Corrigible Assistance in One Round: Pragmatic–Pedagogic Best Response

Elle Lazarski[1] and Jaime Fernández Fisac[1]

Princeton University, Princeton NJ 08544, USA

Abstract. Assistance games formalize human–robot collaboration under asymmetric information: the human knows the goal, while the robot must infer it from observation and interaction in order to assist effectively. In general, computing optimal assistance game strategies online is intractable, since exact solutions require planning in a POMDP. We identify a nontrivial class of assistance games in which pragmatic–pedagogic reasoning resolves goal uncertainty in a single time step, reaching the fixed-point, infinite-horizon equilibrium of the full assistance game through a tractable best-response procedure. Within this class, we show that mainstream inverse optimal control exhibits an inference ceiling that hinders alignment, while pragmatic–pedagogic reasoning breaks the ceiling by turning actions otherwise equivalent under task execution into unambiguous goal-identifying signals. Finally, we validate our theoretical results and proposed method on a simple collaborative block-building example.

Keywords: Human–Robot Interaction · Value Alignment · Mathematical Modeling and Analysis

1 Introduction

As robots become more capable and are deployed in increasingly varied contexts, they must infer and adapt to their users' needs online rather than execute prespecified routines. In artificial intelligence (AI), conventional alignment pipelines like Reinforcement Learning from Human Feedback (RLHF) optimize for human approval of generated outputs, such as through thumbs-up or A/B preference feedback [1]. However, this signal is known to be an imperfect, often problematic proxy, since it tends to neglect the downstream effects of individual decisions [2–4]. In human–robot interaction, the coupling between robot operation and human behavior over time makes it all the more necessary to approach alignment with respect to long-term outcomes instead of isolated robot actions [5–8].

To capture the temporal dimension of alignment, *assistance games* [9, 10] pose the problem as a dynamic two-player collaboration between a human user and a robot assistant, who aims to help realize the human's objective but is *uncertain* about what it is. The resulting equilibrium solutions have been shown to present desirable properties, which extend the notion of optimal information seeking to the two-player setting. In particular, the human's actions often

carry out a *pedagogic* function, strategically conveying information about the goal, while the robot's responses are *pragmatic*, interpreting human cues as purposefully (rather than circumstantially) communicative [11, 12], consistent with modern cognitive science accounts of human teaching and learning [13]. Unfortunately, despite their theoretical strengths, assistance game solutions are largely considered computationally intractable, due to the need to plan in the robot's information space [14].

In this work, we identify a special—but nontrivial—class of assistance games in which pragmatic–pedagogic solutions are simultaneously *highly effective* and *easily computable*, fully disambiguating the human's goal in a single time step. As a result, for this class of problems, an efficient short-horizon approximation (analogous to single-player QMDP [15]) yields an optimal strategy pair for the full-horizon game. We further show that this equilibrium can be readily obtained in a single round of player best responses starting from a mainstream inverse optimal control (IOC)—or inverse reinforcement learning (IRL)—solution. Crucially, however, while IOC often exhibits an *inference ceiling* that impedes one-step alignment, the pragmatic–pedagogic solution fully overcomes this limitation, turning actions that are ambiguous from the standpoint of task execution alone into unambiguous goal-identifying signals. As a result, the pragmatic robot strategy is *maximally empowering*, affording the human the option to convey and achieve any goal regardless of the robot's initial belief, and rendering the robot's operation *corrigible*, a key requirement for robust alignment and long-term safety [16].

Our contributions in this work can be summarized as follows:

1. **One-round convergence to a pragmatic–pedagogic equilibrium.** We quantify the IOC inference ceiling in terms of a posterior belief bound, show that positive pedagogic leverage breaks it, and characterize the class of *action-separable* assistance games—those in which leverage emerges for every goal. For this class, we prove that one round of best responses reaches the fixed-point, infinite-horizon equilibrium of the assistance game.
2. **A practical methodology for tractable pragmatic assistance.** We propose Pragmatic–Pedagogic Best Response (PPBR), an algorithm that directly computes the pragmatic–pedagogic equilibrium of action-separable assistance games, yielding a human-empowering robot strategy.
3. **Empirical validation.** We validate our theoretical results empirically in a collaborative block-building domain, comparing IOC against PPBR under a comprehensive range of initial conditions.

2 Related Work

Traditional IOC/IRL aims to recover the objective pursued by a human "expert" by observing demonstrations of their behavior, assumed to take place in isolation [17–19]. In interactive settings, however, the robot is not a passive observer but an active player whose own actions may affect the human's outcome. The

human may therefore choose actions not only to directly generate utility but also to indirectly improve expected outcomes by influencing the robot.

