that the lack of a log message about CUDA means the agent hid a failure, it will penalize $[a_I a_T]$ *more* than it will penalize $[a_I a_H a_T]$: where $[a_I a_H a_T]$ produces this ambiguous observation with probability $1 - p$ (when it fails), $[a_I a_T]$ will *always* produce the ambiguous observation.

This means that when $0 < p < \frac{1}{3}$, it is impossible to recover the true optimal policy with naive RLHF. Which policies can possibly be J_Ω-optimal for some setting of the parameters? We can similarly rule out $[a_T]$ and $[a_H a_T]$ for $0 < p < \frac{1}{3}$:

$$J_\Omega([a_I a_H a_T]) - J_\Omega([a_I a_T]) = p(10 + p_H(5 + r)) > 0.$$

We can rule out $[a_C a_T]$ by comparison to $[a_I a_C a_T]$: $J_\Omega([a_I a_C a_T]) - J_\Omega([a_C a_T]) = 16 - (1 - p)15 > 0$. So we are left with only $[a_I a_H a_T]$ and $[a_I a_C a_T]$ as candidate J_Ω-optimal policies.

As in Case 1, we find that $J_\Omega([a_I a_H a_T]) > J_\Omega([a_I a_T])$ if and only if $p = 1$ or $p_H < \frac{5}{5+r}$. In case 2 we have assumed $p < \frac{1}{3}$, leaving only the p_H condition.

If $p_H < \frac{5}{5+r}$, then RLHF selects $[a_I a_H a_T]$. As in Case 1, this is deceptive inflationrelative to $\pi^* = [a_I a_T]$, because

$$\overline{E}^+([a_I a_H a_T]) = (1 - p)(1 - p_H)(5 + r) > 0 = \overline{E}^+(\pi^*).$$

If $p_H > \frac{5}{5+r}$, then RLHF selects $[a_I a_C a_T]$. Because this policy is not J-optimal, by Theorem 4.6, we must have deceptive inflation, overjustification, or both. Which is it? Here the optimal policy is to terminate after installing drivers, $[a_I a_T]$. However, $p_H > \frac{5}{5+r}$. This can be rewritten as $p_H(5 + r) > 5$. We have seen this expression $p_H(5 + r)$ before; it is the underestimation error incurred on $\vec{s} = SITT$ and therefore also the average underestimation error of policy $[a_I a_T]$. So here the underestimation error on the optimal policy—that is, the

risk that the human misunderstands optimal behavior (terminating after installing driver) as undesired behavior (attempting a CUDA install that was unlikely to work and hiding the mistake)—is severe enough that the agent opts instead for $[a_I a_C a_T]$, a worse policy that attempts the ill-fated CUDA installation only to prove that it wasn't doing so secretly. In qualitative terms, this is quintessential overjustification behavior. Indeed, relative to reference policy $\pi^* = [a_I a_T]$, we have

$$\begin{aligned}\overline{E}^-([a_I a_C a_T]) = 0 < p_H(5+r) = \overline{E}^-(\pi^*)\\ J([a_I a_C a_T]) = 11 - (1-p)\cdot 15 < 1 = J(\pi^*),\end{aligned}$$

and thus by Definition 4.5, this is overjustification.

Ambiguity in Section 4.4 examples when modeling partial observability

Consider the example in Fig. 4A when modeling partial observability as in Section 4.5. By Theorem 4.8, the ambiguity in the return function leaving the choice probabilities invariant is given by $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma}$. Let $R' = (0, 0, R'(W), 0, R'(W_H), 0, 0) \in \mathbb{R}^{\{S,I,W,L,W_H,L_H,T\}}$ be a reward function that we want to parameterize such that $G' \coloneqq \mathbf{\Gamma} \cdot R'$ ends up in the ambiguity; here, R' is interpreted as a column vector.

We want $\mathbf{B} \cdot G' = 0$. Since the observation sequences $\vec{o} = o_\emptyset o_\emptyset o_\emptyset o_\emptyset$, $\vec{o} = o_\emptyset o_L o_\emptyset o_\emptyset$, $\vec{o} = o_\emptyset o_I o_\emptyset o_\emptyset$, or $\vec{o} = o_\emptyset o_I o_L o_\emptyset$ all cannot involve the states W or W_H, it is clear that they have zero expected return $(\mathbf{B} \cdot G')(\vec{o})$. Set $p'_H \coloneqq B\big(SIW_HT \mid o_\emptyset o_I o_W o_\emptyset\big)$. Then the condition that $\mathbf{B} \cdot G' = 0$ is equivalent to:

$$\begin{aligned}0 = \big(\mathbf{B} \cdot G'\big)(o_\emptyset o_I o_W o_\emptyset) &= \mathop{\mathbf{E}}_{\vec{s} \sim B(\vec{s} \mid o_\emptyset o_I o_W o_\emptyset)} \big[G'(\vec{s})\big] \\ &= p'_H \cdot G'(SIW_HT) + (1 - p'_H) \cdot G'(SIWT) = p'_H \cdot R'(W_H) + (1 - p'_H) \cdot R'(W).\end{aligned}$$

Thus, if $R'(W) = \frac{p'_H}{p'_H - 1} R'(W_H)$, then $G' \in \ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma}$, meaning that $R + R'$ has the same choice probabilities as R and is thus fully feedback-compatible. In particular, if $R'(W_H) \gg 0$ is sufficiently large, then in subsequent policy optimization, there is an incentive to hide the mistakes and π_H will be selected, which is suboptimal with respect to the true reward function R.

Thus Fig. 4A *still retains dangerous ambiguity when modeling partial observability.*

However, the example in Fig. 4B leads to no ambiguity when partial observability is correctly modeled.

To show this in detail, let $G' = \mathbf{\Gamma}(R') \in \ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma}$. We need to show $G' = 0$. Since the human is only uncertain about the state sequences corresponding to the observation sequence $o_\emptyset o_I o_\emptyset o_\emptyset$, the condition $\mathbf{B} \cdot G' = 0$ already implies $G'(\vec{s}) = 0$ for all state sequences except $SIWT$ and $SITT$. From $(\mathbf{B} \cdot G')(o_\emptyset o_I o_\emptyset o_\emptyset) = 0$, one then obtains the equation

$$(1 - p_D) \cdot \big(R'(S) + R'(I) + 2R'(T)\big) + p_D \cdot \big(R'(S) + R'(I) + R'(W) + R'(T)\big) = 0. \quad \text{(C.1)}$$

Table C.12: Experiments showing improved performance of po-aware RLHF

Ex.	p	p_{hide}	p_{default}	**model**	**action**	$\overline{E}^+$	**dec. infl.**	$\overline{E}^-$	**overj.**	**optimal**
A	0.5	0.5	N/A	naive	a_H	1.5	✓	0	×	×
A	0.5	0.5	N/A	po-aware	a_H	1.5	✓	0	×	×
A	0.1	0.9	N/A	naive	a_C	0	×	0	✓	×
A	0.1	0.9	N/A	po-aware	a_T	0	×	5.4	×	✓
B	0.5	N/A	0.9	naive	a_T	4.5	✓	0	✓	×
B	0.5	N/A	0.9	po-aware	a_D	0	×	0.25	×	✓
B	0.5	N/A	0.1	naive	a_V	0	×	0	✓	×
B	0.5	N/A	0.1	po-aware	a_D	0	×	2.25	×	✓

Thus, if one of the two state sequences involved has zero return, then the other has as well, assuming that $0 \neq p_D \neq 1$, and we are done.

To show this, we use that all other state sequences have zero return: $R'(S) + 3R'(T) = 0 = R'(S) + R'(L) + 2R'(T)$, from which $R'(L) = R'(T)$ follows. Then, from $R'(S) + R'(I) + R'(L) + R'(T) = 0$, substituting the previous result gives $R'(S) + R'(I) + 2R'(T) = 0$, and so Equation (C.1) results in $R'(S) + R'(I) + R'(W) + R'(T) = 0$. Overall, this shows $G' = \mathbf{\Gamma}(R') = 0$, and so $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$.

Experimental details

Here, we explain more experimental details for the results in Table 4.1, reproduced here as Table C.12, and Figure 4.5.

The leftmost column ("Ex." for "example") corresponds to Examples A and B in Figure 4.4. p is the success probability upon attempting to install Cuda or NumPy in state I, see Figure C.1. p_{hide} in Example A is the human's belief probability that the agent hid the error message if there is no output after nvidia-driver installation. Similarly, p_{default} in Example B is the human's belief probability that installation was done with default settings if there is no further output after Python installation. Note that lines one and two in the table also correspond to Example C.1, lines three and four to Example C.2, and lines five and six to the first half and seven and eight to the second half of Example C.3, respectively. In all the results *in the table*, we set the penalty to $r = 1$.

The "model" column has value "naive" if the reward learning algorithm is classical RLHF (erroneously assuming full observability) as in P. F. Christiano et al., 2017, and "po-aware" if the human's partial observability is correctly modeled as in Appendix C.3. We initialize the reward function as a list of rewards of states and train it by logistic regression using a dataset that consists of all pairs of state sequences together with the human's choice probabilities

under partial observations. This leads to 28 pairs of distinct trajectories together with choice probabilities. We train the reward model for 300 epochs over a shuffled dataset of 13.5 copies of the 28 pairs with the Adam optimizer, for a total of 113400 training updates.

Once we have the resulting reward model, we use value iteration to find its deterministic optimal policy. All policies choose to install the nvidia-driver (in Example A) and Python (in Example B), and differ in their action in state I, which is given in the column "action". We compute the overestimation error and underestimation error of the resulting policies analytically using the hardcoded environment dynamics, true reward function, observation function, and human belief matrix $\mathbf{B}$. This is given in columns $\overline{E}^+$ and $\overline{E}^-$. Note that these are averages over 10 entire training runs, though since they always result in the same learned policy, there is no variation and we do not state any uncertainty.

The columns "dec. infl.", "overj.", and "optimal" state whether deceptive inflation or overjustification occurs with the learned policy, and whether it is optimal according to the true human's reward function.

For the results in Figure 4.5, we use largely the same procedure as for the table. Instead of fixing the reward penalty r or the belief probabilities p_{hide} and p_{default}, we vary them as hyperparameters for the plots, we fix p to $p = 0.5$, and we restrict ourselves to the analysis of "naive" RLHF.

C.3 Modeling the human in partially observable RLHF

Here, we develop the theory of RLHF with appropriately modeled partial observability, including full proofs of all theorems.

First, we explain how the human can arrive at the belief $B(\vec{s} \mid \vec{o})$ via Bayesian updates. The main theory and the main text in general do not depend on this specific form of the human's belief, but some examples in the appendix do.

Second, we explain our main result: the ambiguity and identifiability of both reward and return functions under observed sequence comparisons. Then, we explain that this theorem means that one could *in principle* design a practical reward learning algorithm that converges on the correct reward function up to the ambiguity characterized in the section before, *if* the human's belief kernel $B(\vec{s} \mid \vec{o})$ is fully known.

Third, we generalize the theory to the case that the human's observations are not necessarily known to the learning system and again characterize precisely when the return function is identifiable from sequence comparisons. We then consider special cases: we show that the fully observable case is covered by our theory, that a deterministic observation kernel $P_{\vec{O}}$ usually leads to non-injective belief matrix $\mathbf{B}$, and that "noise" in the observation kernel $P_{\vec{O}}$ leads, under appropriate assumptions, to the identifiability of the return function.

Our identifiability results require that the learning system knows the human's belief kernel $B(\vec{s} \mid \vec{o})$. Nevertheless, we show that these results are robust to slight misspecifications: a

bound in the error in the specified belief leads to a corresponding bound in the error of the policy evaluation function used for subsequent reinforcement learning.

Next, we provide a very preliminary characterization of the ambiguity in the return function under special cases.

Finally, we study examples of identifiability and non-identifiability of the return function for the case that we *do* model the human's partial observability correctly. This reveals qualitatively interesting cases of identifiability, even when $\mathbf{B}$ is not injective, and catastrophic cases of non-identifiability.

The belief over the state sequence for rational humans

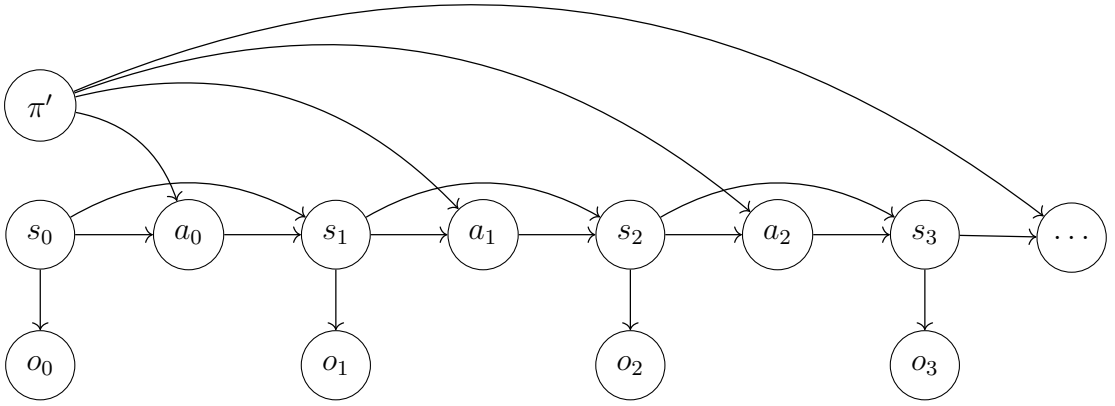
Before we dive into the main theory, we want to explain how the human can iteratively compute the posterior of the state sequence given an observation sequence with successively new observations. This is done by defining a Bayesian network for the joint probability of policy, states, actions, and observations, and doing Bayesian inference over this Bayesian network.

The details of this subsection are only relevant for a few sections in the appendix since it is usually enough to assume that the posterior belief *exists*. Additionally, in the core theory, we do not even assume that $B(\vec{s} \mid \vec{o})$ is a posterior: it is simply any probability distribution. The reason why it can still be interesting to analyze the case when the human is a rational Bayesian reasoner is that one can then analyze RLHF under *generous* assumptions to the human.

We model the human to have a joint distribution $B(\pi, \vec{s}, \vec{a}, \vec{o})$ over the policy π, state sequence $\vec{s} = s_0, \dots, s_T$, action sequence $\vec{a} = a_0, \dots, a_{T-1}$, and observation sequence $\vec{o} = o_0, \dots, o_T$. This is given by a Bayesian network with the following components:

- a policy prior $B(\pi')$;
- the probability of the initial state $B(s_0) \coloneqq P_0(s_0)$;
- action probabilities $B(a \mid s, \pi) \coloneqq \pi(a \mid s)$;
- transition probabilities $B(s_{t+1} \mid s_t, a_t) \coloneqq \mathcal{T}(s_{t+1} \mid s_t, a_t)$;
- and observation probabilities $B(o_t \mid s_t) \coloneqq P_O(o_t \mid s_t)$.

Together, this defines the joint distribution $B(\pi, \vec{s}, \vec{a}, \vec{o})$ over the policy, states, actions, and observations that factorizes according to the following directed acyclic graph:



(C.2)

The following proposition clarifies the iterative Bayesian update of the human's posterior over state sequences, given observation sequences:

Proposition C.4. *Let $t \leq T-1$ and denote by $\hat{s} = s_0, \dots, s_t$ a state sequence* segment *of length $t \geq 0$. Similarly, $\hat{o} = o_0, \dots, o_t$ denotes an observation sequence segment. We have*

$$B(\hat{s}, s_{t+1}, \pi \mid \hat{o}, o_{t+1}) \propto P_O(o_{t+1} \mid s_{t+1}) \cdot \left[\sum_{a_t \in \mathcal{A}} \mathcal{T}(s_{t+1} \mid \hat{s}_t, a_t) \cdot \pi(a_t \mid s_t) \right] \cdot B(\hat{s}, \pi \mid \hat{o}).$$

Thus, the human can iteratively compute $B(\hat{s}, \pi \mid \hat{o})$ from the prior $B(s_0, \pi) = P_0(s_0) \cdot B(\pi')$ using the above Bayesian update.

The posterior over the state sequence can subsequently be computed by

$$B(\hat{s} \mid \hat{o}) = \int_{\pi} B(\hat{s}, \pi \mid \hat{o}).$$

Proof. The proof is essentially just Bayes rule applied to the Bayesian network in Equation (C.2). We repeatedly make use of conditional independences that follow from d-separations in the graph (Geiger, Verma, and Pearl, 1990). More concretely, we have

$$\begin{aligned}
B\big(\hat{s}, s_{t+1}, \pi \,|\, \hat{o}, o_{t+1}\big) &\propto B\big(o_{t+1} \,|\, \hat{s}, s_{t+1}, \pi, \hat{o}\big) \cdot B\big(\hat{s}, s_{t+1}, \pi \,|\, \hat{o}\big) \\
&= P_O\big(o_{t+1} \,|\, s_{t+1}\big) \cdot B\big(s_{t+1} \,|\, \hat{s}, \pi, \hat{o}\big) \cdot B\big(\hat{s}, \pi \,|\, \hat{o}\big) \\
&= P_O\big(o_{t+1} \,|\, s_{t+1}\big) \cdot \left[\sum_{a_t \in \mathcal{A}} B\big(s_{t+1} \,|\, a_t, \hat{s}, \pi, \hat{o}\big) \cdot B\big(a_t \,|\, \hat{s}, \pi, \hat{o}\big) \right] \cdot B\big(\hat{s}, \pi \,|\, \hat{o}\big) \\
&= P_O\big(o_{t+1} \,|\, s_{t+1}\big) \cdot \left[\sum_{a_t \in \mathcal{A}} \mathcal{T}\big(s_{t+1} \,|\, s_t, a_t\big) \cdot \pi\big(a_t \,|\, s_t\big) \right] \cdot B\big(\hat{s}, \pi \,|\, \hat{o}\big).
\end{aligned}$$

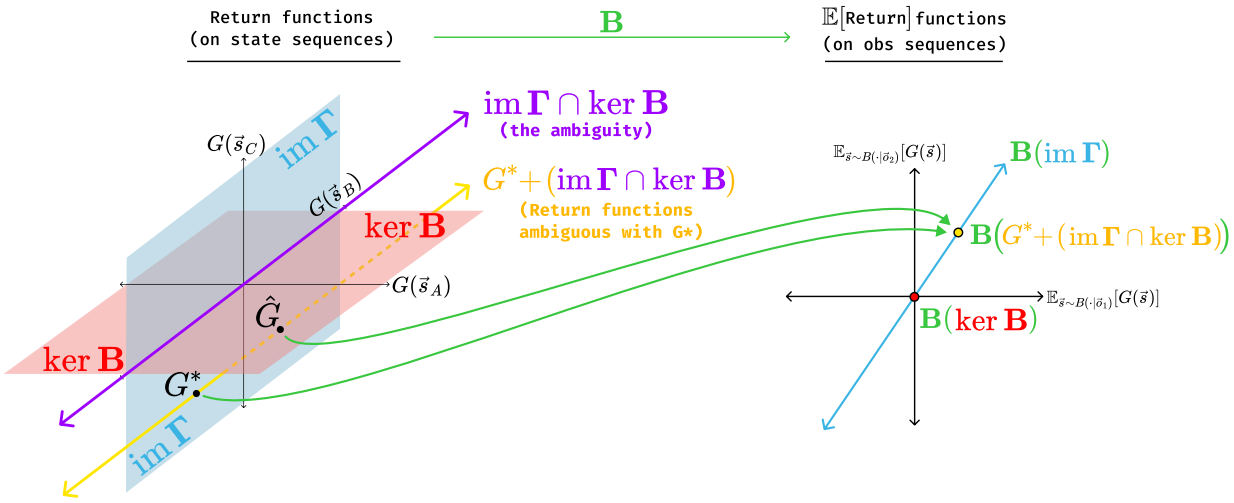


Figure C.3: The linear geometry of ambiguity for a hypothetical example with three state sequences and two observation sequences. G^* is the true return function, and "G" is used in labeling the axes to refer to some arbitrary return function. This is a more accurate geometric depiction of the middle and right spaces in Figure 4.6. The subspace $\operatorname{im}\boldsymbol{\Gamma} \cap \ker \mathbf{B}$ (purple) is the ambiguity in return functions, meaning that adding an element would not change the human's expected return function on observations. Thus the set of return functions that the reward learning system can infer is the affine set $G + (\operatorname{im}\boldsymbol{\Gamma} \cap \ker \mathbf{B})$ (yellow). Note that the planes on the left are drawn to be axis-aligned for ease of visualization; this will not be the case for real MDPs.

In step 1, we used Bayes rule. In step 2, we made use of the independence $o_{t+1} \perp\!\!\!\perp (\hat{s}, \pi, \hat{o}) \mid s_{t+1}$, plugged in the observation kernel, and used the chain rule of probability to compose the second term into a product. In step 3, we marginalized and used, once again, the chain rule of probability. In step 4, we used the independences $s_{t+1} \perp\!\!\!\perp (s_0, \ldots, s_{t-1}, \pi, \hat{o}) \mid (s_t, a)$ and $a_t \perp\!\!\!\perp (s_0, \ldots, s_{t-1}, \hat{o}) \mid (\pi, s_t)$ and plugged in the transition kernel and the policy.

The last formula is just a marginalization over the policy. □

Ambiguity and identifiability of reward and return functions under observation sequence comparisons

In this section, we prove the main theorem of Chapter 4: a characterization of the ambiguity that is left in the reward and return function once the human's Boltzmann-rational choice probabilities are known. We change the formulation slightly by formulating the linear operators "intrinsically" in the spaces they are defined in, instead of using matrix versions. This does not change the general picture, but is a more natural setting when thinking, e.g., about generalizing the results to infinite state sequences. Thus, we define $\mathbf{B} : \mathbb{R}^{\vec{\mathcal{S}}} \to \mathbb{R}^{\vec{\Omega}}$ as

the linear operator given by

$$\big[\mathbf{B}(G)\big](\vec{o}) \coloneqq \mathop{\mathbf{E}}_{\vec{s}\sim B(\vec{s}|\vec{o})}\big[G(\vec{s})\big].$$

Here, $\mathbf{B}$ is the human's belief, which can either be computed as in the previous subsection or simply be any conditional probability distribution. Similarly, we define $\mathbf{\Gamma} : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\vec{\mathcal{S}}}$ as the linear operator given by

$$\big[\mathbf{\Gamma}(R)\big](\vec{s}) \coloneqq \sum_{t=0}^{T} \gamma^t R(s_t).$$

The matrix product $\mathbf{B}\cdot\mathbf{\Gamma}$ then becomes the composition $\mathbf{B}\circ\mathbf{\Gamma} : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\vec{\Omega}}$. Finally, recall that the kernel $\ker \mathbf{A}$ of a linear operator $\mathbf{A}$ is defined as its nullspace, and the image $\operatorname{im}\mathbf{A}$ as the set of elements hit by $\mathbf{A}$. We obtain the following theorem:

Theorem C.5. *Let R be the true reward function and $\tilde{R}$ another reward function. Let $\tilde{G} = \mathbf{\Gamma}(\tilde{R})$ and $G = \mathbf{\Gamma}(R)$ be the corresponding return functions. The following three statements are equivalent:*

(i) The reward function $\tilde{R}$ gives rise to the same vector of choice probabilities as R, i.e

$$\Big(P^{\tilde{R}}\big(\vec{o} \succ \vec{o}'\big)\Big)_{\vec{o},\vec{o}'\in\vec{\Omega}} = \Big(P^{R}\big(\vec{o} \succ \vec{o}'\big)\Big)_{\vec{o},\vec{o}'\in\vec{\Omega}}.$$

(ii) There is a reward function $R' \in \ker(\mathbf{B}\circ\mathbf{\Gamma})$ and a constant $c \in \mathbb{R}$ such that

$$\tilde{R} = R + R' + c.$$

(iii) There is a return function $G' \in \ker\mathbf{B} \cap \operatorname{im}\mathbf{\Gamma}$ and a constant $c' \in \mathbb{R}$ such that

$$\tilde{G} = G + G' + c'.$$

In other words, the ambiguity *that is left in the* reward function *when its observation-based choice probabilities are known is, up to an additive constant, given by* $\ker(\mathbf{B}\circ\mathbf{\Gamma})$*; the ambiguity left in the* return function *is given by* $\ker\mathbf{B} \cap \operatorname{im}\mathbf{\Gamma}$.

Proof. Assume (i). To prove (ii), let σ by the sigmoid function given by $\sigma(x) = \frac{1}{1+\exp(-x)}$. Then by Equation (4.2), the equality of choice probabilities means the following for all $\vec{o}, \vec{o}' \in \vec{\Omega}$:

$$\sigma\Big(\beta\cdot\big(\big[\mathbf{B}(\tilde{G})\big](\vec{o}) - \big[\mathbf{B}(\tilde{G})\big](\vec{o}')\big)\Big) = \sigma\Big(\beta\cdot\big(\big[\mathbf{B}(G)\big](\vec{o}) - \big[\mathbf{B}(G)\big](\vec{o}')\big)\Big).$$

Since the sigmoid function is injective, this implies

$$\big[\mathbf{B}(\tilde{G})\big](\vec{o}) - \big[\mathbf{B}(\tilde{G})\big](\vec{o}') = \big[\mathbf{B}(G)\big](\vec{o}) - \big[\mathbf{B}(G)\big](\vec{o}').$$

Fixing an arbitrary $\vec{o}'$, this implies that there exists a constant c' such that for all $\vec{o} \in \vec{\Omega}$, the following holds:

$$\big[\mathbf{B}(\tilde{G})\big](\vec{o}) - \big[\mathbf{B}(G)\big](\vec{o}') - c' = 0.$$

Noting that $\mathbf{B}(c') = c'$, this implies $\tilde{G} - G - c' \in \ker(\mathbf{B})$. Now, define the constant reward function

$$c \coloneqq c' \cdot \frac{1-\gamma}{1-\gamma^{T+1}}.$$

We obtain

$$\begin{aligned}\big[\mathbf{\Gamma}(c)\big](\vec{s}) &= \sum_{t=0}^{T} \gamma^t \cdot c \\ &= c' \cdot \frac{1-\gamma}{1-\gamma^{T+1}} \cdot \sum_{t=0}^{T} \gamma^t \\ &= c'.\end{aligned}$$

Thus, we have

$$\mathbf{\Gamma}(\tilde{R} - R - c) = \tilde{G} - G - c' \in \ker(\mathbf{B}),$$

implying $R' \coloneqq \tilde{R} - R - c \in \ker(\mathbf{B} \circ \mathbf{\Gamma})$. This shows (ii).

That (ii) implies (iii) follows by applying $\mathbf{\Gamma}$ to both sides of the equation.

Now assume (iii), i.e. $\tilde{G} = G + G' + c'$ for a constant $c' \in \mathbb{R}$ and a return function $G' \in \ker(\mathbf{B}) \cap \operatorname{im} \mathbf{\Gamma}$. This implies $\mathbf{B}(\tilde{G}) = \mathbf{B}(G) + c'$. Thus, for all $\vec{o}, \vec{o}' \in \vec{\Omega}$, we have

$$\big[\mathbf{B}(\tilde{G})\big](\vec{o}) - \big[\mathbf{B}(\tilde{G})\big](\vec{o}') = \big[\mathbf{B}(G)\big](\vec{o}) - \big[\mathbf{B}(G)\big](\vec{o}'),$$

which implies the equal choice probabilities after multiplying with β and applying the sigmoid function σ on both sides. Thus, (iii) implies (i). □

Corollary C.6. *The following two statements are equivalent:*

(i) $\ker(\mathbf{B} \circ \mathbf{\Gamma}) = 0$.

(ii) The data $\Big(P^R\big(\vec{o} \succ \vec{o}'\big)\Big)_{\vec{o},\vec{o}' \in \vec{\Omega}}$ *determine the reward function* R *up to an additive constant.*

Proof. That (i) implies (ii) follows immediately from the implication from (i) to (ii) within the preceding theorem.

Now assume (ii). Let $R' \in \ker(\mathbf{B} \circ \mathbf{\Gamma})$. Define $\tilde{R} \coloneqq R + R'$. Then the implication from (ii) to (i) within the preceding theorem implies that $\tilde{R}$ and R have the same choice probabilities. Thus, the assumption (ii) in this corollary implies that R' is a constant. Since $\mathbf{\Gamma}$ and $\mathbf{B}$ map nonzero constants to nonzero constants, the fact that $R' \in \ker(\mathbf{B} \circ \mathbf{\Gamma})$ implies that $R' = 0$, showing that $\ker(\mathbf{B} \circ \mathbf{\Gamma}) = \{0\}$. □

As mentioned in the main text, the previous result already leads to the non-identifiability of R whenever $\mathbf{\Gamma}$ is not injective, corresponding to the presence of zero-initial potential shaping (Skalse, Farrugia-Roberts, et al. (2023), Lemma B.3). Thus, we now strengthen the previous result so that it deals with the identifiability of the *return* function, which is sufficient for the purpose of policy optimization:

Corollary C.7. *Consider the following four statements (which can each be true or false):*

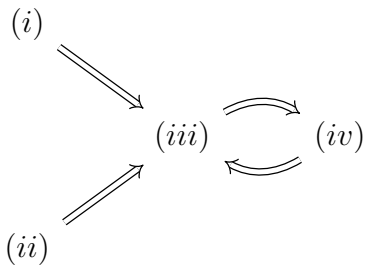
(i) $\ker \mathbf{B} = \{0\}$.

(ii) $\ker\big(\mathbf{B} \circ \mathbf{\Gamma}\big) = \{0\}$.

(iii) $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$.

(iv) The data $\Big(P^R\big(\vec{o} \succ \vec{o}'\big)\Big)_{\vec{o},\vec{o}' \in \vec{\Omega}}$ *determine the return function* $G = \mathbf{\Gamma}(R)$ *on sequences* $\vec{s} \in \vec{\mathcal{S}}$ *up to a constant independent of* $\vec{s}$.

Then the following implications, and no other implications, are true:



In particular, all of (i), (ii), and (iii) are sufficient conditions for identifying the return function from the choice probabilities.

Proof. That (i) implies (iii) is trivial. That (ii) implies (iii) is a simple linear algebra fact: Assume (ii) and that $G' \in \ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma}$. Then $G' = \mathbf{\Gamma}(R')$ for some $R' \in \mathbb{R}^{\mathcal{S}}$ and

$$0 = \mathbf{B}(G') = \mathbf{B}\left(\mathbf{\Gamma}(R')\right) = (\mathbf{B} \circ \mathbf{\Gamma})(R').$$

By (ii), this implies $R' = 0$ and therefore $G' = \mathbf{\Gamma}(R') = 0$, showing (iii).

That (iii) implies (iv) immediately follows from the implication from (i) to (iii) in Theorem C.5.

Now, assume (iv). To prove (iii), assume $G' \in \ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma}$. Then the implication from (iii) to (i) in Theorem C.5 implies that $G + G'$ induces the same observation-based choice probabilities as G. Thus, (iv) implies $G + G' = G + c'$ for some constant c', which implies $G' = c'$. Since $G' \in \ker \mathbf{B}$, this implies $0 = \mathbf{B}(G') = \mathbf{B}(c') = c'$ and thus $G' = 0$. Thus, we showed $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$.

We now show that no other implication holds in general. Example C.35 will show that (ii) does not imply (i). We now show that (i) does also not imply (ii), from which it will logically follow that (iii) does neither imply (i) nor (ii). Namely, consider the following simple MDP with time horizon $T = 1$:

$$a \longrightarrow b \tag{C.3}$$

In this MDP, every state sequence starts in a, deterministically transitions to b, and then ends. This means that $\vec{s} = ab$ is the only sequence. Now, let $R' \in \mathbb{R}^{\{a,b\}}$ be the reward function given by

$$R'(a) = 1, \quad R'(b) = \frac{-1}{\gamma}.$$

We obtain

$$\big[\mathbf{\Gamma}(R')\big](\vec{s}) = R'(a) + \gamma R'(b) = 1 + \gamma \cdot \frac{-1}{\gamma} = 0.$$

Thus, $\mathbf{\Gamma}(R') = 0$, $(\mathbf{B} \circ \mathbf{\Gamma})(R') = 0$, and, therefore, $\ker\big(\mathbf{B} \circ \mathbf{\Gamma}\big) \neq \{0\}$. Thus, (ii) does not hold. However, it is possible to choose $B(\vec{s} \mid \vec{o})$ such that (i) holds: e.g., if $\Omega = \mathcal{S}$ and $B(\vec{s} \mid \vec{o}) \coloneqq \delta_{\vec{o}}(\vec{s})$, then $\ker \mathbf{B} = \{0\}$ since this operator is the identity. □

The ambiguity in reward learning in practice

In this section, we point out that Theorem C.5 is not just a theoretical discussion: When $\mathbf{B}$ and the inverse temperature parameter β are known, then it is possible to design a reward learning algorithm that learns the true reward function up to the ambiguity $\ker(\mathbf{B} \circ \mathbf{\Gamma})$ in the infinite data limit. In doing so, we essentially use the loss function proposed in P. F. Christiano et al. (2017).

Namely, assume $\mathcal{D}$ is a data distribution of observation sequences $\vec{o} \in \vec{\Omega}$ such that all sequences in $\vec{\Omega}$ have a strictly positive probability of being sampled; for example, $\mathcal{D}$ could use an exploration policy and the observation sequence kernel $P_{\vec{O}}$. For each pair of observation sequences $(\vec{o}, \vec{o}')$, we then get a conditional distribution $P(\mu \mid \vec{o}, \vec{o}')$ over a one-hot encoded human choice $\mu \in \{(1,0),(0,1)\}$, with probability

$$P\big(\mu = (1,0) \mid \vec{o}, \vec{o}'\big) = P^R\big(\vec{o} \succ \vec{o}'\big).$$

Together, this gives rise to a dataset $(\vec{o}_1, \vec{o}'_1, \mu_1), \ldots, (\vec{o}_N, \vec{o}'_N, \mu_N)$ of observation sequences plus a human choice.

Now assume we learn a reward function $R_\theta : \mathcal{S} \to \mathbb{R}$ that is differentiable in the parameter θ and that can represent all possible reward functions $R \in \mathbb{R}^{\mathcal{S}}$. Let $G_\theta \coloneqq \mathbf{\Gamma}(R_\theta)$ be the corresponding return function. Write $\mu_k = (\mu_k^{(1)}, \mu_k^{(2)})$. As in P. F. Christiano et al. (2017), we define its loss over the dataset above by

$$\widetilde{\mathcal{L}}(\theta) = -\frac{1}{N} \sum_{k=1}^{N} \mu_k^{(1)} \cdot \log P^{R_\theta}\big(\vec{o}_k \succ \vec{o}'_k\big) + \mu_k^{(2)} \cdot \log P^{R_\theta}\big(\vec{o}'_k \succ \vec{o}_k\big).$$

Note that by Equation (4.2), this loss function essentially uses $\mathbf{B}$ and also the inverse temperature parameter β in its definition. This means that these need to be explicitly represented to be able to use the loss function in practice.

Proposition C.8. *The loss function $\widetilde{\mathcal{L}}$ is differentiable. Furthermore, in the infinite datalimit its minima are precisely given by parameters θ such that $R_\theta = R + R' + c$ for $R' \in \ker\big(\mathbf{B} \circ \mathbf{\Gamma}\big)$ and $c \in \mathbb{R}$, or equivalently $G_\theta = G + G' + c'$ for $G' \in \ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma}$ and $c' \in \mathbb{R}$.*

Proof. The differentiability of the loss function follows from the differentiability of multiplication with the matrix $\mathbf{B}$, see Equation (4.2), and of the reward function R_θ in its parameter θ that we assumed.