Assistance games, initially introduced under the name cooperative inverse reinforcement learning (CIRL), formalize this idea as a cooperative game with asymmetric information [9, 10]. In particular, the human knows the true goal, while the robot must infer it from observation and interaction. Solving a CIRL game can be reduced to solving a partially observable Markov decision process (POMDP), in which the robot's belief over goals serves as a sufficient statistic for optimal decision-making [10]. However, this reduction inherits the computational burden of POMDP planning, barring its use in practical problems.

Subsequent research in assistance games showed that optimal human–robot strategy pairs must satisfy a specific dynamic programming relation expressed as a fixed point in a *pragmatic–pedagogic Bellman recursion*, whereby the human chooses actions pedagogically to convey strategically relevant information to a suitably attuned robot partner, and the robot in turn interprets human actions pragmatically, treating them as cues purposefully chosen to convey information [11]. Despite an exponential complexity improvement over the naïve POMDP reduction, the resulting belief-space planning is still intractable for runtime computation of assistance strategies.

Here, we build on the theoretical insights established by prior assistance game efforts to investigate runtime-computable robot strategies that preserve cooperative structure while enabling online assistance. We combine classical short-horizon approximation ideas from decision theory [15] and truncated iterated-best-response approximations from behavioral game theory [20–22], and show that they allow us to exactly solve the full-horizon assistance game for a class of problems in which the robot's uncertainty can be resolved in a single time step.

Finally, some work in the AI alignment literature has brought into question the usefulness of pragmatic robot behavior, due to potentially higher sensitivity to modeling assumptions (in particular, whether or not the human user intends to behave pedagogically) [23]. Our results shed new light on the matter, suggesting an alternative perspective: regardless of modeling accuracy, robot strategies derived from pragmatic–pedagogic solutions make the robot comparatively more responsive to human action cues, increasing the human's effective controllability over outcomes and mitigating known overconfidence issues with non-pragmatic (sometimes called "literal") goal-inferring robots (typically IOC) [7, 24, 25].

3 Problem Formulation: Assistance Games

We study assistance games $\mathcal{G}$ in which a human and a robot take turns acting in a shared environment to achieve the human's objective. Critically, this objective is known to the human but, a priori, unknown to the robot. Let

$$\mathcal{G} = \langle \mathcal{S}, \{\mathcal{A}^H, \mathcal{A}^R\}, T_s, \{\Theta, R\}, P_0, \gamma \rangle,$$

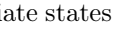
where $\mathcal{S}$ denotes the space of world states; $\mathcal{A}^H$, $\mathcal{A}^R$ are the human and robot action spaces; Θ is a set of possible goal parameters encoding the human's objective; $T_s(s_{t+1} \mid s_t, a_t^H, a_t^R)$ is the transition probability measure; $R(s_t, a_t^H, a_t^R; \theta)$

is the shared reward function, parameterized by $\theta \in \Theta$ (whose value is only observed by the human); $P_0(s_0, \theta)$ is the initial joint distribution over states and goals; and $\gamma \in [0, 1]$ is a discount factor. At each time step t, the human selects $a_t^H \in \mathcal{A}^H$ first, after which the robot observes a_t^H and selects $a_t^R \in \mathcal{A}^R$; the joint action (a_t^H, a_t^R) then induces a successor state through T_s.

The robot maintains a Bayesian belief $b_t \in \Delta(\Theta)$ over candidate goal hypotheses, which is updated after each observed a_t^H. Following the assistance game literature, P_0 is known to both players, which means that b_t can be computed by *either* player as a sufficient statistic of the robot's information state after observing $(s_0, a_0^H, a_0^R, \ldots, s_t, a_t^H)$ under any particular human policy. Since the robot's belief may inform its behavior, it is *also* strategically relevant to the human. Therefore, the human's optimal assistance game strategy will in general be a stochastic policy $\pi^H(a_t^H \mid s_t, b_t; \theta)$, whereas the robot—not privy to the true θ, but observing a_t^H before selecting its own action—must choose a policy $\pi^R(a_t^R \mid s_t, b_t, a_t^H)$. Solving the assistance game amounts to finding a team strategy $\pi = (\pi^H, \pi^R)$ that maximizes the expected time-discounted return

$$J(\pi^H, \pi^R) := \mathop{\mathbb{E}}_{\substack{(s_0,\theta)\sim P_0 \\ \tau\sim(\pi,T)}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t^H, a_t^R; \theta) \right],$$

where $\tau := (s_0, a_0^H, a_0^R, s_1, a_1^H, a_1^R, \ldots)$ denotes a trajectory, and the distribution over trajectories is sequentially induced by policies $\pi = (\pi^H, \pi^R)$ and (s_t, b_t) transitions $T = (T_s, T_b)$.