For the second statement, let $N(\vec{o}, \vec{o}')$ be the number of times that the pair $(\vec{o}, \vec{o}')$ appears in the dataset, and let $N(\vec{o}, \vec{o}', 1)$ be the number of times that the human choice is $\mu = (1, 0)$ and the sampled pair is $(\vec{o}, \vec{o}')$, and similar for 2 instead of 1. We obtain

$$\begin{aligned}
\widetilde{\mathcal{L}}(\theta) = &- \sum_{\vec{o}, \vec{o}' \in \vec{\Omega}} \frac{N(\vec{o}, \vec{o}')}{N} \cdot \left[\frac{N(\vec{o}, \vec{o}', 1)}{N(\vec{o}, \vec{o}')} \log P^{R_\theta}\big(\vec{o} \succ \vec{o}'\big) \right. \\
&\left. + \frac{N(\vec{o}, \vec{o}', 2)}{N(\vec{o}, \vec{o}')} \log P^{R_\theta}\big(\vec{o}' \succ \vec{o}\big) \right] \\
\approx &\mathop{\mathbf{E}}_{\vec{o}, \vec{o}' \sim \mathcal{D}} \left[\mathrm{CE}\left(P^{R}\big(\vec{o} \preceq \vec{o}'\big) \parallel P^{R_\theta}\big(\vec{o} \preceq \vec{o}'\big) \right) \right] \\
=: &\mathcal{L}(\theta).
\end{aligned}$$

Here, CE is the crossentropy between the two binary distributions. Since we assumed that $\mathcal{D}$ gives a positive probability to all observation sequences in $\vec{\Omega}$, and since the cross entropy is generally minimized exactly when the second distribution equals the first, the loss function $\mathcal{L}(\theta)$ is minimized if and only if R_θ gives rise to the same choice probabilities as R for all pairs of observation sequences. Theorem C.5 then gives the result. □

Identifiability of return functions when human observations are not known

Corollary C.7 assumes that the choice probabilities of each observation sequence pair are known to the reward learning algorithm. However, this requires the algorithm to know what the human observed. In some applications, this is a reasonable assumption, e.g. if the human's observations are themselves produced by an algorithm that can feed the observations also back to the learning algorithm. In general, however, the observations happen in the physical world, and are only known probabilistically via the observation kernel P_O. The

learning system *does* however have access to the full state sequences that generate the observation sequences. This leads to knowledge of the following choice probabilities for $\vec{s}, \vec{s}' \in \vec{\mathcal{S}}$:

$$P^R\big(\vec{s} \succ \vec{s}'\big) \coloneqq \mathop{\mathbf{E}}_{\vec{o},\vec{o}' \sim P_{\vec{O}}(\cdot|\vec{s},\vec{s}')} \Big[P^R\big(\vec{o} \succ \vec{o}'\big)\Big],^1 \tag{C.4}$$

where the observation-based choice probabilities are given as in Equation (4.2). In other words, the learning algorithm can only infer an aggregate of the observation-based choice probabilities. Again, we can ask a question similar to the ones before, extending the investigations in the previous section:

Question C.9. *Assume the vector of choice probabilities* $\Big(P^R(\vec{s} \succ \vec{s}')\Big)_{\vec{s},\vec{s}' \in \vec{\mathcal{S}}}$ *is known. Additionally, assume that it is known that the human's observations are governed by* P_O*, and that the human is Boltzmann rational with inverse temperature parameter* β *and beliefs* $B(\vec{s} \mid \vec{o})$*, see Equation* (C.4)*. Does this data identify the return function* $G : \vec{\mathcal{S}} \to \mathbb{R}$*?*

If the observation-based choice probabilities from Equation (4.2) would be known, then Corollary C.7 would provide the answer to this question. Thus, similar to how we previously inverted the belief operator $\mathbf{B}$, we are now simply tasked with inverting the expectation over observation sequences. This leads us to the following definition:

Definition C.10 (Ungrounding Operator)**.** *The* ungrounding operators $\mathbf{O} : \mathbb{R}^{\vec{\Omega}} \to \mathbb{R}^{\vec{\mathcal{S}}}$ *and* $\mathbf{O} \otimes \mathbf{O} : \mathbb{R}^{\vec{\Omega} \times \vec{\Omega}} \to \mathbb{R}^{\vec{\mathcal{S}} \times \vec{\mathcal{S}}}$ *are defined by*

$$\big[\mathbf{O}(v)\big](\vec{s}) \coloneqq \mathop{\mathbf{E}}_{\vec{o} \sim P_{\vec{O}}(\vec{o}|\vec{s})} \big[v(\vec{o})\big], \quad \big[(\mathbf{O} \otimes \mathbf{O})(C)\big](\vec{s}, \vec{s}') \coloneqq \mathop{\mathbf{E}}_{\vec{o},\vec{o}' \sim P_{\vec{O}}(\cdot|\vec{s},\vec{s}')} \big[C(\vec{o}, \vec{o}')\big].$$

Here, $v \in \mathbb{R}^{\vec{\Omega}}$ *is an arbitrary vector, and* $C \in \mathbb{R}^{\vec{\Omega} \times \vec{\Omega}}$ *is also an arbitrary vector, where the notation can remind of "Choice" since the inputs to* $\mathbf{O} \otimes \mathbf{O}$ *are, in practice, vectors of observation-based Boltzmann-rational choice probabilities.*

Formally, $\mathbf{O} \otimes \mathbf{O}$ is the Kronecker product of $\mathbf{O}$ with itself, but it is not necessary to understand this fact to follow the discussion. Ultimately, to be able to recover the observation-based choice probabilities, what matters is that $\mathbf{O} \otimes \mathbf{O}$ is injective on whole vectors of these choice probabilities. The injectivity of $\mathbf{O}$ is a *sufficient condition* for this, which explains its usefulness. We show this in the following lemma:

Lemma C.11. $\mathbf{O} : \mathbb{R}^{\vec{\Omega}} \to \mathbb{R}^{\vec{\mathcal{S}}}$ *is injective if and only if* $\mathbf{O} \otimes \mathbf{O} : \mathbb{R}^{\vec{\Omega} \times \vec{\Omega}} \to \mathbb{R}^{\vec{\mathcal{S}} \times \vec{\mathcal{S}}}$ *is injective.*

Proof. This is a general property of the Kronecker product of a linear operator with itself. For completeness, we demonstrate the calculation in our special case. First, assume that $\mathbf{O}$ is injective. Assume that $(\mathbf{O} \otimes \mathbf{O})(C) = 0$ for some $C \in \mathbb{R}^{\vec{\Omega} \times \vec{\Omega}}$. We need to show $C = 0$.

[1] We excuse the following abuse of notation: these choice probabilities *run through* the observations of the human and are not the same as the choice probabilities from Equation (4.1).

For all pairs of state sequences $(\vec{s}, \vec{s}')$, we have

$$\begin{aligned}
0 = \big[(\mathbf{O} \otimes \mathbf{O})(C)\big](\vec{s}, \vec{s}') &= \mathop{\mathbf{E}}_{\vec{o},\vec{o}' \sim P_{\vec{O}}(\cdot|\vec{s},\vec{s}')} \big[C(\vec{o}, \vec{o}')\big] \\
&= \mathop{\mathbf{E}}_{\vec{o} \sim P_{\vec{O}}(\vec{o}|\vec{s})} \left[\mathop{\mathbf{E}}_{\vec{o}' \sim P_{\vec{O}}(\vec{o}'|\vec{s}')} \big[C(\vec{o}, \vec{o}')\big] \right] \\
&= \mathop{\mathbf{E}}_{\vec{o} \sim P_{\vec{O}}(\vec{o}|\vec{s})} \Big[C'_{\vec{s}'}(\vec{o})\Big] \\
&= \Big[\mathbf{O}\left(C'_{\vec{s}'}\right)\Big](\vec{s}),
\end{aligned}$$

where $C'_{\vec{s}'}(\vec{o}) \coloneqq \mathbf{E}_{\vec{o}' \sim P_{\vec{O}}(\vec{o}'|\vec{s}')} \big[C(\vec{o}, \vec{o}')\big]$. By the injectivity of $\mathbf{O}$, we obtain $C'_{\vec{s}'} = 0$ for all $\vec{s}'$. This means that for all $\vec{s}'$ and $\vec{o}$, we have

$$0 = C'_{\vec{s}'}(\vec{o}) = \mathop{\mathbf{E}}_{\vec{o}' \sim P_{\vec{O}}(\vec{o}'|\vec{s}')} \big[C(\vec{o}, \vec{o}')\big] = \Big[\mathbf{O}\left(C''_{\vec{o}}\right)\Big](\vec{s}'),$$

where $C''_{\vec{o}}(\vec{o}') \coloneqq C(\vec{o}, \vec{o}')$. Again, by the injectivity of $\mathbf{O}$, we obtain $C''_{\vec{o}} = 0$ for all $\vec{o}$, leading to $C = 0$. That proves the direction from left to right.

To prove the other direction, assume that $\mathbf{O}$ is *not* injective. This means there exists $0 \neq C \in \mathbb{R}^{\vec{\Omega}}$ such that $\mathbf{O}(C) = 0$. Define $C \otimes C \in \mathbb{R}^{\vec{\Omega} \times \vec{\Omega}}$ by

$$(C \otimes C)(\vec{o}, \vec{o}') \coloneqq C(\vec{o})C(\vec{o}').$$

Then clearly, $C \otimes C \neq 0$. We are done if we can show that $(\mathbf{O} \otimes \mathbf{O})(C \otimes C) = 0$ since that establishes that $\mathbf{O} \otimes \mathbf{O}$ is also not injective. For any $\vec{s}, \vec{s}' \in \vec{\mathcal{S}}$, we have

$$\begin{aligned}
\Big[(\mathbf{O} \otimes \mathbf{O})(C \otimes C)\Big](\vec{s}, \vec{s}') &= \mathop{\mathbf{E}}_{\vec{o},\vec{o}' \sim P_{\vec{O}}(\cdot|\vec{s},\vec{s}')} \Big[(C \otimes C)(\vec{o}, \vec{o}')\Big] \\
&= \mathop{\mathbf{E}}_{\vec{o},\vec{o}' \sim P_{\vec{O}}(\cdot|\vec{s},\vec{s}')} \Big[C(\vec{o}) \cdot C(\vec{o}')\Big] \\
&= \mathop{\mathbf{E}}_{\vec{o} \sim P_{\vec{O}}(\vec{o}|\vec{s})} \big[C(\vec{o})\big] \cdot \mathop{\mathbf{E}}_{\vec{o}' \sim P_{\vec{O}}(\vec{o}'|\vec{s}')} \big[C(\vec{o}')\big] \\
&= \big[\mathbf{O}(C)\big](\vec{s}) \cdot \big[\mathbf{O}(C)\big](\vec{s}') \\
&= 0 \cdot 0 \\
&= 0.
\end{aligned}$$

This finishes the proof. □

We now state and prove the following extension of Corollary C.7:

Theorem C.12. *Consider the following statements (which can each be true or false):*

1. $\mathbf{O} : \mathbb{R}^{\vec{\Omega}} \to \mathbb{R}^{\vec{\mathcal{S}}}$ *is an injective linear operator:* $\ker \mathbf{O} = \{0\}$.

2. $\mathbf{O} \otimes \mathbf{O} : \mathbb{R}^{\vec{\Omega} \times \vec{\Omega}} \to \mathbb{R}^{\vec{\mathcal{S}} \times \vec{\mathcal{S}}}$ *is an injective linear operator:* $\ker \mathbf{O} \otimes \mathbf{O} = \{0\}$.
3. $\mathbf{O} \otimes \mathbf{O}$ *is injective on vectors of observation-based choice probabilities* $\Big(P^R\big(\vec{o} \succ \vec{o}'\big)\Big)_{\vec{o},\vec{o}'}$ *over the set of return functions* $G \in \mathbb{R}^{\vec{\mathcal{S}}}$.
4. *The data of state-based choice probabilities* $\Big(P^R\big(\vec{s} \succ \vec{s}'\big)\Big)_{\vec{s},\vec{s}' \in \vec{\mathcal{S}}}$ *from Equation* (C.4) *determine the data of observation-based choice probabilities* $\Big(P^R\big(\vec{o} \succ \vec{o}'\big)\Big)_{\vec{o},\vec{o}' \in \vec{\Omega}}$ *from Equation* (4.2).

Then the following implications hold and 3 does not imply 2:

$$1 \rightleftharpoons 2 \implies 3 \implies 4.$$

Consequently, if any of the conditions 1, 2, or 3 hold, and additionally any of the conditions (i), (ii) or (iii) from Corollary C.7, then the data $\Big(P^R\big(\vec{s} \succ \vec{s}'\big)\Big)_{\vec{s},\vec{s}' \in \vec{\Omega}}$ *determine the return function* G *on sequences* $\vec{s} \in \vec{\mathcal{S}}$ *up to a constant independent of* $\vec{s}$.

Proof. That 1 and 2 are equivalent was shown in Lemma C.11. That 2 implies 3 is clear. To prove that 3 implies 4, simply put both sets of choice probabilities into a vector. Then Equation (C.4) and Definition C.10 show the following equality of vectors in $\mathbb{R}^{\vec{\mathcal{S}} \times \vec{\mathcal{S}}}$:

$$\Big(P^R\big(\vec{s} \succ \vec{s}'\big)\Big)_{\vec{s},\vec{s}'} = \big(\,\mathbf{O} \otimes \mathbf{O}\,\big)\Big(\Big(P^R\big(\vec{o} \succ \vec{o}'\big)\Big)_{\vec{o},\vec{o}'}\Big).$$

The injectivity of $\mathbf{O} \otimes \mathbf{O}$ on such inputs ensures that the observation-based choice probabilities can be recovered using this equation.

We now show that (3) does not imply (2). Again, we use the simple MDP from Equation (C.3), but this time with a different observation kernel. Namely, we choose

$$P_O(o^{(a)} \mid a) = P_O(o^{(a)\prime} \mid a) = \frac{1}{2}, \quad P_O(o^{(b)} \mid b) = 1,$$

where $o^{(a)\prime} \neq o^{(a)}$ and $o^{(a)} \neq o^{(b)} \neq o^{(a)\prime}$. This results in two possible observation sequences: $o^{(a)}o^{(b)}$ and $o^{(a)\prime}o^{(b)}$. Thus, $\mathbb{R}^{\vec{\Omega}}$ is two-dimensional, whereas $\mathbb{R}^{\vec{\mathcal{S}}}$ is only one-dimensional. Consequently, $\mathbf{O} : \mathbb{R}^{\vec{\Omega}} \to \mathbb{R}^{\vec{\mathcal{S}}}$ cannot be injective, so $\ker \mathbf{O} \neq \{0\}$, so (2) does not hold since (1) and (2) are equivalent. However, (3) still holds: Since there is only one state sequence, Equation (4.2) shows that the only vector of choice probabilities has $1/2$ in all its entries, irrespective of the return function G. Thus, $\mathbf{O} \otimes \mathbf{O}$ has only one input of observation-based choice probabilities, and is thus automatically injective on its inputs.

The final result of identifiability of the return function G follows using Corollary C.7. □

Simple special cases: full observability, deterministic $P_{\vec{O}}$, and noisy $P_{\vec{O}}$

In this section, we analyze three simple special cases of the general theory.

Theorem 3.9 (together with Lemma B.3) from Skalse, Farrugia-Roberts, et al. (2023), reproduced as a corollary below, is a special case of our theorem:

Corollary C.13 (Skalse, Farrugia-Roberts, et al. (2023))**.** *Assume the human directly observes the true sequences, and the choice probabilities are given by*

$$P^R\big(\vec{s} \succ \vec{s}'\big) = \sigma\Big(\beta\big(G(\vec{s}) - G(\vec{s}')\big)\Big).$$

This data determines the return function $G = \mathbf{\Gamma}(R)$ on state sequences $\vec{s} \in \vec{\mathcal{S}}$ up to a constant independent on $\vec{s}$.

Proof. We can embed this case into the one of Theorem C.12 by defining the observation kernel as $P_{\vec{O}}(\vec{s}' \mid \vec{s}) = \delta_{\vec{s}}(\vec{s}')$ (i.e., the correct sequence is deterministically observed) and defining the human's belief as $B(\vec{s}' \mid \vec{s}) = \delta_{\vec{s}}(\vec{s}')$ (i.e., the human knows that the observation reflects the true sequence). This shows that $P(\vec{s} \succ \vec{s}')$ is of the form of Equation (C.4). The result follows from Theorem C.12: the operators $\mathbf{O}$ and $\mathbf{B}$ are the identity in this case, due to the defining property of the Kronecker delta, and so they are injective. □

The following proposition shows that Corollary C.13 is essentially the *only* example of deterministic observation kernel $P_{\vec{O}}$ for which $\mathbf{B}$ is injective. Note, however, that in some situations, we can have $\operatorname{im}\mathbf{\Gamma} \cap \ker \mathbf{B} = \{0\}$ even if $\mathbf{B}$ is not injective, see Example C.35.

Proposition C.14. *Assume $P_{\vec{O}}$, the observation kernel on the level of sequences, is deterministic and* not *injective. Then $\mathbf{O}$ is automatically injective. However, $\mathbf{B}$ is not injective.*

Proof. To show that $\mathbf{O}$ is injective, assume $v \in \mathbb{R}^{\vec{\Omega}}$ is such that $\mathbf{O}(v) = 0$. Then for all $\vec{s} \in \vec{\mathcal{S}}$, we get

$$0 = \big[\,\mathbf{O}(v)\big](\vec{s}) = \mathop{\mathbf{E}}_{\vec{o} \sim P_{\vec{O}}(\vec{o}|\vec{s})}\big[v(\vec{o})\big] = v\big(\vec{O}(\vec{s})\big).$$

Since $\vec{O} : \vec{\mathcal{S}} \to \vec{\Omega}$ is by definition surjective, we obtain $v = 0$.

$\vec{O} : \vec{\mathcal{S}} \to \vec{\Omega}$ is by definition surjective, and here assumed to be non-injective, which implies that $\vec{\mathcal{S}}$ has a higher cardinality than $\vec{\Omega}$. Thus, $\mathbf{B} : \mathbb{R}^{\vec{\mathcal{S}}} \to \mathbb{R}^{\vec{\Omega}}$ cannot be injective. □

In the following, we analyze a simple case that guarantees identifiability. It requires that the observation kernel is "well-behaved" of a form where the observations are simply "noisy states", and that the human is a Bayesian reasoner with any prior $B(\vec{s})$ that supports every state sequence $\vec{s} \in \vec{\mathcal{S}}$.

Definition C.15 (Noise in the Observation Kernel)**.** *Then we say that there is* noise in the observation kernel $P_O : \vec{\mathcal{S}} \to \Delta(\vec{\Omega})$ *if $\vec{\mathcal{S}} = \vec{\Omega}$ and if $\mathbf{O}$ is an injective linear operator.*

Proposition C.16. *Assume that $\vec{\mathcal{S}} = \vec{\Omega}$. Furthermore, assume that $B(\vec{s} \mid \vec{o})$ is given by the posterior with likelihood $P_{\vec{O}}(\vec{o} \mid \vec{s})$ and any prior $B(\vec{s})$ with $B(\vec{s}) > 0$ for all $\vec{s} \in \vec{\mathcal{S}}$. Then there is noise in the observation kernel if and only if $\mathbf{B}$ is injective.*

Proof. Assume $\mathbf{O}$ is injective. To show that $\mathbf{B}$ is injective, assume there is $G' \in \mathbb{R}^{\vec{\mathcal{S}}}$ with $\mathbf{B}(G') = 0$. Then for all $\vec{o} \in \vec{\Omega}$, we have

$$\begin{aligned} 0 = \big[\mathbf{B}(G')\big](\vec{o}) &= \mathop{\mathbf{E}}_{\vec{s} \sim B(\vec{s} \mid \vec{o})}\big[G'(\vec{s})\big] = \sum_{\vec{s}} B(\vec{s} \mid \vec{o}) G'(\vec{s}) \propto \sum_{\vec{s}} P_{\vec{O}}(\vec{o} \mid \vec{s}) \cdot \big(B(\vec{s}) \cdot G'(\vec{s})\big) \\ &= \big[\mathbf{O}^T (B \odot G')\big](\vec{o}). \end{aligned}$$

Here, $\mathbf{O}^T$ is the transpose of $\mathbf{O}$ and $B \odot G'$ is the componentwise product of the prior B with the return function G'. Since $\mathbf{O}$ is injective and thus invertible, $\mathbf{O}^T$ is as well. Thus, $B \odot G' = 0$, which implies $G' = 0$ since the prior gives positive probability to all state sequences. Thus, $\mathbf{B}$ is injective.

For the other direction, assume $\mathbf{B}$ is injective. To show that $\mathbf{O}$ is injective, let $v \in \mathbb{R}^{\vec{\Omega}}$ be any vector with $\mathbf{O}(v) = 0$. We do a similar computation as above: for all $\vec{s} \in \mathbb{R}^{\vec{\mathcal{S}}}$, we have

$$\begin{aligned} 0 = \big[\mathbf{O}(v)\big](\vec{s}) &= \mathop{\mathbf{E}}_{\vec{o} \sim P_{\vec{O}}(\vec{o} \mid \vec{s})}\big[v(\vec{o})\big] = \sum_{\vec{o}} P_{\vec{O}}(\vec{o} \mid \vec{s}) v(\vec{o}) \propto \sum_{\vec{o}} B(\vec{s} \mid \vec{o}) \cdot \big(P_{\vec{O}}(\vec{o}) \cdot v(\vec{o})\big) \\ &= \Big[\mathbf{B}^T \big(P_{\vec{O}} \odot v\big)\Big](\vec{s}). \end{aligned}$$

Here, $\mathbf{B}^T$ is the transpose of $\mathbf{B}$, $P_{\vec{O}}(\vec{o})$ is the denominator in Bayes rule, and $P_{\vec{O}} \odot v$ is the vector with components $P_{\vec{O}}(\vec{o}) \cdot v(\vec{o})$. From the injectivity and thus invertibility of $\mathbf{B}$, it follows that $\mathbf{B}^T$ is invertible as well, and so $P_{\vec{O}} \odot v = 0$, which implies $v = 0$. Thus, $\mathbf{O}$ is injective. □

Corollary C.17. *When there is noise in the observation kernel and the human is a Bayesian reasoner with some prior B such that $B(\vec{s}) > 0$ for all $\vec{s} \in \vec{\mathcal{S}}$, then the return function is identifiable from choice probabilities of state sequences even if the learning system does not know the human's observations.*

Proof. This follows from the injectivity of $\mathbf{O}$, the injectivity of $\mathbf{B}$ that we proved in Proposition C.16, and Theorem C.12. □

Remark C.18. *We mention the following caveat: intuitively, one could think that $\mathbf{O}$ (and thus $\mathbf{B}$, by Proposition C.16) will be injective if every $\vec{s}$ is identifiable from infinitely many i.i.d. samples from $P_{\vec{O}}(\vec{o} \mid \vec{s})$. A counterexample is the following:*

$$\mathbf{O} = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 3/8 & 3/8 & 1/4 \end{pmatrix}.$$

In this case, the rows are linearly dependent with coefficients $1/2, 1/2$ *and* -1*. Consequently,* $\mathbf{O}$ *and* $\mathbf{B}$ *are not injective, and so if this observation kernel comes from a multi-armed bandit with three states, then Corollary C.7 shows that the return function is not identifiable.*

Nevertheless, the distributions $P_{\vec{O}}(\cdot \mid \vec{s})$ *(given by the rows) all differ from each other, and so infinitely many i.i.d. samples identify the state sequence* $\vec{s}$*.*

Robustness of return function identifiability under belief misspecification

We now again look at the case where the observations that the human observes are known to the reward learning system, as in Appendix C.3. Furthermore, we assume that $\mathbf{B} : \mathbb{R}^{\vec{\mathcal{S}}} \to \mathbb{R}^{\vec{\Omega}}$ is such that $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$. In this case, we can apply Corollary C.7 and identify the true return function G from $\mathbf{B}(G)$, which, in turn, can be identified up to an additive constant from the observation-based choice probabilities with the argument as for Proposition 4.1.

In this section, we investigate what happens when the human belief model is slightly misspecified. In other words: the learning system uses a perturbed matrix $\mathbf{B}_{\mathbf{\Delta}} \coloneqq \mathbf{B} + \mathbf{\Delta}$ with some small perturbation $\mathbf{\Delta}$. How much will the inferred return function deviate from the truth? To answer this, we first need to outline some norm theory of linear operators.

Some norm theory for linear operators

In this section, let V, W be two finite-dimensional inner product-spaces. In other words, V and W each have inner products $\langle \cdot, \cdot \rangle$ and there are linear isomorphisms $V \cong \mathbb{R}^k$, $W \cong \mathbb{R}^m$ such that the inner products in V and W correspond to the standard scalar products in $\mathbb{R}^k$ and $\mathbb{R}^m$. The reason that we don't directly work with $\mathbb{R}^k$ and $\mathbb{R}^m$ itself is that we will later apply the analysis to the case that $V = \operatorname{im} \mathbf{\Gamma} \subseteq \mathbb{R}^{\vec{\mathcal{S}}}$. Let in this whole section $\mathbf{A} : V \to W$ be a linear operator and $\mathbf{\Delta} : V \to W$ be a perturbance, so that $\mathbf{A}_{\mathbf{\Delta}} \coloneqq \mathbf{A} + \mathbf{\Delta}$ is a perturbed version of $\mathbf{A}$.

The inner products give rise to a norm on V and W defined by

$$\|v\| = \sqrt{\langle v, v \rangle}, \quad \|w\| = \sqrt{\langle w, w \rangle}.$$

As is well known, for each linear operator $\mathbf{A} : V \to W$ there exists a unique, basis-independent *adjoint* (generalizing the notion of a transpose) $\mathbf{A}^T : W \to V$ such that for all $v \in V$ and $w \in W$, we have

$$\langle \mathbf{A}\, v, w \rangle = \left\langle v, \mathbf{A}^T\, w \right\rangle.$$

Let us recall the following fact that is often used in linear regression:

Lemma C.19. *Assume* $\mathbf{A} : V \to W$ *is injective. Then* $\mathbf{A}^T\, \mathbf{A} : V \to V$ *is invertible and* $(\mathbf{A}^T\, \mathbf{A})^{-1}\, \mathbf{A}^T$ *is a left inverse of* $\mathbf{A}$*.*

Proof. To show that $\mathbf{A}^T\,\mathbf{A}$ is invertible, we only need to show that it is injective. Thus, let $0 \neq x \in V$. Then

$$\big\langle x, \mathbf{A}^T\,\mathbf{A}\,x \big\rangle = \langle \mathbf{A}\,x, \mathbf{A}\,x \rangle = \|\,\mathbf{A}\,x\|^2 > 0,$$

where the last step followed from the injectivity of $\mathbf{A}$. Thus, $\mathbf{A}^T\,\mathbf{A}\,x \neq 0$, and so $\mathbf{A}^T\,\mathbf{A}$ is injective, and thus invertible. Consequently, $(\mathbf{A}^T\,\mathbf{A})^{-1}\,\mathbf{A}^T$ is a well-defined operator. That it is the left inverse of $\mathbf{A}$ is clear. □

Definition C.20 (Operator Norm)**.** *The* norm of an operator $\mathbf{A} : V \to W$ *is given by*

$$\|\,\mathbf{A}\,\| \coloneqq \max_{x,\ \|x\|=1} \|\,\mathbf{A}\,x\|.$$

It has the following well-known properties, where $\mathbf{A}, \mathbf{B}$ *and* $\mathbf{C}$ *are matrices of compatible sizes:*

$$\|\,\mathbf{A}+\mathbf{B}\,\| \leq \|\,\mathbf{A}\,\| + \|\,\mathbf{B}\,\|, \quad \|\,\mathbf{C}\,\mathbf{A}\,\| \leq \|\,\mathbf{C}\,\| \cdot \|\,\mathbf{A}\,\|, \quad \|\,\mathbf{A}^T\,\| = \|\,\mathbf{A}\,\|.$$

To study how a perturbance in $\mathbf{A}$ (and thus $\mathbf{A}^T\,\mathbf{A}$) transfers into a perturbance of $\big(\,\mathbf{A}^T\,\mathbf{A}\,\big)^{-1}$, we will use the following theorem:

Theorem C.21 (El Ghaoui (2002))**.** *Let* $\mathbf{B} : V \to V$ *be an invertible operator. Let* $\rho < \|\,\mathbf{B}^{-1}\,\|^{-1}$*. Let* $\mathbf{\Delta} : V \to V$ *be any operator with* $\|\,\mathbf{\Delta}\,\| \leq \rho$*. Then* $\mathbf{B}+\mathbf{\Delta}$ *is invertible and we have*

$$\big\|(\mathbf{B}+\mathbf{\Delta})^{-1} - \mathbf{B}^{-1}\,\big\| \leq \frac{\rho \cdot \|\,\mathbf{B}^{-1}\,\|}{\|\,\mathbf{B}^{-1}\,\|^{-1} - \rho}.$$

Proof. See El Ghaoui (2002), Section 7 and in particular Equation 7.2. Note that the reference defines $\|\,\mathbf{A}\,\|$ to be the largest singular value of $\mathbf{A}$; by the well-known min-max theorem, this is equivalent to Definition C.20. □

We will apply this theorem to $\mathbf{A}^T\,\mathbf{A}$, which raises the question about the size of the perturbance in $\mathbf{A}^T\,\mathbf{A}$ for a given perturbance in $\mathbf{A}$. This is clarified in the following lemma. Before stating it, for a given perturbance ρ, define

$$\widetilde{\rho}(\mathbf{A}) \coloneqq \rho \cdot \big(2 \cdot \|\,\mathbf{A}\,\| + \rho\big),$$

which depends on $\mathbf{A}$ and ρ. Also, recall that for a given perturbance $\mathbf{\Delta}$, we define $\mathbf{A}_{\mathbf{\Delta}} \coloneqq \mathbf{A}+\mathbf{\Delta}$. We obtain:

Lemma C.22. *Assume that* $\|\,\mathbf{\Delta}\,\| \leq \rho$*. Then*

$$\|\,\mathbf{A}_{\mathbf{\Delta}}^T\,\mathbf{A}_{\mathbf{\Delta}} - \mathbf{A}^T\,\mathbf{A}\,\| \leq \widetilde{\rho}(\mathbf{A}).$$

Proof. We have

$$\big\|\,\mathbf{A}_{\mathbf{\Delta}}^T\,\mathbf{A}_{\mathbf{\Delta}} - \mathbf{A}^T\,\mathbf{A}\,\big\| = \big\|(\mathbf{A}+\mathbf{\Delta})^T(\mathbf{A}+\mathbf{\Delta}) - \mathbf{A}^T\,\mathbf{A}\,\big\|$$

$$
\begin{aligned}
&= \| \mathbf{A}^T \boldsymbol{\Delta} + \boldsymbol{\Delta}^T \mathbf{A} + \boldsymbol{\Delta}^T \boldsymbol{\Delta} \| \\
&\leq \| \mathbf{A} \| \cdot \| \boldsymbol{\Delta} \| + \| \boldsymbol{\Delta} \| \cdot \| \mathbf{A} \| + \| \boldsymbol{\Delta} \|^2 \\
&\leq \rho \cdot \Big(2 \cdot \| \mathbf{A} \| + \rho \Big) \\
&= \widetilde{\rho}(\mathbf{A}). \qquad \square
\end{aligned}
$$

To be able to apply Theorem C.21 to $\mathbf{A}^T \mathbf{A}$, we need to make sure that $\widetilde{\rho}(\mathbf{A})$ is bounded above by $\big\|(\mathbf{A}^T \mathbf{A}\,)^{-1}\|^{-1}$. The next lemma clarifies what condition ρ needs to satisfy for $\widetilde{\rho}(\mathbf{A})$ to obey that bound. For this, define

$$
\tau(\mathbf{A}) \coloneqq -\| \mathbf{A} \| + \sqrt{\| \mathbf{A} \|^2 + \big\|(\mathbf{A}^T \mathbf{A})^{-1}\big\|^{-1}}, \tag{C.5}
$$

which only depends on $\mathbf{A}$.

Lemma C.23. *Assume $\rho < \tau(\mathbf{A})$. Then*

$$
\widetilde{\rho}(\mathbf{A}) < \big\|(\mathbf{A}^T \mathbf{A})^{-1}\big\|^{-1}.
$$

Proof. Note that $\rho = \tau(\mathbf{A})$ is the positive solution to the following quadratic equation in the indeterminate ρ:

$$
\rho^2 + 2 \cdot \| \mathbf{A} \| \cdot \rho - \big\|(\mathbf{A}^T \mathbf{A})^{-1}\big\|^{-1} = \widetilde{\rho}(\mathbf{A}) - \big\|(\mathbf{A}^T \mathbf{A})^{-1}\big\|^{-1} = 0.
$$

Since this is a convex parabola, we get the inequality $\widetilde{\rho}(\mathbf{A}) - \big\|(\mathbf{A}^T \mathbf{A})^{-1}\big\|^{-1} < 0$ whenever we have $0 \leq \rho < \tau(\mathbf{A})$, which shows the result. $\square$

Finally, we put it all together to obtain a bound on the perturbance of $\big(\mathbf{A}^T \mathbf{A}\,\big)^{-1} \mathbf{A}^T$. For this, set

$$
C(\mathbf{A}, \rho) \coloneqq \frac{\widetilde{\rho}(\mathbf{A}) \cdot \Big\| \big(\mathbf{A}^T \mathbf{A}\,\big)^{-1} \Big\|}{\Big\| \big(\mathbf{A}^T \mathbf{A}\,\big)^{-1} \Big\|^{-1} - \widetilde{\rho}(\mathbf{A})} \cdot \Big(\| \mathbf{A} \| + \rho \Big) + \Big\| \big(\mathbf{A}^T \mathbf{A}\,\big)^{-1} \Big\| \cdot \rho. \tag{C.6}
$$

We obtain:

Proposition C.24. *Assume $\| \boldsymbol{\Delta} \| \leq \rho < \tau(\mathbf{A})$. Then $\mathbf{A}_{\boldsymbol{\Delta}}^T \mathbf{A}_{\boldsymbol{\Delta}}$ is invertible, and we have*

$$
\Big\| \big(\mathbf{A}_{\boldsymbol{\Delta}}^T \mathbf{A}_{\boldsymbol{\Delta}}\big)^{-1} \mathbf{A}_{\boldsymbol{\Delta}}^T - \big(\mathbf{A}^T \mathbf{A}\,\big)^{-1} \mathbf{A}^T \Big\| \leq C(\mathbf{A}, \rho).
$$