Running example: We consider a collaborative block-building task in which the human's latent goal $\theta^\star \in \{$  $\}$ specifies a target Tetris shape. The human is indifferent to where the structure is built and what direction it faces, as long as it is upright. To avoid known action-multiplicity artifacts in Boltzmann likelihood models [26], we represent states and actions in the quotient spaces induced by the problem's underlying symmetries: states that differ only by an SE(2) transform (translation and rotation on the ground plane) are treated as equivalent, as are actions that induce equivalent state changes. In our two-goal example, the quotient action space $\mathcal{A}(s)$ in most intermediate states of interest ($s =$  , ...) contains two placements (SIDE/$\leftarrow$ and UP/$\uparrow$) and possibly the destruction of an existing block with nothing on top ($\nleftarrow$ or $\updownarrow$).

Suboptimal play: While other actions are possible in principle (e.g., placing a new block far away from the existing ones), we exclude them from our analysis because they are *exponentially* less likely under all hypotheses. A robot observing such actions would typically lose situational confidence and choose not to act [7].

4 Approach: One Time Step, One Best-Response Round

4.1 Alignment in One Time Step

Extending the QMDP approximation in single-agent POMDPs to the two-player setting, we decompose the assistance game's horizon into a present uncertain

phase and a hypothesized later phase in which the robot has gained full certainty about the human's goal. Although optimistic in general, this approximation is in fact exact whenever both players can independently arrive at an unambiguous correspondence between goals and actions, so the robot *will* become fully confident after a single time step—as we show in the next section.

Perfect-Information Subgame. To support the QMDP decomposition, we introduce two oracle value functions that characterize the fully observed phase of planning. First, the *solipsistic value* $V^{H0}(s_t;\theta)$ represents the maximum expected reward-to-go from state s_t when the human works alone, taking one action per time step with no robot assistance. This single-player MDP over $\mathcal{A}^H$ yields the solipsistic state–action value

$$Q^{H0}(s_t, a_t^H;\theta) := r(s_t, a_t^H;\theta) + \gamma\, \mathbb{E}\big[V^{H0}(s_{t+1};\theta)\big],$$

with $s_{t+1} \sim T(\cdot \mid s_t, a_t^H)$.

Second, the *team value* $V^{\text{team}}(s_t;\theta)$ represents the maximum expected reward-to-go from state s_t when both players are privy to the human's true goal and coordinate optimally. This is computed by solving a fully observed MDP with joint action space $\mathcal{A}^H \times \mathcal{A}^R$, yielding the team state–action value

$$Q^{\text{team}}(s_t, a_t^H, a_t^R;\theta) := r(s_t, a_t^H, a_t^R;\theta) + \gamma\, \mathbb{E}\big[V^{\text{team}}(s_{t+1};\theta)\big],$$

with $s_{t+1} \sim T(\cdot \mid s_t, a_t^H, a_t^R)$.

4.2 Equilibrium in One Round

We operationalize pragmatic–pedagogic inference from the perspective of a robot helper that reasons about human behavior using a truncated hierarchy of *hypothesized* player models. Specifically, the level-3 robot, denoted R3, simulates H0, R1, and H2 as follows.

H0: Solipsistic Human Expert. Under a given hypothesis θ, H0 acts solely to optimize for that objective without considering how their actions influence the robot's belief or behavior. Hence, H0 serves as the "expert demonstrator" in classical IOC/IRL methods. Define

$$\pi^{H0}(a_t^H \mid s_t, \theta) \propto \exp\big(\beta^H\, Q^{H0}(s_t, a_t^H;\theta)\big),$$

where $\beta^H > 0$ is an inverse temperature or "rationality" parameter controlling how strongly H0 favors high-value actions under θ: as $\beta^H \to 0$, the policy approaches a uniform distribution over all actions; as $\beta^H \to \infty$, the policy concentrates on optimal actions, converging to a uniform distribution over only the $\arg\max$ set.

R1: Naïve Robot Learner. R1 assumes the human behaves as H0 and updates its belief b_t by Bayes' rule under likelihood $\pi^{H0}(a_t^H \mid s_t, \theta)$, yielding $\tilde{b}_{t+}(\cdot \mid a_t^H)$. Define $Q^{R1}(s_t, a_t^R; \tilde{b}_{t+}) := \mathbb{E}_{\theta\sim\tilde{b}_{t+}}\big[r(s_t, a_t^R;\theta) + \gamma\, V^{H0}(s_{t+1};\theta)\big]$, and

$$\pi^{R1}(a_t^R \mid s_t, a_t^H, \tilde{b}_{t+}) \propto \exp\big(\beta^R\, Q^{R1}(s_t, a_t^R; \tilde{b}_{t+})\big),$$

where $\beta^R > 0$ is the robot's inverse-temperature parameter. The use of V^{H0} rather than V^{team} as the continuation value is a deliberate modeling choice: it reflects the IOC/IRL worldview that the human acts in isolation, without anticipating robot assistance. Under this assumption, R1's help is one-shot; it may assist at the current time step, but its model of the future is one in which the human continues to act alone.