Proof. The invertibility of $\mathbf{A}_{\boldsymbol{\Delta}}^T \mathbf{A}_{\boldsymbol{\Delta}}$ follows from Theorem C.21, Lemma C.22 and Lemma C.23. We get

$$
\Big\| \big(\mathbf{A}_{\boldsymbol{\Delta}}^T \mathbf{A}_{\boldsymbol{\Delta}}\big)^{-1} \mathbf{A}_{\boldsymbol{\Delta}}^T - \big(\mathbf{A}^T \mathbf{A}\,\big)^{-1} \mathbf{A}^T \Big\|
$$

$$
\begin{aligned}
&= \left\| \left[\left(\mathbf{A}_{\mathbf{\Delta}}^T \mathbf{A}_{\mathbf{\Delta}} \right)^{-1} - \left(\mathbf{A}^T \mathbf{A} \right)^{-1} \right] \cdot \mathbf{A}_{\mathbf{\Delta}}^T + \left(\mathbf{A}^T \mathbf{A} \right)^{-1} \cdot \left(\mathbf{A}_{\mathbf{\Delta}}^T - \mathbf{A}^T \right) \right\| \\
&\leq \left\| \left(\mathbf{A}_{\mathbf{\Delta}}^T \mathbf{A}_{\mathbf{\Delta}} \right)^{-1} - \left(\mathbf{A}^T \mathbf{A} \right)^{-1} \right\| \cdot \| \mathbf{A}_{\mathbf{\Delta}} \| + \left\| \left(\mathbf{A}^T \mathbf{A} \right)^{-1} \right\| \cdot \| \mathbf{\Delta} \| \\
&\leq \frac{\widetilde{\rho}(\mathbf{A}) \cdot \left\| \left(\mathbf{A}^T \mathbf{A} \right)^{-1} \right\|}{\left\| \left(\mathbf{A}^T \mathbf{A} \right)^{-1} \right\|^{-1} - \widetilde{\rho}(\mathbf{A})} \cdot \Big(\| \mathbf{A} \| + \rho \Big) + \left\| \left(\mathbf{A}^T \mathbf{A} \right)^{-1} \right\| \cdot \rho \\
&= C(\mathbf{A}, \rho).
\end{aligned}
$$

In the second-to-last step, we used Theorem C.21. □

The constant $C(\mathbf{A}, \rho)$, defined in Equation (C.6), has a fairly complicated form. In the following proposition, we find an easier-to-study upper bound in a special case:

Proposition C.25. *Assume that* $\rho \leq \| \mathbf{A} \|$ *and* $\rho \leq -\| \mathbf{A} \| + \sqrt{\| \mathbf{A} \|^2 + 1/2 \cdot \left\| (\mathbf{A}^T \mathbf{A})^{-1} \right\|^{-1}}$.[2] *Then we have*

$$
C(\mathbf{A}, \rho) \leq \rho \cdot \left\| (\mathbf{A}^T \mathbf{A})^{-1} \right\| \cdot \Big[12 \cdot \| \mathbf{A} \|^2 \cdot \left\| (\mathbf{A}^T \mathbf{A})^{-1} \right\| + 1 \Big].
$$

Proof. The second assumption gives, as in the proof of Lemma C.23, that $\widetilde{\rho}(\mathbf{A}) \leq 1/2 \cdot \left\| (\mathbf{A}^T \mathbf{A})^{-1} \right\|^{-1}$. Together with $\rho \leq \| \mathbf{A} \|$, the result follows. □

Application to bounds in the error of the return Function

We now apply the results from the preceding section to our case. Define $\mathrm{r}(\mathbf{B}) : \operatorname{im} \mathbf{\Gamma} \to \mathbb{R}^{\vec{\Omega}}$ as the restriction of the belief operator $\mathbf{B}$ to $\operatorname{im} \mathbf{\Gamma}$. Assume that $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$, which is, according to Corollary C.7, a sufficient condition for identifiability. Note that this condition means that $\mathrm{r}(\mathbf{B})$ is injective. Thus, Lemma C.19 ensures that $\mathrm{r}(\mathbf{B})^T \mathrm{r}(\mathbf{B})$ is invertible and that $\big(\mathrm{r}(\mathbf{B})^T \mathrm{r}(\mathbf{B})\big)^{-1} \mathrm{r}(\mathbf{B})^T$ is a left inverse of $\mathrm{r}(\mathbf{B})$.

Consequently, from the equation

$$
\mathrm{r}(\mathbf{B})(G) = \mathbf{B}(G)
$$

we obtain

$$
G = \big(\mathrm{r}(\mathbf{B})^T \mathrm{r}(\mathbf{B})\big)^{-1} \mathrm{r}(\mathbf{B})^T (\mathbf{B}(G)).
$$

This is the concrete formula with which G can be identified from $\mathbf{B}(G)$. When perturbing $\mathbf{B}$, this leads to a corresponding perturbance in $\big(\mathrm{r}(\mathbf{B})^T \mathrm{r}(\mathbf{B})\big)^{-1} \mathrm{r}(\mathbf{B})^T$ whose size influences the maximal error in the inference of G. This, in turn, influences the size of the error in J_G, the policy evaluation function, where

$$
J_G(\pi) \coloneqq \mathop{\mathbf{E}}_{\vec{s} \sim P^\pi(\vec{s})} \big[G(\vec{s}) \big].
$$

We obtain:

[2] Note the factor 1/2 compared to the definition of $\tau(\mathbf{A})$ in Equation (C.5).

Theorem C.26. *Let G be the true reward function, $\mathbf{B}$ the belief operator corresponding to the human's true belief model $B(\vec{s} \mid \vec{o})$, and $\mathbf{B}(G)$ be the resulting observation-based return function. Assume that $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$, so that $\mathrm{r}(\mathbf{B})^T \mathrm{r}(\mathbf{B})$ is invertible. Let $\mathbf{\Delta} : \mathbb{R}^{\vec{\mathcal{S}}} \to \mathbb{R}^{\vec{\Omega}}$ be a perturbation satisfying $\| \mathbf{\Delta} \| \leq \rho$, where ρ satisfies the following two properties:*

$$\rho \leq \big\|\mathrm{r}(\mathbf{B})\big\|, \quad \rho \leq -\big\|\mathrm{r}(\mathbf{B})\big\| + \sqrt{\big\|\mathrm{r}(\mathbf{B})\big\|^2 + 1/2 \cdot \Big\|\big(\mathrm{r}(\mathbf{B})^T \mathrm{r}(\mathbf{B})\big)^{-1}\Big\|^{-1}}.$$

Let $\mathbf{B}_{\mathbf{\Delta}} \coloneqq \mathbf{B} + \mathbf{\Delta}$ be the misspecified belief operator. The first claim is that $\mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})^T \mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})$ is invertible under these conditions.

Now, assume that the learning system infers the return function

$$\tilde{G} \coloneqq \big(\mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})^T \mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})\big)^{-1} \mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})^T (\mathbf{B}(G)).^3$$

Then there is a polynomial $Q(X, Y)$ of degree five such that

$$\|\tilde{G} - G\| \leq \|G\| \cdot Q\Big(\big\|(\mathrm{r}(\mathbf{B})^T \mathrm{r}(\mathbf{B}))^{-1}\big\|, \|\mathrm{r}(\mathbf{B})\|\Big) \cdot \rho.$$

Thus, for all policies π, we obtain

$$\Big|J_{\tilde{G}}(\pi) - J_G(\pi)\Big| \leq \|G\| \cdot Q\Big(\big\|(\mathrm{r}(\mathbf{B})^T \mathrm{r}(\mathbf{B}))^{-1}\big\|, \|\mathrm{r}(\mathbf{B})\|\Big) \cdot \rho.$$

In particular, for sufficiently small perturbances ρ, the error in the inferred policy evaluation function $J_{\tilde{G}}$ becomes arbitrarily small.

Proof. That $\mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})^T \mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})$ is invertible follows immediately from Proposition C.24 by using that $\|\mathrm{r}(\mathbf{\Delta})\| \leq \| \mathbf{\Delta} \|$ and that $\mathrm{r}(\mathbf{B}_{\mathbf{\Delta}}) = \mathrm{r}(\mathbf{B})_{\mathrm{r}(\mathbf{\Delta})}$, together with the second bound on ρ (which implies the assumed bound in Proposition C.24).

We have

$$\begin{aligned}
\Big|J_{\tilde{G}}(\pi) - J_G(\pi)\Big| &= \Big|\mathop{\mathbf{E}}_{\vec{s} \sim P^\pi(\vec{s})}\big[(\tilde{G} - G)(\vec{s})\big]\Big| \\
&\leq \mathop{\mathbf{E}}_{\vec{s} \sim P^\pi(\vec{s})}\Big[\big|(\tilde{G} - G)(\vec{s})\big|\Big] \\
&\leq \max_{\vec{s} \in \vec{\mathcal{S}}} \big|(\tilde{G} - G)(\vec{s})\big| \\
&\leq \|\tilde{G} - G\| \\
&= \Big\|\Big[\big(\mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})^T \mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})\big)^{-1} \mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})^T - \big(\mathrm{r}(\mathbf{B})^T \mathrm{r}(\mathbf{B})\big)^{-1} \mathrm{r}(\mathbf{B})^T\Big] \cdot \mathbf{B}(G)\Big\| \\
&\leq \big\|\big(\mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})^T \mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})\big)^{-1} \mathrm{r}(\mathbf{B}_{\mathbf{\Delta}})^T - \big(\mathrm{r}(\mathbf{B})^T \mathrm{r}(\mathbf{B})\big)^{-1} \mathrm{r}(\mathbf{B})^T\big\| \cdot \big\|\mathbf{B}(G)\big\| \\
&\leq C(\mathrm{r}(\mathbf{B}), \rho) \cdot \|\mathrm{r}(\mathbf{B})(G)\| \\
&\leq C(\mathrm{r}(\mathbf{B}), \rho) \cdot \|\mathrm{r}(\mathbf{B})\| \cdot \|G\|.
\end{aligned}$$

In the second to last step, we used Proposition C.24. By Proposition C.25, we can define the polynomial $Q(X, Y)$ by

$$Q(X, Y) = XY \cdot \Big[12XY^2 + 1\Big],$$

which is of degree five.

The last claim follows from $\lim_{\rho \to 0} \rho = 0$. □

Remark C.27. *In the case of a square matrix* $\mathbf{B}$ *that is injective, we can apply Theorem C.21 directly to* $\mathbf{B}^{-1}$ *(which is now invertible) and obtain the following simplification of Theorem C.26 for the case that* $\| \mathbf{\Delta} \| \leq \rho \leq \frac{1}{2} \cdot \| \mathbf{B}^{-1} \|^{-1}$*:*

$$\big| J_{\tilde{G}}(\pi) - J_G(\pi) \big| \leq \rho \cdot 2 \cdot \| \mathbf{B} \| \cdot \|G\| \cdot \| \mathbf{B}^{-1} \|^2.$$

The polynomial is then only of degree 3.

Preliminary characterizations of the ambiguity

Recall the sequence of functions

$$\mathbb{R}^{\mathcal{S}} \xrightarrow{\mathbf{\Gamma}} \mathbb{R}^{\vec{\mathcal{S}}} \xrightarrow{\mathbf{B}} \mathbb{R}^{\vec{\Omega}}.$$

In this section, we clarify $\operatorname{im} \mathbf{\Gamma}$ and $\ker \mathbf{B}$ in special cases, as their intersection is the crucial ambiguity in Theorem C.5.

The following proposition shows that for deterministic $P_{\vec{O}}$ and a rational human, $\ker \mathbf{B}$ decomposes into hyperplanes defined by normal vectors of probabilities of sequences mapping to the same observation sequence:

Proposition C.28. *Assume the human reasons as in Appendix C.3. Assume* $P_{\vec{O}}$ *is deterministic. Let* $B(\vec{s})$ *be the distribution of sequences under the human's belief over the policy, given by* $B(\vec{s}) = \int_{\pi'} B(\pi') P^{\pi'}(\vec{s})$ *for some policy prior* $B(\pi')$*. For each* $\vec{o}$*, let* $B_{\vec{o}} \coloneqq [B(\vec{s})]_{\vec{s}:\ \vec{O}(\vec{s}) = \vec{o}} \in \mathbb{R}^{\{\vec{s} \in \vec{\mathcal{S}} \ \mid \ \vec{O}(\vec{s}) = \vec{o}\}}$ *be the vector of probabilities of sequences that are observed as* $\vec{o}$*.*

Let G' *be a return function. For each* $\vec{o} \in \vec{\Omega}$*, define the restriction* $G'_{\vec{o}} \in \mathbb{R}^{\{\vec{s} \in \vec{\mathcal{S}} | \vec{O}(\vec{s}) = \vec{o}\}}$ *by* $G'_{\vec{o}}(\vec{s}) \coloneqq G'(\vec{s})$ *for all* $\vec{s} \in \{\vec{s} \in \vec{\mathcal{S}} \mid \vec{O}(\vec{s}) = \vec{o}\}$*. Assume that* $B(\vec{s} \mid \vec{o})$ *is the Bayesian posterior. Then* $G' \in \ker \mathbf{B}$ *if and only if the property*

$$B_{\vec{o}} \cdot G'_{\vec{o}} = 0$$

holds for all $\vec{o} \in \vec{\Omega}$*.*

Proof. For a deterministic observation kernel $P_{\vec{O}}$, by Bayes rule we have

$$\begin{aligned} B(\vec{s} \mid \vec{o}) &= \frac{P_{\vec{O}}(\vec{o} \mid \vec{s}) \cdot B(\vec{s})}{\sum_{\vec{s}'} P_{\vec{O}}(\vec{o} \mid \vec{s}') \cdot B(\vec{s}')} \\ &= \frac{\delta_{\vec{o}}\big(\vec{O}(\vec{s})\big) \cdot B(\vec{s})}{\sum_{\vec{s}'} \delta_{\vec{o}}\big(\vec{O}(\vec{s}')\big) \cdot B(\vec{s}')} \\ &= \begin{cases} 0, & \vec{O}(\vec{s}) \neq \vec{o} \\ \frac{B(\vec{s})}{\sum_{\vec{s}':\ \vec{O}(\vec{s}')=\vec{o}} B(\vec{s}')}, & \vec{O}(\vec{s}) = \vec{o}. \end{cases} \end{aligned}$$

Thus, for any return function G' and any observation sequence $\vec{o}$, we have

$$\begin{aligned} \big[\mathbf{B}(G')\big](\vec{o}) &= \mathop{\mathbf{E}}_{\vec{s}\sim B(\vec{s}|\vec{o})}\big[G'(\vec{s})\big] \\ &= \sum_{\vec{s}} B(\vec{s} \mid \vec{o}) G'(\vec{s}) \\ &= \sum_{\vec{s}:\ \vec{O}(\vec{s})=\vec{o}} \frac{B(\vec{s})}{\sum_{\vec{s}':\ \vec{O}(\vec{s}')=\vec{o}} B(\vec{s}')} G'(\vec{s}) \\ &= \left(\sum_{\vec{s}':\ \vec{O}(\vec{s}')=\vec{o}} B(\vec{s}') \right)^{-1} \cdot \sum_{\vec{s}:\ \vec{O}(\vec{s})=\vec{o}} B(\vec{s}) G'(\vec{s}). \end{aligned}$$

Thus, we have $G' \in \ker \mathbf{B}$ if and only if

$$B_{\vec{o}} \cdot G'_{\vec{o}} = \sum_{\vec{s}:\ \vec{O}(\vec{s})=\vec{o}} B(\vec{s}) G'(\vec{s}) = 0$$

for all $\vec{o}$. That was to show. □

Remark C.29. *One can interpret the previous proposition as follows:*

As long as $\vec{O}$ is injective, we have $\big|\{\vec{s} \in \vec{\mathcal{S}} \mid \vec{O}(\vec{s}) = o\}\big| = 1$ for all $\vec{o}$, meaning that $B_{\vec{o}}$ and $G'_{\vec{o}}$ have only one entry. Thus, $B_{\vec{o}} \cdot G'_{\vec{o}} = 0$ implies $G'_{\vec{o}} = 0$. If that holds for all $\vec{o}$, then $G' \in \ker \mathbf{B}$ implies $G' = 0$, meaning $\mathbf{B}$ is injective.

However, as soon as there is an $\vec{o}$ with $k_{\vec{o}} \coloneqq \big|\{\vec{s} \in \vec{\mathcal{S}} \mid \vec{O}(\vec{s}) = o\}\big| > 1$, the equation $B_{\vec{o}} \cdot G'_{\vec{o}} = 0$ leads to $k_{\vec{o}} - 1$ free parameters in $G'_{\vec{o}}$. $G'_{\vec{o}}$ can then be chosen freely in the hyperplane of vectors orthogonal to $B_{\vec{o}}$ without moving out of the kernel of $\mathbf{B}$.

Another way of writing Proposition C.28 is to write $\ker \mathbf{B}$ as a direct sum of these hyperplanes perpendicular to $B_{\vec{o}}$:

$$\ker \mathbf{B} = \bigoplus_{\vec{o}:\ |\vec{O}^{-1}(\vec{o})|\geq 2} B_{\vec{o}}^{\perp}.$$

Recall that a return function G is called *time-separable* if there exists a reward function R such that $\mathbf{\Gamma}(R) = G$.

Before we discuss time-separability in more interesting examples, we want to talk about one simple case where all return functions are time-separable. We leave a general characterization of $\operatorname{im}\mathbf{\Gamma}$ to future work.

Proposition C.30. *Let there be an ordering $\vec{s}^{(1)}, \vec{s}^{(2)}, \dots$ of all sequences in $\vec{\mathcal{S}}$, and a function $\phi : \vec{\mathcal{S}} \to \mathcal{S}$ from sequences to states such that $\phi(\vec{s}) \in \vec{s}$ and $\phi(\vec{s}^{(k)}) \notin \vec{s}^{(i)}$ for all $i < k$. Then every return function is time-separable.*

Proof. Let G be a return function. Initialize $R(s) = 0$ for all s and inductively update it for all $i = 1, 2, \dots$:

$$R\big(\phi(\vec{s}^{(i)})\big) \coloneqq \left(\sum_{t\colon\, s_t^{(i)} = \phi(\vec{s}^{(i)})} \gamma^t\right)^{-1} \cdot \left(G(\vec{s}^{(i)}) - \sum_{t\colon\, s_t^{(i)} \neq \phi(\vec{s}^{(i)})} \gamma^t \cdot R\big(s_t^{(i)}\big)\right),$$

where the inductive definition always uses R as it is defined by that point in time. Once $R\big(\phi(\vec{s}^{(i)})\big)$ is defined, but not yet any future values $R\big(\phi(\vec{s}^{(k)})\big)$, $k > i$, we have

$$\begin{aligned}
\big[\mathbf{\Gamma}(R)\big](\vec{s}^{(i)}) &= \sum_{t=0}^{T} \gamma^t \cdot R\big(s_t^{(i)}\big) \\
&= \left(\sum_{t\colon\, s_t^{(i)} = \phi(\vec{s}^{(i)})} \gamma^t\right) \cdot R\big(\phi(\vec{s}^{(i)})\big) + \sum_{t\colon\, s_t^{(i)} \neq \phi(\vec{s}^{(i)})} \gamma^t \cdot R\big(s_t^{(i)}\big) \\
&= G(\vec{s}^{(i)}).
\end{aligned}$$

Furthermore, the property $\phi(\vec{s}^{(k)}) \notin \vec{s}^{(i)}$ for all $i < k$ ensures that changes to the reward function for $k > i$ do not affect the value of $\big[\mathbf{\Gamma}(R)\big](\vec{s}^{(i)})$. This shows $\mathbf{\Gamma}(R) = G$, and thus G is time-separable. □

Corollary C.31. *In a multi-armed bandit, every return function is time-separable.*

Proof. In a multi-armed bandit, states and sequences are equivalent, and so we can choose $\phi(s) = s$ for every state/sequence s. The result follows from Proposition C.30.

Alternatively, simply directly notice that in a multi-armed bandit, $\mathbf{\Gamma}$ is the identity mapping, and so for every return/reward function R, we have $\mathbf{\Gamma}(R) = R$. □

Examples supplementing Section 4.5

In this whole section, the inverse temperature parameter in the human choice probabilities is given by $\beta = 1$. We now consider four more mathematical examples of Corollary C.7 and Theorem C.12. In the first example, the ambiguity is so bad that the reward inference can become worse than simply maximizing J_Ω as in naive RLHF. In Example C.33, there is simply "noise" in the observations and the human's belief, the matrices $\mathbf{B}$ and $\mathbf{O}$ are injective, and identifiability works, as in Corollary C.17. In the third example, the matrix $\mathbf{B}$ is not injective and identifiability fails, which is a minimal example showing the limits of our main theorems. In the fourth example, the matrix $\mathbf{B}$ is not injective, but $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$, and so identifiability works. This example is interesting in that the identifiability simply emerges through different distributions of *delay* that are caused by the different unobserved events.

In this section, both the linear operators $\mathbf{B} : \mathbb{R}^{\vec{\mathcal{S}}} \to \mathbb{R}^{\vec{\Omega}}$ and $\mathbf{O} : \mathbb{R}^{\vec{\Omega}} \to \mathbb{R}^{\vec{\mathcal{S}}}$ are considered as matrices

$$\mathbf{O} = \left(P_{\vec{O}}(\vec{o} \mid \vec{s})\right)_{\vec{s},\vec{o}} \in \mathbb{R}^{\vec{\mathcal{S}} \times \vec{\Omega}}, \quad \mathbf{B} = \left(B(\vec{s} \mid \vec{o})\right)_{\vec{o},\vec{s}} \in \mathbb{R}^{\vec{\Omega} \times \vec{\mathcal{S}}}.$$

Notice that both have a swap in their indices.

Example C.32. *Theorem 4.8 shows that the remaining ambiguity from the human's choice probabilities is given by* $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma}$*, but it doesn't explain how to proceed given this ambiguity. Without further inductive biases, some reward functions within the ambiguity of the true reward function can be even worse than simply maximizing* J_Ω*.*

E.g., consider a multi-armed bandit with three actions a, b, c*, observation-kernel* $o = O(a) = O(b) \neq O(c) = c$ *and reward function* $R(a) = R(b) < R(c)$*. If the human belief is given by* $B(a \mid o) = p = 1 - B(b \mid o)$*, then* $R' = \alpha \cdot (p - 1, p, 0) \in \mathbb{R}^{\{a,b,c\}}$ *is in the ambiguity for all* $\alpha \in \mathbb{R}$*, and so* $\tilde{R} \coloneqq R + R'$ *is compatible with the choice probabilities. However, for* $\alpha \ll 0$*, we have* $\tilde{R}(a) > \tilde{R}(b)$ *and* $\tilde{R}(a) > \tilde{R}(c)$*, and so optimizing against this reward function leads to a suboptimal policy.*

In contrast, maximizing J_Ω *leads to the correct policy since* a*,* b*, and* c *all obtain their ground truth reward in this example. This generally raises the question of how to tie-break reward functions in the ambiguity, or how to act conservatively given the uncertainty, in order to consistently improve upon the setting in Section 4.4.*

Example C.33. *This example is a special case of Corollary C.17. Consider a multi-armed bandit with two actions (which are automatically also states and sequences)* a *and* b*. In this case, the reward function and return function is the same.*

We assume there to be two possible observations $o^{(a)}, o^{(b)}$ *and the observation kernel to be non-deterministic, with probabilities*

$$P_O(o^{(j)} \mid i) = \begin{cases} 2/3, & \text{if } i = j, \\ 1/3, & \text{else.} \end{cases}$$

If we assume the human forms Bayesian posterior beliefs as in Appendix C.3 and to have a policy prior $B(\pi')$ *such that* $B(a) = \int_\pi \pi(a) B(\pi') d\pi = 1/2$ *and* $B(b) = 1/2$*, then it is easy*

to show that the human's belief is the "reversed" observation kernel:

$$B(j \mid o^{(i)}) = P_O(o^{(i)} \mid j).$$

We obtain

$$\mathbf{O} = \mathbf{B} = \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix} = \frac{1}{3} \cdot \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

These matrices are injective since they are invertible:

$$\mathbf{O}^{-1} = \mathbf{B}^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

More generally, even if the human does not form fully rational posterior beliefs, it is easy to imagine that the matrix $\mathbf{B}$ *can end up being invertible. Thus, Corollary C.7 guarantees that the reward function can be inferred up to an additive constant from the choice probabilities of observations, and Theorem C.12 shows that this even works when the learning system does not know what the human observed.*

In the rest of this example, we explicitly walk the reader through the process of how the reward function can be inferred, in the general case that the observations are not known. In the process, we essentially recreate the proof of the theorems for this special case. For this aim, we first want to compute the choice probabilities $P^R\big(i \succ j\big)$ *that the learning system has access to in the limit of infinite data. We assume that the reward function is given by* $R(a) = -1$ *and* $R(b) = 2$*. We compute:*

$$\mathbf{B}(R) = \frac{1}{3} \cdot \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

In other words, we have $\mathbf{E}_{s \sim B(s|o^{(a)})}[R(s)] = 0$ *and* $\mathbf{E}_{s \sim B(s|o^{(b)})}[R(s)] = 1$*. From this, we can compute the observation-based choice probabilities* $\widetilde{P}_{o^{(i)}o^{(j)}} = \sigma\big(\mathbf{B}(R)(o^{(i)}) - \mathbf{B}(R)(o^{(j)})\big)$*, see Equation* (4.2)*, and obtain:*

$$\widetilde{P}_{o^{(a)}o^{(a)}} = \widetilde{P}_{o^{(b)}o^{(b)}} = \frac{1}{2}, \quad \widetilde{P}_{o^{(a)}o^{(b)}} = \frac{1}{1+e}, \quad \widetilde{P}_{o^{(b)}o^{(a)}} = \frac{e}{1+e}.$$

We can now determine the final choice probabilities $P_{ij} \coloneqq P^R\big(i \succ j\big)$ *again by a matrix-vector product, with the indices ordered lexicographically, see Equation* (C.4)*. Here,* $\mathbf{O} \otimes \mathbf{O}$ *is the Kronecker product of the matrix* $\mathbf{O}$ *with itself:*

$$P = (\mathbf{O} \otimes \mathbf{O}) \cdot \widetilde{P} = \frac{1}{9} \cdot \begin{pmatrix} 4 & 2 & 2 & 1 \\ 2 & 4 & 1 & 2 \\ 2 & 1 & 4 & 2 \\ 1 & 2 & 2 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1/2 \\ 1/(1+e) \\ e/(1+e) \\ 1/2 \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1/3 \cdot (2+e)/(1+e) \\ 1/3 \cdot (1+2e)/(1+e) \\ 1/2 \end{pmatrix}.$$

For example, the second entry in P is $P_{ab} = P^R\big(a \succ b\big) = \frac{2+e}{3\cdot(1+e)}$. This is the likelihood that, for ground-truth actions a, b, the human will prefer a after only receiving observations $o^{(a)}$ or $o^{(b)}$ according to $\mathbf{O}$ and following a Boltzman-rational policy based on the belief of the real action, see Equation (C.4).

Over time, the learning system will be able to estimate these probabilities based on repeated human choices, assuming all state-pairs are sampled infinitely often. The question of identifiability is whether the original reward function R can be inferred from that data, given that the learning system knows $\mathbf{O}$ and $\mathbf{B}$. We assume that the learning system doesn't a priori *know R or any of the intermediate steps in the computation. First, $\widetilde{P}$ can be inferred by inverting $\mathbf{O} \otimes \mathbf{O}$:*

$$\widetilde{P} = (\mathbf{O}\otimes\mathbf{O})^{-1}\cdot P = \begin{pmatrix} 4 & -2 & -2 & 1 \\ -2 & 4 & 1 & -2 \\ -2 & 1 & 4 & -2 \\ 1 & -2 & -2 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1/2 \\ 1/3\cdot(2+e)/(1+e) \\ 1/3\cdot(1+2e)/(1+e) \\ 1/2 \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1/(1+e) \\ e/(1+e) \\ 1/2 \end{pmatrix}.$$

The learning system wants to use this to infer $\mathbf{B}(\tilde{R})$ (for the later-to-be inferred reward function $\tilde{R}$ that may differ from the true reward function R) and uses the equation

$$\widetilde{P}_{o^{(a)}o^{(b)}} = \frac{\exp\big(\mathbf{B}(\tilde{R})(o^{(a)})\big)}{\exp\big(\mathbf{B}(\tilde{R})(o^{(a)})\big) + \exp\big(\mathbf{B}(\tilde{R})(o^{(b)})\big)},$$

which can be rearranged to

$$\mathbf{B}(\tilde{R})(o^{(a)}) = \log\frac{\widetilde{P}_{o^{(a)}o^{(b)}}}{1-\widetilde{P}_{o^{(a)}o^{(b)}}} + \mathbf{B}(\tilde{R})(o^{(b)}) = \log\frac{1/(1+e)}{e/(1+e)} + \mathbf{B}(\tilde{R})(o^{(b)}) = \mathbf{B}(\tilde{R})(o^{(b)}) - 1.$$

This relation is all which can be inferred about $\mathbf{B}(\tilde{R})(o^{(a)})$ and $\mathbf{B}(\tilde{R})(o^{(b)})$; the precise value cannot be determined and $\mathbf{B}(\tilde{R})(o^{(b)})$ is a free parameter. One can check that for $\mathbf{B}(\tilde{R})(o^{(b)}) = 1$ this coincides with the true value $\mathbf{B}(R)$. Finally, one can invert $\mathbf{B}$ to infer $\tilde{R}$ from this:

$$\begin{aligned}
\tilde{R} &= \mathbf{B}^{-1}\cdot\mathbf{B}(\tilde{R}) \\
&= \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{B}(\tilde{R})(o^{(b)}) - 1 \\ \mathbf{B}(\tilde{R})(o^{(b)}) \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{B}(\tilde{R})(o^{(b)}) - 2 \\ 1 + \mathbf{B}(\tilde{R})(o^{(b)}) \end{pmatrix} \\
&= \begin{pmatrix} -1 \\ 2 \end{pmatrix} + \begin{pmatrix} \mathbf{B}(\tilde{R})(o^{(b)}) - 1 \\ \mathbf{B}(\tilde{R})(o^{(b)}) - 1 \end{pmatrix} \\
&= R + \begin{pmatrix} \mathbf{B}(\tilde{R})(o^{(b)}) - 1 \\ \mathbf{B}(\tilde{R})(o^{(b)}) - 1 \end{pmatrix}.
\end{aligned}$$

Thus, the inferred and true reward functions differ maximally by a constant, as predicted in Theorem C.12.

In the following example, we work out a case where the reward function is so ambiguous that any policy is optimal to some reward function consistent with the human feedback:

Example C.34. *Consider a multi-armed bandit with exactly three actions/states a, b, c. We assume a deterministic observation kernel with $o \coloneqq O(a) = O(c) \neq O(b) = b$. Assume the human has some arbitrary beliefs $B(a \mid o), B(c \mid o) = 1 - B(a \mid o)$, and can identify b: $B(b \mid b) = 1$. Then if the human makes observation comparisons with a Boltzman-rational policy, as in Theorem C.5, the resulting reward function is so ambiguous that some reward functions consistent with the feedback place the highest value on action a, no matter the true reward function R. Thus, even if the true reward function R regards a as the worst action, a can result from the reward learning and subsequent policy optimization process.*

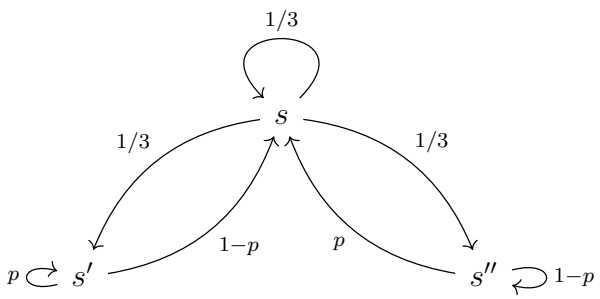
Proof. The matrix $\mathbf{B} : \mathbb{R}^{\{a,b,c\}} \to \mathbb{R}^{\{o,b\}}$ is given by

$$\mathbf{B} = \begin{pmatrix} B(a \mid o) & 0 & B(c \mid o) \\ 0 & 1 & 0 \end{pmatrix}.$$

Its kernel is given by reward functions R' with $R'(b) = 0$ and $R'(c) = -\frac{B(a|o)}{B(c|o)} R'(a)$, with $R'(a)$ a free parameter. Theorem C.5 shows that, up to an additive constant, the reward functions consistent with the feedback of observation comparisons are given by $\tilde{R} = R + R'$ for any $R' \in \ker \mathbf{B}$. Thus, whenever the free parameter $R'(a)$ satisfies $R'(a) > R(b) - R(a)$ and $R'(a) > B(c \mid o) \cdot \big(R(c) - R(a)\big)$, we obtain $\tilde{R}(a) > \tilde{R}(b)$ and $\tilde{R}(a) > \tilde{R}(c)$, showing the claim. □

We now investigate another example where $\mathbf{B}$ is not injective, and yet, identifiability works because $\mathbf{B} \circ \mathbf{\Gamma} \neq \{0\}$. We saw such cases already in Example C.41, but include this additional example since it shows a conceptually interesting case: two different states lead to the exact same observations, but can be disambiguated since they lead to different amounts of *delay* until a more informative observation is made again.

Example C.35. *In this example, we assume that the human knows the policy π that generates the state sequences (corresponding to a policy prior $B(\pi') = \delta_\pi(\pi')$ concentrated on π), which together with knowledge of the transition dynamics of the environment determines the true state transition probabilities $\mathcal{T}^\pi(s' \mid s) = \sum_{a \in \mathcal{A}} \mathcal{T}(s' \mid s, a) \cdot \pi(a \mid s)$. We consider an environment with three states s, s', s'' and the following transition dynamics $\mathcal{T}^\pi$, where $p \neq 1/2$ is a probability:*



We assume that $P_0(s) = 1$. Furthermore, we assume deterministic observations and $s = O(s) \neq O(s') = O(s'') =: o$.

Assume the time horizon T is 3, i.e., there are timesteps $0, 1, 2, 3$. Assume that the human forms the belief over the true state sequence by Bayesian posterior updates as in Appendix C.3. In this case, $\ker \mathbf{B} \neq \{0\}$ by Proposition C.14. However, we will now show that $\ker(\mathbf{B} \circ \mathbf{\Gamma}) = \{0\}$. If the human makes Boltzmann-rational comparisons of observation sequences, then this implies the identifiability of the return function up to an additive constant by Corollary C.7.[4]

Thus, let $R' \in \ker(\mathbf{B} \circ \mathbf{\Gamma})$, i.e., $\Big[\mathbf{B}\left(\mathbf{\Gamma}(R')\right)\Big](\vec{o}) = 0$ for every observation sequence $\vec{o}$. For $\vec{o} = ssss$ being the observation sequence that only consists of state s, this implies $R'(s) = 0$. Consequently, for general observation sequences $\vec{o}$, we have:

$$0 = \Big[\mathbf{B}\left(\mathbf{\Gamma}(R')\right)\Big](\vec{o}) = \mathop{\mathbf{E}}_{\vec{s}\sim B(\vec{s}|\vec{o})}\left[\sum_{t=0}^{3} \delta_{s'}(s_t) \cdot \gamma^t\right] \cdot R'(s') + \mathop{\mathbf{E}}_{\vec{s}\sim B(\vec{s}|\vec{o})}\left[\sum_{t=0}^{3} \delta_{s''}(s_t) \cdot \gamma^t\right] \cdot R'(s'').$$

Now we specialize this equation to the two observation sequences $\vec{o}^{(1)} = soss$ and $\vec{o}^{(2)} = soos$. We start by considering $\vec{o}^{(1)}$. This is consistent with the two state sequences $\vec{s}^{(1),(s')} = ss'ss$ and $\vec{s}^{(1),(s'')} = ss''ss$. We have posterior probabilities

$$B\big(\vec{s}^{(1),(s')} \mid \vec{o}^{(1)}\big) = 1 - p, \quad B\big(\vec{s}^{(1),(s'')} \mid \vec{o}^{(1)}\big) = p,$$

and therefore

$$0 = \Big[\mathbf{B}\left(\mathbf{\Gamma}(R')\right)\Big](\vec{o}^{(1)}) = (1-p) \cdot \gamma \cdot R'(s') + p \cdot \gamma \cdot R'(s''),$$

and so

$$R'(s') = \frac{p}{p-1} \cdot R'(s''). \tag{C.7}$$

Similarly, $\vec{o}^{(2)}$ is consistent with the sequences $\vec{s}^{(2),(s')} = ss's's$ and $\vec{s}^{(2),(s'')} = ss''s''s$. They have posterior probabilities

$$B\big(\vec{s}^{(2),(s')} \mid \vec{o}^{(2)}\big) = \frac{1}{2}, \quad B\big(\vec{s}^{(2),(s'')} \mid \vec{o}^{(2)}\big) = \frac{1}{2},$$

leading to

$$0 = \frac{1}{2} \cdot (\gamma + \gamma^2) \cdot R'(s') + \frac{1}{2} \cdot (\gamma + \gamma^2) \cdot R'(s'').$$

Together with Equation (C.7)*, we obtain*

$$R'(s'') = -R'(s') = \frac{p}{1-p} \cdot R'(s''),$$

which implies $R'(s'') = 0$ because $p \neq \frac{1}{2}$, and thus also $R'(s') = 0$. Overall, we have showed $R' = 0$, and so $\mathbf{B} \circ \mathbf{\Gamma}$ is injective. This means that reward functions are identifiable in this example up to an additive constant, see Corollary C.7.