H2: Pedagogic Human User. H2 anticipates R1's belief update and subsequent behavior. For each candidate human action a_t^H, H2 simulates its downstream consequences: R1 updates its belief assuming H0 behavior and then selects a response, which H2 evaluates under each goal. Define $Q^{H2}(s_t, a_t^H; \theta, b_t) := \mathbb{E}_{a_t^R \sim \pi^{R1}(\cdot|s_t, a_t^H, \tilde{b}_{t+}(\cdot|a_t^H))}\big[Q^{\text{team}}(s_t, a_t^H, a_t^R; \theta)\big]$, and

$$\pi^{H2}(a_t^H \mid s_t, b_t, \theta) \propto \exp\big(\beta^H\, Q^{H2}(s_t, a_t^H; \theta, b_t)\big).$$

R3: Pragmatic Robot Helper. After observing a_t^H, R3 updates its belief b_t by Bayes' rule using the H2 likelihood model $\pi^{H2}(a_t^H \mid s_t, b_t, \theta)$, yielding b_{t+}. It then selects its response—which is executed in the real environment—by maximizing expected value under the posterior:

$$a_t^{R3\star} := \arg\max_{a_t^R \in \mathcal{A}^R} \mathop{\mathbb{E}}_{\theta \sim b_{t+}} \big[Q^{\text{team}}(s_t, a_t^H, a_t^R; \theta)\big].$$

We will later show that, for the problem class we consider, truncating the hierarchy at R3 already reaches the pragmatic–pedagogic fixed-point equilibrium. Thus, there is no need to iterate further levels (H4, R5, and so on).

5 Analysis: Pragmatic–Pedagogic One-Step Alignment

In this section, we establish important properties of the R3–H2 solution and the conditions under which it allows us to exactly recover the pragmatic–pedagogic equilibrium of the full assistance game.

5.1 IOC Inference Ceiling and Incorrigibility

We begin by examining the fundamental alignment limitations of IOC/IRL in the assistance context. Let

$$\mathcal{A}^{H0\star}(s_t, \theta) := \arg\max_{a_t^H \in \mathcal{A}^H} Q^{H0}(s_t, a_t^H; \theta)$$

denote the set of H0-optimal actions under goal θ in state s_t. In general, an IOC posterior ceiling appears whenever one goal's H0-optimal action set is a *subset* of the H0-optimal action set for another goal. As such, every H0-optimal action in the smaller-set goal is also optimal for the larger-set goal, so the human cannot unambiguously communicate the smaller-set goal in one step.

Two Candidate Goals We first formalize this pathology in the simplest possible setting with only two candidate goal hypotheses.

Lemma 1 (Two-goal IOC inference ceiling). *Suppose $a^\dagger \in \mathcal{A}^H$ is H0-optimal for only two goals $\theta^\star, \theta'$, with $a^\dagger \in \mathcal{A}^{H0\star}(s_t, \theta^\star) \subseteq \mathcal{A}^{H0\star}(s_t, \theta')$. Let $|\mathcal{A}^{H0\star}(s_t, \theta^\star)| = k$ and $|\mathcal{A}^{H0\star}(s_t, \theta')| = m$, where $1 \leq k \leq m$. Then, R1's posterior belief in goal $\theta^\star$ after observing $a^\dagger$ is bounded above by the limit*

$$b^{\infty}_{R1}(\theta^\star) := \lim_{\beta^H \to \infty} b^{+}_{R1}(\theta^\star \mid a^\dagger) = \frac{m\, b^{-}(\theta^\star)}{m\, b^{-}(\theta^\star) + k\, b^{-}(\theta')}, \tag{1}$$

which bounds $b^{+}_{R1}(\theta^\star \mid a^\dagger)$ strictly below 1.

Proof. As $\beta^H \to \infty$, $\pi^{H0}(a^\dagger \mid s_t, \theta^\star) \to \frac{1}{k}$ and $\pi^{H0}(a^\dagger \mid s_t, \theta') \to \frac{1}{m}$. Substituting these likelihoods into the R1 Bayes update and normalizing yields (1). □

Remark 1. If the two goals assign equal H0 likelihood to $a^\dagger$, R1 learns no new information to disambiguate between $\theta^\star$ and θ'. In particular, when the optimal action sets are identical, (1) leaves the prior unaltered, yielding $\lim_{\beta^H \to \infty} b^{+}_{R1}(\theta^\star \mid a^\dagger) = b^{-}(\theta^\star)$.