[4] We assume that the learning system knows what the human observes, which is valid since P_O is deterministic. Alternatively, one can argue with Proposition C.14 that $\mathbf{O}$ is automatically injective, meaning one can apply Theorem C.12.

C.4 Issues of naively applying RLHF under partial observability

In this section, we study the naive application of RLHF under partial observability. Thus, most of it takes a step back from the general theory of *appropriately modeled* partial observability in RLHF.

We first briefly explain what happens when the learning system incorrectly assumes that the human observes the full environment state. We show that as a consequence, the system is incentivized to infer what we call the *observation return function* G_Ω, which evaluates a state sequence based on the human's belief of the state sequence given the human's observations. In the policy optimization process, the policy is then selected to maximize J_Ω, an expectation over G_Ω. In an interlude, we then briefly analyze the unrealistic case that the human, when evaluating a policy π, fully knows the complete specification of that policy and all of the environment and engages in rational Bayesian reasoning; in this case, $J_\Omega = J$ is the true policy evaluation function.

Realistically, however, maximizing J_Ω can lead to failure modes. Accordingly, we show that a suboptimal policy that is optimal according to J_Ω causes deceptive inflation, overjustification, or both. (For examples, see Appendix C.2, where we expand on the analysis of the main examples in the main text.) Finally, we study further concrete examples where maximizing J_Ω reveals deceptive and overjustifying behavior by the resulting policy.

Optimal policies under RLHF with deterministic partial observations maximize J_Ω

Assume that $P_{\vec{O}}$ is deterministic and that the human makes Boltzmann-rational sequence comparisons between observation sequences. The true choice probabilities are then given by (See Equations (4.2) and (C.4)):

$$P^R\big(\vec{s} \succ \vec{s}'\big) = \sigma\Big(\beta \cdot \Big(\big(\mathbf{B} \cdot G\big)\big(\vec{O}(\vec{s})\big) - \big(\mathbf{B} \cdot G\big)\big(\vec{O}(\vec{s}')\big)\Big)\Big) \tag{C.8}$$

Now, assume that the learning system does *not model the situation correctly*. In particular, we assume:

- The system is not aware that the human only observes observation sequences $\vec{O}(\vec{s})$ instead of the full state sequences.
- The system does not model that the human's return function is *time-separable*, i.e., comes from a reward function R over environment states.

The learning system then thinks that there is a return function $\tilde{G} \in \mathbb{R}^{\vec{S}}$ such that the choice probabilities are given by the following faulty formula:

$$P^R\big(\vec{s} \succ \vec{s}'\big) \coloneqq \sigma\Big(\beta\big(G(\vec{s}) - G(\vec{s}')\big)\Big)$$

Now, assume that the learning system has access to the choice probabilities and wants to infer G. Inverting the sigmoid function and then plugging in the true choice probabilities from Equation (C.8), we obtain:

$$\begin{aligned}\tilde{G}(\vec{s}) &= \frac{1}{\beta}\log\frac{P^R(\vec{s}\succ\vec{s}')}{P^R(\vec{s}'\succ\vec{s})} + \tilde{G}(\vec{s}')\\ &= \frac{1}{\beta}\Big[\beta\cdot\Big(\big(\mathbf{B}\cdot G\big)\big(\vec{O}(\vec{s})\big) - \big(\mathbf{B}\cdot G\big)\big(\vec{O}(\vec{s}')\big)\Big)\Big] + \tilde{G}(\vec{s}')\\ &= \big(\mathbf{B}\cdot G\big)\big(\vec{O}(\vec{s})\big) + C(\vec{s}').^5\end{aligned}$$

Here, $C(\vec{s}')$ is some quantity that does not depend on $\vec{s}$. Now, fix $\vec{s}'$ as a reference sequence. Then for varying $\vec{s}$, $C(\vec{s}')$ is simply an additive constant. Consequently, up to an additive constant, this determines the return function that the learning system is incentivized to infer. We call it the *observation return function* since it is the return function based on the human's observations:

$$G_\Omega(\vec{s}) \coloneqq \big(\mathbf{B}\cdot G\big)\big(\vec{O}(\vec{s})\big).$$

This return function is not necessarily time-separable, but we assume that time-separability is not modeled correctly by the learning system. Now, define the resulting policy evaluation function J_Ω by

$$J_\Omega(\pi) \coloneqq \mathop{\mathbf{E}}_{\vec{s}\sim P^\pi(\vec{s})}\big[G_\Omega(\vec{s})\big].$$

This is the policy evaluation function that would be optimized if the learning system erroneously inferred the return function G_Ω.

Interlude: when the human knows the policy and is a Bayesian reasoner, then $J_\Omega = J$

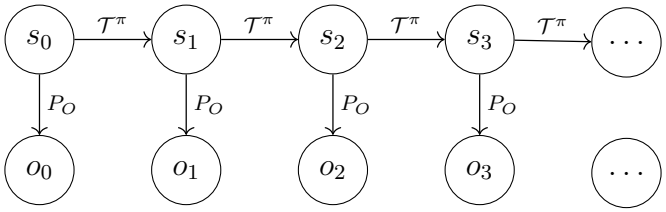
In this section, we briefly consider what would happen if in J_Ω, the human's belief B would make use of the true policy and be a rational Bayesian posterior as in Appendix C.3. We will show that under these conditions, we have $J_\Omega = J$. Since these are unrealistic assumptions, no other section depends on this result.

For the analysis, we drop the assumption that the observation sequence kernel $P_{\vec{O}}$ is deterministic, and assume that J_Ω is given as follows:

$$J_\Omega(\pi) \coloneqq \mathop{\mathbf{E}}_{\vec{s}\sim P^\pi(\vec{s})}\left[\mathop{\mathbf{E}}_{\vec{o}\sim P_{\vec{O}}(\vec{o}|\vec{s})}\left[\mathop{\mathbf{E}}_{\vec{s}'\sim B^\pi(\vec{s}'|\vec{o})}\big[G(\vec{s}')\big]\right]\right]. \tag{C.9}$$

[5] Note that in the case of non-deterministic observation kernels and choice probabilities given as in Equation (C.4), this argument does not work since the logarithm cannot be swapped with the outer expectation of the choice probabilities.

In this formula, $B^\pi(\vec{s} \mid \vec{o}) \coloneqq B(\vec{s} \mid \vec{o}, \pi)$ with B being the joint distribution from Appendix C.3. Formally, this is the posterior of the joint distribution $B(\vec{s}, \vec{o} \mid \pi)$ that is given by the following hidden Markov model:



(C.10)

Here, $\mathcal{T}^\pi(s' \mid s) \coloneqq \sum_{a \in \mathcal{A}} \mathcal{T}(s' \mid s, a) \cdot \pi(a \mid s)$. s_0 is sampled according to the known initial distribution $P_0(s_0)$. The human's posterior $B^\pi(\vec{s}' \mid \vec{o})$ is then the true posterior in this HMM. We obtain:

Proposition C.36. *Let π be a policy that is known to the human. Then $J_\Omega(\pi) = J(\pi)$.*

Proof. By Equation (C.9), we have

$$
\begin{aligned}
J_\Omega(\pi) &= \mathop{\mathbf{E}}_{\vec{s} \sim P^\pi(\vec{s})} \left[\mathop{\mathbf{E}}_{\vec{o} \sim P_{\vec{O}}(\vec{o} \mid \vec{s})} \left[\mathop{\mathbf{E}}_{\vec{s}' \sim B^\pi(\vec{s}' \mid \vec{o})} \left[G(\vec{s}') \right] \right] \right] \\
&\overset{(1)}{=} \sum_{\vec{s}} P^\pi(\vec{s}) \sum_{\vec{o}} P_{\vec{O}}(\vec{o} \mid \vec{s}) \sum_{\vec{s}'} B^\pi(\vec{s}' \mid \vec{o}) G(\vec{s}') \\
&\overset{(2)}{=} \sum_{\vec{s}'} \left[\sum_{\vec{o}} B^\pi(\vec{s}' \mid \vec{o}) \left[\sum_{\vec{s}} P_{\vec{O}}(\vec{o} \mid \vec{s}) P^\pi(\vec{s}) \right] \right] G(\vec{s}') \\
&\overset{(3)}{=} \sum_{\vec{s}'} \left[\sum_{\vec{o}} B^\pi(\vec{s}' \mid \vec{o}) B^\pi(\vec{o}) \right] G(\vec{s}') \\
&\overset{(4)}{=} \sum_{\vec{s}'} \left[\sum_{\vec{o}} P^\pi(\vec{s}') P_{\vec{O}}(\vec{o} \mid \vec{s}') \right] G(\vec{s}') \\
&\overset{(5)}{=} \sum_{\vec{s}'} P^\pi(\vec{s}') G(\vec{s}') \\
&\overset{(6)}{=} \sum_{\vec{s}} P^\pi(\vec{s}) G(\vec{s}) \\
&\overset{(7)}{=} J(\pi).
\end{aligned}
$$

In step (1), we wrote the expectations out in terms of sums. In step (2), we reordered them. In step (3), we observed that the inner sum over $\vec{s}$ evaluates to the marginal distribution $B^\pi(\vec{o})$ of the observation sequence $\vec{o}$ in the HMM in Equation (C.9). In step (4), we used Bayes rule in the inner sum. This is possible since $B^\pi(\vec{s}' \mid \vec{o})$ is the true posterior when π is known. In step (5), we pull $P^\pi(\vec{s}')$ out and notice that the remaining inner sum evaluates to

1. Step (6) is a relabeling and step (7) the definition of the true policy evaluation function J. □

Proof of Theorem 4.6

We first prove the following lemma.

Lemma C.37. *Let π and π_{ref} be two policies. If $J(\pi) < J(\pi_{ref})$ and $J_\Omega(\pi) > J_\Omega(\pi_{ref})$, then relative to π_{ref}, π must exhibit deceptive inflation, overjustification, or both.*

Proof. We start by establishing a quantitative relationship between the average overestimation and underestimation errors $\overline{E}^+$ and $\overline{E}^-$ as defined in Definition 4.3, the true policy evaluation function J, and the observation evaluation function J_Ω defined in Equation (4.4). Define $\Delta : \vec{\mathcal{S}} \to \mathbb{R}$ by $\Delta(\vec{s}) = G_\Omega(\vec{s}) - G(\vec{s})$, where G_Ω is as defined in Equation (4.3). Consider the quantity

$$E^+(\vec{s}) - E^-(\vec{s}) = \max\big(0, \Delta(\vec{s})\big) - \max\big(0, -\Delta(\vec{s})\big).$$

If $\Delta(\vec{s}) > 0$, then the first term is $\Delta(\vec{s})$ and the second one is 0. If $\Delta(\vec{s}) < 0$, then the first term is zero and the second one is $\Delta(\vec{s})$. If $\Delta(\vec{s}) = 0$, then both terms are zero. In all cases the right-hand side is equal to $\Delta(\vec{s})$. Unpacking the definition of Δ again, we have that for all $\vec{s}$,

$$E^+(\vec{s}) - E^-(\vec{s}) = G_\Omega(\vec{s}) - G(\vec{s}). \tag{C.11}$$

For any policy π, if we take the expectation of both sides of this equation over the on-policy distribution admitted by π, P^π, we get

$$\overline{E}^+(\pi) - \overline{E}^-(\pi) = J_\Omega(\pi) - J(\pi). \tag{C.12}$$

We now prove the lemma. Let π and π_{ref} be two policies, and assume that $J(\pi) < J(\pi_{\text{ref}})$ and $J_\Omega(\pi) \geq J_\Omega(\pi_{\text{ref}})$. Equivalently, we have $J_\Omega(\pi) - J_\Omega(\pi_{\text{ref}}) \geq 0$ and $J(\pi_{\text{ref}}) - J(\pi) > 0$, which we combine to state

$$\Big(J_\Omega(\pi) - J_\Omega(\pi_{\text{ref}})\Big) + \Big(J(\pi_{\text{ref}}) - J(\pi)\Big) > 0. \tag{C.13}$$

Rearranging terms yields

$$\Big(J_\Omega(\pi) - J(\pi)\Big) - \Big(J_\Omega(\pi_{\text{ref}}) - J(\pi_{\text{ref}})\Big) > 0.$$

These two differences inside parentheses are equal to the right-hand side of (C.12) for π and π_{ref}, respectively. We substitute the left-hand side of (C.12) twice to obtain

$$\Big(\overline{E}^+(\pi) - \overline{E}^-(\pi)\Big) - \Big(\overline{E}^+(\pi_{\text{ref}}) - \overline{E}^-(\pi_{\text{ref}})\Big) > 0.$$

Rearranging terms again yields

$$\left(\overline{E}^{+}(\pi) - \overline{E}^{+}(\pi_{\text{ref}})\right) + \left(\overline{E}^{-}(\pi_{\text{ref}}) - \overline{E}^{-}(\pi)\right) > 0. \tag{C.14}$$

If $\overline{E}^{+}(\pi) - \overline{E}^{+}(\pi_{\text{ref}}) > 0$ then we have $\overline{E}^{+}(\pi) > \overline{E}^{+}(\pi_{\text{ref}})$ and, by assumption, $J_\Omega(\pi) > J_\Omega(\pi_{\text{ref}})$. By Definition 4.4, this means π exhibits deceptive inflation relative to π_{ref}.

If $\overline{E}^{-}(\pi_{\text{ref}}) - \overline{E}^{-}(\pi) > 0$ then we have $\overline{E}^{-}(\pi) < \overline{E}^{-}(\pi_{\text{ref}})$ and, by assumption, $J(\pi) < J(\pi_{\text{ref}})$. By Definition 4.5, this means π exhibits overjustification relative to π_{ref}.

At least one of the two differences in parentheses in (C.14) must be positive, otherwise their sum would not be positive. Thus π must exhibit deceptive inflation relative to π_{ref}, overjustification relative to π_{ref}, or both. □

We can now combine earlier results to prove Theorem 4.6, repeated here for convenience:

Theorem C.38. *Assume that P_O is deterministic. Let π^*_Ω be an optimal policy according to a naive application of RLHF under partial observability, and let π^* be an optimal policy according to the true objective J. If π^*_Ω is not J-optimal, then relative to π^*, π^*_Ω must exhibit deceptive inflation, overjustification, or both.*

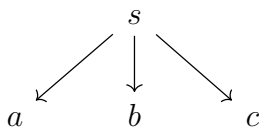
Proof. Because P_O is deterministic, π^*_Ω must be optimal with respect to J_Ω by Proposition 4.2 (proved in Appendix C.4). Thus $J_\Omega(\pi^*_\Omega) \geq J_\Omega(\pi^*)$. Since π^* is J-optimal and π^*_Ω is not, $J(\pi^*) < J(\pi^*_\Omega)$. By Lemma C.37, relative to π^*, π^*_Ω must exhibit deceptive inflation, overjustification, or both. □

Further examples supplementing Section 4.4

In this section, we present further mathematical examples supplementing those in Section 4.4. We found many of them before finding the examples we discuss in the main text, and show the same and additional conceptual features with somewhat less polish. We again assume that $P_{\vec{O}}$ is deterministic.

Example C.39. *In the main text, we have assumed a model where the human obeys Eq.* (4.2) *and showed that a naive application of RLHF can lead to suboptimal policies, and the specific failure modes of deceptive inflation and overjustification. What if the human makes the choices in a different way? Specifically, assume that all we know is that $P^R(\vec{o} \succ \vec{o}') + P^R(\vec{o}' \succ \vec{o}) = 1$. Can the human generally choose these choice probabilities in such a way that RLHF is incentivized to infer a reward function whose optimal policies are also optimal for R? The answer is no.*

Take the following example:



In this example, there is a fixed start state s and three actions a, b, c that also serve as the final states. The time horizon is $T = 1$, so the only state sequences are sa, sb, sc. Assume $\mathcal{T}(a \mid s, a) = 1$, $\mathcal{T}(b \mid s, b) = 1$, $\mathcal{T}(c \mid s, c) = 1 - \epsilon$, $\mathcal{T}(a \mid s, c) = \epsilon$, i.e., selecting action c sometimes leads to state a. Also, assume $a = O(a) \neq O(b) = O(c) =: o$ and $R(a) = R(b) < R(c)$.