Running example: The human's first action—placing the anchor block—is necessarily uninformative to the robot. Since by assumption the human is indifferent to position and north–south–east–west orientation, all feasible initial block placements belong to the same equivalence class, and placing a block is the unique optimal action for any $\theta \in \{$  ,  $\}$ (cf. Remark 1).

Since both goals require one block above and one block adjacent to the anchor, either robot follow-up $a^R_t \in \{\leftarrow, \uparrow\}$ has the same expected value. Without loss of generality, we assume the robot places above the anchor, $a^R_t = \uparrow$; the alternate case is symmetric. We then fix the resulting partial state $s_t =$ as the reference state for the remainder of the paper.

From Lemma 1, with $k = 1$ (since $\leftarrow$ is the unique H0-optimum under) and $m = 2$ (since $\{\leftarrow, \uparrow\}$ are H0-optimal under ), the IOC ceiling is

$$b^{\infty}_{R1}(\text{}) = \lim_{\beta^H \to \infty} b^{+}_{R1}(\theta = \text{} \mid a^H_t = \leftarrow) = \frac{2b^{-}(\text{})}{1 + b^{-}(\text{})}.$$

When $b^{-}(\text{}) < \frac{1}{3}$, this ceiling lies below the robot's one-step decision boundary $b = \frac{1}{2}$. Importantly, no amount of assumed human rationality ($\beta^H \gg 1$) can push the one-step IOC posterior beyond $\frac{1}{2}$. This means that, if the human's true goal is $\theta^\star =$ and the robot starts off placing more than $\frac{2}{3}$ belief on , there is no course of action available to the human to prevent the robot from building the wrong structure.

Specifically, if the human places a block on top of the partial structure ($a^H_t = \uparrow$), the IOC robot's optimal response will place a block on the side ($a^R_t = \leftarrow$); and if the human places a block on the side ($a^H_t = \leftarrow$), the IOC robot will confidently follow up with a final block on top ($a^R_t = \uparrow$); either scenario results in completing the structure. In other words, even in simple assistance games, the IOC robot strategy is short-term *incorrigible* from seemingly benign initial conditions.

Multi-Goal Assistance Games We now express the inference ceiling in complete generality.

Lemma 2 (General IOC inference ceiling). *Let $a^\dagger \in \mathcal{A}^H$ be an observed human action at state s_t. For each goal $\theta \in \Theta$, R1's posterior after observing $a^\dagger$ is bounded above by*

$$b^{\infty}_{R1}(\theta \mid a^\dagger) := \lim_{\beta^H \to \infty} b^{+}_{R1}(\theta \mid a^\dagger) = \frac{\dfrac{b^-(\theta)}{|\mathcal{A}^{H0\star}(s_t,\theta)|}\,\mathbb{1}\big[a^\dagger \in \mathcal{A}^{H0\star}(s_t,\theta)\big]}{\displaystyle\sum_{\tilde\theta\in\Theta} \frac{b^-(\tilde\theta)}{|\mathcal{A}^{H0\star}(s_t,\tilde\theta)|}\,\mathbb{1}\big[a^\dagger \in \mathcal{A}^{H0\star}(s_t,\tilde\theta)\big]}. \quad (2)$$

In particular, $b^{\infty}_{R1}(\theta \mid a^\dagger) < 1$ whenever $a^\dagger$ is H0-optimal under θ and at least one other goal with positive prior probability.

Proof. As $\beta^H \to \infty$, the H0 softmax policy converges to the uniform distribution over the optimal action set:

$$\pi^{H0}(a^\dagger \mid s_t, \theta) \;\to\; \frac{1}{|\mathcal{A}^{H0\star}(s_t,\theta)|}\,\mathbb{1}\big[a^\dagger \in \mathcal{A}^{H0\star}(s_t,\theta)\big].$$

Substituting into the R1 Bayes update and normalizing yields (2). □

5.2 Pedagogic Leverage

We next show that pragmatic–pedagogic reasoning can break the belief ceiling experienced by the IOC robot R1. Since H2 evaluates actions by anticipating R1's response, actions that are H0-equivalent for a competing hypothesis need not remain equivalent under H2's reasoning.

Definition 1 (Pedagogic advantage). *For any two human actions $a^\dagger, \tilde a \in \mathcal{A}^H$, the pedagogic advantage of $a^\dagger$ against $\tilde a$ under θ from conditions (s_t, b_t) is the difference in their expected value in the presence of an IOC robot:*

$$A_\theta(s_t, b_t; a^\dagger, \tilde a) := Q^{H2}(s_t, a^\dagger; \theta, b_t) - Q^{H2}(s_t, \tilde a; \theta, b_t). \quad (3)$$

The pedagogic advantage measures how strongly H2 prefers $a^\dagger$ relative to $\tilde a$ under θ. We emphasize that, in the Boltzmann-rational setting, Q^{H2} is a function of both players' rationality parameters: β^H influences R1's belief update after observing a^H, and β^R determines its subsequent response probabilities. This is in contrast with Q^{H0}, which depends on neither β^H nor β^R.

We now establish a quantitative measure of how strongly an action relates to a single goal.

Definition 2 (Pedagogic leverage). *The pedagogic leverage of an action $a^\dagger$ toward a goal θ is equal to the product of its smallest pedagogic advantage under θ and its greatest disadvantage under the strongest competing hypothesis $\tilde\theta$:*

$$L(a^\dagger, \theta; s_t, b_t) := \left[\min_{\tilde a \neq a^\dagger} A_\theta(s_t, b_t, a^\dagger, \tilde a)\right]_+ \cdot \left[-\max_{\tilde\theta \neq \theta} \min_{\tilde a \neq a^\dagger} A_{\tilde\theta}(s_t, b_t, a^\dagger, \tilde a)\right]_+, \quad (4)$$

where $[x]_+ := \max\{0, x\}$.

We now show that whenever an observed human action has positive pedagogic leverage toward a particular goal (Definition 2), R3's posterior concentrates on that goal in the limit $\beta^H \to \infty$.

Proposition 1 (Pedagogic leverage enables full one-step alignment). *Fix any $\beta^R > 0$, and suppose that $L(a_t^H, \theta^\star; s_t, b_t) > 0$. Then,*

$$b_{R3}^{\infty}(\theta^\star) := \lim_{\beta^H \to \infty} b_{R3}^{+}(\theta^\star \mid a_t^H) = 1.$$

Proof. Positive pedagogic leverage implies that a_t^H strictly dominates every alternative action under $\theta^\star$, so $\pi^{H2}(a_t^H \mid s_t, b_t, \theta^\star) \to 1$ as $\beta^H \to \infty$. It also implies that, for every $\tilde{\theta} \neq \theta^\star$, there exists some $\tilde{a} \neq a_t^H$ with $A_{\tilde{\theta}}(s_t, b_t; a_t^H, \tilde{a}) < 0$, so a_t^H is not H2-optimal under $\tilde{\theta}$. Hence, $\pi^{H2}(a_t^H \mid s_t, b_t, \tilde{\theta}) \to 0$ as $\beta^H \to \infty$. Substituting these limiting likelihoods into R3's Bayes update yields $\lim_{\beta^H \to \infty} b_{R3}^{+}(\theta^\star \mid a_t^H) = 1$. □

Remark 2. Leverage is defined in terms of a single action $a^\dagger$ that strictly dominates every alternative under $\theta^\star$. If instead a *set* of actions $\mathcal{A}_{\theta^\star}^\dagger$ forms an H2-equivalent optimal action class, Proposition 1 holds for every $a_t^H \in \mathcal{A}_{\theta^\star}^\dagger$, provided that each action is H2-suboptimal under every competing goal. In that case, as $\beta^H \to \infty$, $\pi^{H2}(a_t^H \mid s_t, b_t, \theta^\star) \to 1/|\mathcal{A}_{\theta^\star}^\dagger| > 0$, while $\pi^{H2}(a_t^H \mid s_t, b_t, \tilde{\theta}) \to 0$ for all $\tilde{\theta} \neq \theta^\star$, so R3's posterior again concentrates on $\theta^\star$.

Corollary 1 (Pedagogic leverage breaks the IOC ceiling). *If Lemma 1 holds for $\theta^\star, \theta'$ and action $a^\dagger$ has strictly positive pedagogic leverage toward $\theta^\star$ from (s_t, b_t), then*

$$\lim_{\beta^H \to \infty} b_{R1}^{+}(\theta^\star \mid a^\dagger) = \frac{m\, b^-(\theta^\star)}{m b^-(\theta^\star) + k b^-(\theta')} \quad \textit{while} \quad \lim_{\beta^H \to \infty} b_{R3}^{+}(\theta^\star \mid a^\dagger) = 1.$$

5.3 Action-Separable Assistance Games and Corrigibility

We have shown that positive pedagogic leverage yields full one-step alignment (Proposition 1). We now analyze the specific structural conditions under which leverage will emerge, and we prove that in all such conditions, it in fact enables the computation of the full assistance game equilibrium through a simple procedure (Theorem 1). We first build intuition with the two-goal case.

Two Candidate Goals In assistance games, the human's ability to achieve desired outcomes is mediated at least in part by the robot's follow-up to their actions. We say that the human is *empowered* by the robot to the extent that the robot's response to the human's actions increases their causal influence on the state's evolution [27, 28]. It is precisely the notion of empowerment that allows for ceiling-breaking.