Since b and c have the same observation o, the human choice probabilities do not make a difference between them, and so RLHF is incentivized to infer a reward function $\tilde{R}$ with $\tilde{R}(b) = \tilde{R}(c) =: \tilde{R}(o)$. If $\tilde{R}(o) > \tilde{R}(a)$, then the policy optimal under $\tilde{R}$ will produce action b since this deterministically leads to observation o, whereas c does not. If $\tilde{R}(o) < \tilde{R}(a)$, then the policy optimal under $\tilde{R}$ will produce action a. In both cases, the resulting policy is suboptimal compared to π^*, which deterministically chooses action c.

In the coming examples, it will also be useful to look at the *misleadingness* of state sequences:

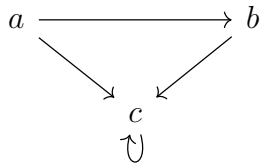
Definition C.40 (Misleadingness)**.** *Let $\vec{s} \in \vec{\mathcal{S}}$ be a state sequence. Then its* misleadingness *is defined by*

$$\mathrm{M}(\vec{s}) \coloneqq G_\Omega(\vec{s}) - G(\vec{s}) = \mathop{\mathbf{E}}_{\vec{s}' \sim B(\vec{s}' \mid \vec{O}(\vec{s}))} \big[G(\vec{s}') - G(s)\big].$$

We call a state sequence positively misleading *if $M(\vec{s}) > 0$, which means the sequence appears better than it is, and* negatively misleading *if $\mathrm{M}(\vec{s}) < 0$. The* misleadingness vector *is given by $\mathrm{M} \in \mathbb{R}^{\vec{\mathcal{S}}}$.*

Note that the misleadingness is related to E^+ and E^-, as defined in Definition 4.3: If $\mathrm{M}(\vec{s}) > 0$ then $\mathrm{M}(\vec{s}) = E^+(\vec{s})$, and if $\mathrm{M}(\vec{s}) < 0$ then $\mathrm{M}(\vec{s}) = -E^-(\vec{s})$.

Example C.41. *In this example, we assume the human is a Bayesian reasoner as in Appendix C.3. Consider the MDP that is suggestively depicted as follows:*



The MDP has states $\mathcal{S} = \{a, b, c\}$ and actions $\mathcal{A} = \{b, c\}$. The transition kernel is given by $\mathcal{T}(c \mid a, c) = 1$ and $\mathcal{T}(b \mid a, b) = 1$, meaning that the action determines whether to transition from a to b or c. All other transitions are deterministic and do not depend on the action, as depicted. We assume an initial state distribution P_0 over states with probabilities $p_a = P_0(a), p_b = P_0(b), p_c = P_0(c)$. The true reward function $R \in \mathbb{R}^{\{a,b,c\}}$ and discount factor $\gamma \in [0, 1)$ are, for now, kept arbitrary. The time horizon is $T = 2$, meaning we have four possible state sequences acc, abc, bcc, ccc.

Furthermore, assume that $o \coloneqq O(a) = O(b) \neq O(c) = c$, i.e., c is observed and a and b are ambiguous.

Finally, assume that the human has a policy prior $B(\lambda)$, where $\lambda = \pi_\lambda(c \mid a)$ is the likelihood that the policy chooses action c when in state a, which is a parameter that determines the entire policy.

We claim the following:

1. *If $p_b \neq \gamma \cdot \mathbf{E}_{\lambda\sim B(\lambda)}[\lambda] \cdot p_a$, then $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$, so there is no return function ambiguity under appropriately modeled partially observable RLHF, see Corollary C.7.*
2. *There are true reward functions R for which optimizing J_Ω leads to a suboptimal policy according to the true policy evaluation function J, a case of misalignment. Thus, a naive application of RLHF under partial observability fails, see Section 4.4.*
3. *The failure modes are related to hiding negative information (deception) and purposefully revealing information while incuring a loss (overjustifying behavior).*

Proof. Write $p \coloneqq B(bcc \mid occ)$, the human's posterior probability of state sequence bcc for observation sequence occ. We have $1 - p = B(acc \mid occ)$.

Consider the linear operators $\mathbf{\Gamma} : \mathbb{R}^{\{a,b,c\}} \to \mathbb{R}^{\{abc,bcc,ccc,acc\}}$ and $\mathbf{B} : \mathbb{R}^{\{abc,bcc,ccc,acc\}} \to \mathbb{R}^{\{ooc,occ,ccc\}}$ defined in the main text. When ordering the states, state sequences, and observation sequences as we just wrote down, we obtain

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & \gamma & \gamma^2 \\ 0 & 1 & \gamma + \gamma^2 \\ 0 & 0 & 1 + \gamma + \gamma^2 \\ 1 & 0 & \gamma + \gamma^2 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & p & 0 & 1-p \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{B} \circ \mathbf{\Gamma} = \begin{pmatrix} 1 & \gamma & \gamma^2 \\ 1-p & p & \gamma + \gamma^2 \\ 0 & 0 & 1 + \gamma + \gamma^2 \end{pmatrix}.$$

By Corollary C.7, if $\mathbf{B} \circ \mathbf{\Gamma}$ is injective, then there is no reward function ambiguity. Clearly, this is the case if and only if $p \neq \gamma \cdot (1 - p)$. From Bayes rule, we have

$$p = \frac{B(bcc)}{B(acc) + B(bcc)}, \quad 1 - p = \frac{B(acc)}{B(acc) + B(bcc)}.$$

So the condition for injectivity holds if and only if

$$B(bcc) \neq \gamma \cdot B(acc).$$

Now, notice

$$B(bcc) = \int_\lambda B(\lambda) \cdot B(bcc \mid \lambda) d\lambda = \int_\lambda B(\lambda) \cdot p_b d\lambda = p_b$$

and

$$B(acc) = \int_\lambda B(\lambda) B(acc \mid \lambda) d\lambda = \int_\lambda B(\lambda) \cdot p_a \cdot \lambda d\lambda = p_a \cdot \mathop{\mathbf{E}}_{\lambda\sim B(\lambda)}\big[\lambda\big].$$

This shows the first result.

For the second statement, we explicitly compute J_Ω up to an affine transformation, which does not change the policy ordering. Let R be the true reward function, $G = \mathbf{\Gamma}(R)$

the corresponding return function, and $\mathbf{B}(G)$ the resulting return function at the level of observations. For simplicity, assume $R(c) = 0$, which can always be achieved by adding a constant. We have:

$$\begin{aligned} J_\Omega(\lambda) &= \mathop{\mathbf{E}}_{\vec{s}\sim P^\lambda(\vec{s})}\Big[\mathbf{B}(G)\big(\vec{O}(\vec{s})\big)\Big] \\ &= P^\lambda(abc)\cdot\mathbf{B}(G)(ooc) + P^\lambda(bcc)\cdot\mathbf{B}(G)(occ) \\ &\quad + P^\lambda(ccc)\cdot\mathbf{B}(G)(ccc) + P^\lambda(acc)\cdot\mathbf{B}(G)(occ) \\ &= p_a\cdot(1-\lambda)\cdot G(abc) + p_b\cdot\mathbf{B}(G)(occ) \\ &\quad + p_c\cdot G(ccc) + p_a\cdot\lambda\cdot\mathbf{B}(G)(occ) \\ &\propto \lambda\cdot\Big[\mathbf{B}(G)(occ) - G(abc)\Big]. \end{aligned}$$

We have

$$G(abc) = R(a)+\gamma R(b), \quad \mathbf{B}(G)(occ) = (1-p)\cdot G(acc)+p\cdot G(bcc) = (1-p)\cdot R(a)+p\cdot R(b).$$

Thus, the condition $\mathbf{B}(G)(occ) > G(abc)$ is equivalent to

$$R(a) < \frac{p-\gamma}{p}\cdot R(b).$$

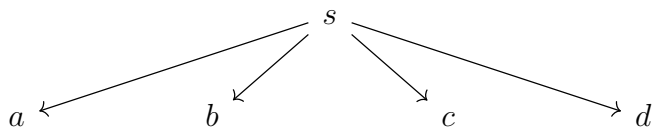
Thus, we have

$$\arg\max_{\lambda\in[0,1]} J_\Omega(\lambda) = \begin{cases} 1, \text{ if } R(a) < \frac{p-\gamma}{p}\cdot R(b), \\ 0, \text{ else.} \end{cases}$$

Now consider the case $R(b) > 0$. In this case, $\lambda = 0$ gives rise to the optimal policy according to G since going to b gives extra reward that one misses when going to c directly. However, when $R(a) \ll 0$, then J_Ω selects for $\lambda = 1$. Intuitively, the policy tries to "hide that the episode started in a" by going directly to c, which leads to ambiguity between acc and bcc. This is a case of deceptive inflation as in Theorem 4.6.

Now, consider the case $R(b) < 0$. In this case, $\lambda = 1$ gives rise to the optimal policy according to G. However, when $R(a) \gg 0$, then J_Ω selects for $\lambda = 0$. Intuitively, the policy tries to "reveal that the episode started with a" by going to b, which is positive information to the human, but negative from the perspective of optimizing G. As in Theorem 4.6, we see that this is a case of overjustification. □

Example C.42. *In this example, we consider an MDP that's similar to a multi-armed bandit with four states/actions a, b, c, d and observation kernel $O(a) = O(b) \neq O(c) = O(d)$. Formally, we can imagine that it is given by the MDP*



with $R(s) = 0$ and a time-horizon of $T = 1$. In this example, we reveal that misleadingness and non-optimality (according to the true reward R, or J) are in principle orthogonal concepts. We consider the following four example cases. In each one, we vary some environment parameters and then determine a^*_Ω, the action that results from optimizing J_Ω (corresponding to a naive application of RLHF under partial observability, see Section 4.4), its misleadingness $\mathrm{M}(a^*_\Omega)$ (see Definition C.40), and the action a^* that would result from optimizing J. If $a^*_\Omega = a^*$, then J_Ω selects for the optimal action. For simplicity, we can imagine that the human has a uniform prior over what action results eventually (out of the action taken and potentially a deviation defined by ϵ, see below) is taken before making an observation, i.e. $B(a) = B(b) = B(c) = B(d) = \frac{1}{4}$.

(a) Assume $R(a) > R(c) > R(d) \gg R(b)$. Also assume that action d leads with probability $\epsilon > 0$ to state b, whereas all other actions lead deterministically to the specified state. Then $a^*_\Omega = c$, $\mathrm{M}(c) < 0$ and $a^* = a$.

(b) Assume $R(d) > R(a) > R(c) \gg R(b)$. Again, assume there is a small probability $\epsilon > 0$ that action d leads to state b. Then $a^*_\Omega = c$, $\mathrm{M}(c) > 0$, and $a^* = d$ or $a^* = a$, depending on the size of ϵ.

(c) Assume $R(a) > R(b) > R(c) > R(d)$. Additionally, assume that there is a large *probability $\epsilon > 0$ that action a leads to state d, whereas all other actions lead to what's specified. If ϵ is large enough, then $a^* = b$. Additionally, we have $a^*_\Omega = b$ and $\mathrm{M}(b) > 0$.*

(d) Assume $R(a) > R(b) > R(c) > R(d)$. Also, assume some probability $\epsilon > 0$ that action b leads to state d, whereas all other actions lead deterministically to what's specified. Then $a^*_\Omega = a$, $\mathrm{M}(a) < 0$, and $a^* = a$.

Overall, we notice:

- *Example (a) shows a high regret and negative misleadingness of $a^*_\Omega = c$. The action is better then it seems, but action a would be better still but cannot be selected because it can be confused with the very bad action b.*
- *Example (b) shows a high regret and high misleadingness of $a^*_\Omega = c$. The action is worse than it seems and also not optimal.*
- *Example (c) shows zero regret and high misleadingness of $a^*_\Omega = b$. The action is worse than it seems because it can be confused with a, but it is still the optimal action because a can turn into d.*
- *Example (d) shows zero regret negative misleadingness of $a^*_\Omega = a$. The action is chosen even though it seems worse than it is, and is also optimal.*

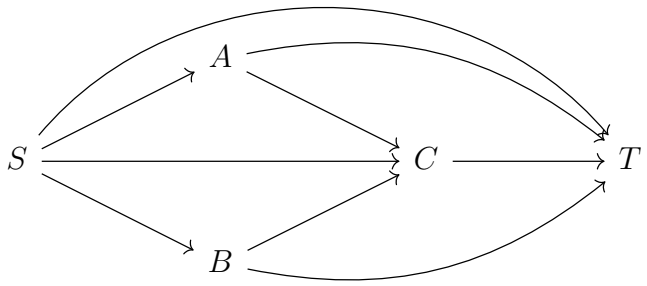
Thus, we showed all combinations of regret and misleadingness of the action optimized for under J_Ω.

We can also notice the following: Examples (a) and (b) only differ in the placement of $R(d)$. In particular, the reason *that $a^*_\Omega = c$ is structurally the same in both, but the misleadingness changes. This indicates that misleadingness is not* on its own *contributing to what J_Ω optimizes for.*

The following is the smallest example we found with the following properties:

- There is a unique start state and terminal state.
- A naive application of RLHF fails in a way that shows deception and overjustification.
- Modeling partial observability resolves the problems.

Example C.43. *Consider the following graph:*



This depicts an MDP with start state S, terminal state T and possible state sequences $STTT, SATT, SACT, SCTT, SBCT, SBTT$ and no discount, i.e. $\gamma = 1$. Assume that S, B, C are observed, i.e. $O(S) = S$, $O(B) = B$, $O(C) = C$, and that A and T are ambiguous: $O(A) = O(T) = X$. Then there are five observation sequences $SXXX$, $SXCX$, $SCXX$, $SBCX$, $SBXX$. Assume that the human can identify all observation sequences except $SXXX$, with belief $b = B(STTT \mid SXXX)$ and $1 - b = B(SATT \mid SXXX)$.

Then the return function is identifiable under these conditions when the human's belief is correctly modeled. However, for some choices of the true reward function R and transition dynamics of this MDP, we can obtain deceptive or overjustified behavior for a naive application of RLHF.

Proof. We apply Corollary C.7. We order states, state sequences, and observation sequences as follows:

$$\begin{aligned}
\mathcal{S} &= S, A, B, C, T, \\
\vec{\mathcal{S}} &= STTT, SATT, SACT, SCTT, SBCT, SBTT, \\
\vec{\Omega} &= SXXX, SXCX, SCXX, SBCX, SBXX.
\end{aligned}$$

As can easily be verified, with this ordering the matrices $\mathbf{B} \in \mathbb{R}^{\vec{\Omega} \times \vec{\mathcal{S}}}$ and $\mathbf{\Gamma} \in \mathbb{R}^{\vec{\mathcal{S}} \times \mathcal{S}}$ are given by:

$$\mathbf{B} = \begin{pmatrix} b & 1-b & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{\Gamma} = \begin{pmatrix} 1 & 0 & 0 & 0 & 3 \\ 1 & 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 2 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 2 \end{pmatrix}.$$

To show identifiability, we need to show that $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$. Clearly, the kernel of $\mathbf{B}$ is given by all return functions in $\mathbb{R}^{\vec{\mathcal{S}}}$ that are multiples of $G' = (b-1, b, 0, 0, 0, 0)$. Assume $G' \in \operatorname{im} \mathbf{\Gamma}$, meaning there is a reward function $R' \in \mathbb{R}^{\vec{\mathcal{S}}}$ with $\mathbf{\Gamma} \cdot R' = G'$. We need to deduce from this a contradiction. The assumption means we obtain the following equations:

$$\begin{aligned}
&(i) \quad R'(S) + 3R'(T) = b - 1, \\
&(ii) \quad R'(S) + R'(A) + 2R'(T) = b, \\
&(iii) \quad R'(S) + R'(A) + R'(C) + R'(T) = 0, \\
&(iv) \quad R'(S) + R'(C) + 2R'(T) = 0, \\
&(v) \quad R'(S) + R'(B) + R'(C) + R'(T) = 0 \\
&(vi) \quad R'(S) + R'(B) + 2R'(T) = 0
\end{aligned}$$

(iii) and (v) together imply $R'(A) = R'(B)$; (iv) and (vi) together imply $R'(B) = R'(C)$; (v) and (vi) together imply $R'(C) = R'(T)$; so together, we have $R'(A) = R'(T)$. Thus, replacing $R'(A)$ in (ii) by $R'(T)$ and comparing (i) and (ii), we obtain $b - 1 = b$, a contradiction. Overall, this shows $\ker \mathbf{B} \cap \operatorname{im} \mathbf{\Gamma} = \{0\}$, and thus identifiability of the return function by Corollary C.7.

Now we investigate the case of unmodeled partial observability.

For demonstrating overjustification, assume deterministic transition dynamics in which every arrow in the diagram can be chosen by the policy. Also, assume $R(A) \ll 0$, $R(T) > 0$, $R(S) = 0$, $R(B) = 0$, and $R(C) = 0$. Then the optimal policy chooses the state sequence $STTT$. However, this trajectory has low observation value since $G_\Omega(STTT) = (\mathbf{B} \cdot G)(SXXX) = bG(STTT) + (1-b)G(SATT)$, which is low since $R(A) \ll 0$. J_Ω then selects for the suboptimal policies choosing $SBTT$ or $SCTT$, which is overjustified behavior that makes sure that the human does not think state A was accessed.

For demonstrating deception, assume that $R(A) \gg 0$, $R(T) < 0$, $R(S) = R(B) = R(C) = 0$ and that the transition dynamics are such that when the policy *attempts* to transition from S to A, it will sometimes transition to B, with all other transitions deterministic. In this case, the optimal behavior attempts to enter state A since this has very high value. J_Ω, however, will select for the policy that chooses $STTT$. This is deceptive behavior. □