Consider the two-goal setting in which $|\mathcal{A}^{H0\star}(s_t, \theta^\star)| = k$ and $|\mathcal{A}^{H0\star}(s_t, \theta')| = m$, where $1 \leq k \leq m$, leading to the IOC inference ceiling (Lemma 1). Furthermore, the oracle robot policy (with perfect knowledge of the human's goal) is

$$\pi_\theta^{R\star}(a^R \mid s_t, a^\dagger) \propto \mathbb{1}\left[a^R \in \arg\max_{a' \in \mathcal{A}^R} Q^{\text{team}}(s_t, a^\dagger, a'; \theta)\right].$$

The pragmatic (R3) robot will *break the IOC ceiling* if there exists a cue $a^\dagger \in \mathcal{A}^{H0\star}(s_t, \theta^\star)$ such that

$$\mathop{\mathbb{E}}_{a^R \sim \pi_{\theta^\star}^{R\star}} Q^{\text{team}}\big(s_t, a^\dagger, a^R; \theta^\star\big) > \mathop{\mathbb{E}}_{a^R \sim \pi_{\theta'}^{R\star}} Q^{\text{team}}\big(s_t, a^\dagger, a^R; \theta^\star\big). \tag{5}$$

This inequality captures human empowerment toward goal $\theta^\star$ in the sense that the robot is able to better empower the human when the robot believes $\theta^\star$ over θ': the team strictly benefits from the robot's $\theta^\star$-aligned response over its θ'-aligned response after observing $a^\dagger$, so the human's action meaningfully shapes the state's evolution. Note that (5) concerns team Q-values, not actions. Even if $\pi_{\theta^\star}^{R\star}(\cdot \mid s_t, a^\dagger) \neq \pi_{\theta'}^{R\star}(\cdot \mid s_t, a^\dagger)$, the two responses could yield the same value under $\theta^\star$, in which case equality holds in (5) and observing $a^\dagger$ carries no information for one-step goal resolution.

Running example: Return to the two-goal Tetris example with $\theta \in \{$  ,  $\}$ and $a_t^H, a_t^R \in \{\leftarrow, \uparrow\}$. We once again analyze the robot's inference from reference state $s_t =$, after the first two blocks have been placed. We use an additive team reward: each correct placement gives $+1$, and each incorrect placement gives -1, so the joint reward lies in $\{-2, 0, +2\}$. In the one-step setting ($\gamma = 0$), $Q(s_t, a_t^H, a_t^R; \theta) = R^\theta(a_t^H, a_t^R)$, with

$$R^{\text{  }} = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}, \qquad R^{\text{  }} = \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix},$$

where rows denote $a_t^H \in \{\leftarrow, \uparrow\}$ and columns denote $a_t^R \in \{\leftarrow, \uparrow\}$.

Let the observed human action be $a_t^H =\leftarrow$. In this example, (5) holds for $a^\dagger =\leftarrow$ and $\theta^\star =$  : if the robot responds optimally for , the team receives value 2, whereas if it responds optimally for  , the team receives value 0. This strict value gap makes the cue pedagogically salient. Direct evaluation of Q^{H2} using $R^{\text{  }}, R^{\text{  }}$ and the R1 softmax response indeed gives

$$A_{\text{  }}(s_t, b_t; \leftarrow, \uparrow) > 0 \quad \text{and} \quad A_{\text{  }}(s_t, b_t; \leftarrow, \uparrow) < 0 \quad \forall \beta^H, \beta^R,$$

so $\leftarrow$ acquires leverage for (Definition 2). Hence, by Corollary 1,

$$b_{R3}^+(\text{} \mid a_t^H =\leftarrow) \to 1 \qquad \text{as } \beta^H \to \infty$$

from any prior $b^-(\text{}) \in (0, 1)$.

In particular, even in the prior regime $b^-(\text{  }) < \frac{1}{3}$, the pragmatic (R3) robot's optimal response will place a block on the side ($a_t^R =\leftarrow$) as $\beta^H \to \infty$, completing

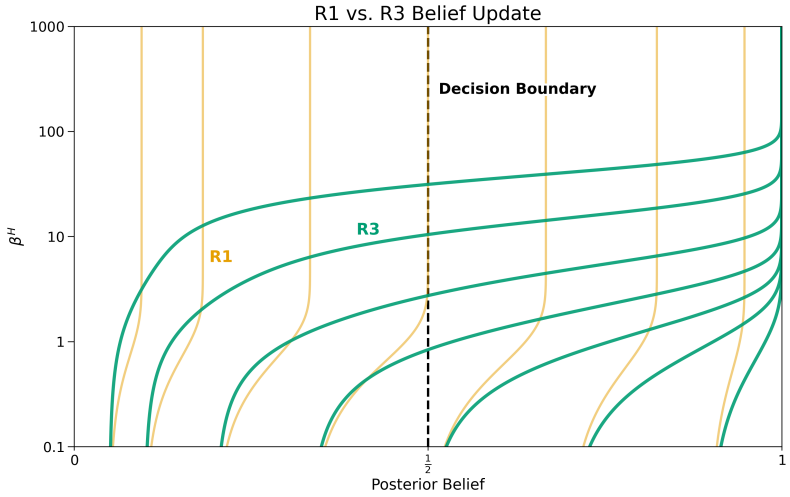


Fig. 1: Posterior belief in 🏗 after observing $a_t^H = \leftarrow$ as a function of human rationality β^H for representative priors $b^-(🏗) \in (0,1)$. The IOC update (R1, orange) saturates at the posterior ceiling of Lemma 1, while the pragmatic update (R3, green; $\beta^R = 1$) can cross the decision boundary $b = \frac{1}{2}$ (black dashed line) and approach certainty in 🏗 as β^H increases. This illustrates one-step corrigibility under pragmatic–pedagogic reasoning.

🏗. Therefore, R3 is one-step *corrigible* from any initial conditions. The emergent convention of $a_t^H = \leftarrow$ signaling 🏗 (and, by symmetry, $a_t^H = \uparrow$ signaling 🏗) can be seen as a Schelling focal point.[1] See Figure 1.

Multi-Goal Assistance Games The two-goal condition (5) generalizes to any $|\Theta| \geq 2$: for cue $a_\theta^\dagger$ to acquire leverage toward θ, the team must strictly benefit from the θ-aligned response over every competitor-aligned response at $a_\theta^\dagger$. That is, for every $\tilde{\theta} \neq \theta$ with $a_\theta^\dagger \in \mathcal{A}^{H0\star}(s_t, \theta)$,

$$\mathop{\mathbb{E}}_{a^R \sim \pi_\theta^{R\star}} Q^{\text{team}}\big(s_t, a_\theta^\dagger, a^R; \theta\big) > \mathop{\mathbb{E}}_{a^R \sim \pi_{\tilde{\theta}}^{R\star}} Q^{\text{team}}\big(s_t, a_\theta^\dagger, a^R; \theta\big). \tag{6}$$

Mirroring the two-goal case, inequality (6) is the structural condition under which pedagogic leverage emerges for a given goal $\theta^\star$. We can collect the leveraging actions for the different goals into a leverage map.

[1] Coordination problems with multiple equilibria often admit *Schelling focal points* [29, 30], or prominent solutions that players gravitate toward based on shared intuition or salience—for example, choosing "12:00 PM" as a default meeting time. Here, the convention emerges from pragmatic–pedagogic reasoning.

Definition 3 (Leverage map). *Given an assistance game $\mathcal{G}$, the leverage map at (s_t, b_t) is a set-valued map $\mathcal{L} : \Theta \rightrightarrows \mathcal{A}^H(s_t)$ that assigns to each goal θ the set of human actions $a^\dagger$ with positive pedagogic leverage toward θ.*

By Definition 2, the sets $\{\mathcal{L}(\theta)\}_{\theta\in\Theta}$ are mutually exclusive: an action (or set of actions—cf. Remark 2) with positive leverage toward θ is H2-suboptimal under every other goal.

Definition 4 (Action-separable assistance game). *An assistance game $\mathcal{G}$ is action-separable at (s_t, b_t) if $\mathcal{L}(\theta) \neq \emptyset$ for every $\theta \in \Theta$.*

In an action-separable assistance game, a single observed human action uniquely indicates a particular goal. Algorithm 1 constructs an injective assignment by choosing one leveraging action from $\mathcal{L}(\theta)$ for each goal θ. Furthermore, we prove that for this class, the full equilibrium of the assistance game can be reached in a single iteration of pragmatic–pedagogic reasoning (Theorem 1).

5.4 Vanishing Pedagogic Leverage and Scaling Behavior

We are primarily interested in the limiting behavior of rational agents, which we note is obtained by having the agents reason with noisily rational models of each other. In particular, pedagogic leverage is fundamentally a finite-temperature phenomenon: it emerges precisely because R1 is modeled as *noisily* rational (finite β^R), affording H2 a Q^{H2} gradient that the high-β^H softmax can amplify. As long as $\beta^H \to \infty$, even a vanishing amount of leverage is exploited with probability one by H2. Nonetheless, it is relevant to show why leverage vanishes and characterize the resulting β^H, β^R scaling laws in our running example.

Recall that pedagogic leverage is built from advantages A_θ (Definition 2); we thus study their scaling behavior. The H2 softmax policy and the definition of pedagogic advantage together yield the log-odds identity

$$\log \frac{\pi^{H2}(a^\dagger \mid s_t, b_t, \theta)}{\pi^{H2}(\tilde{a} \mid s_t, b_t, \theta)} = \beta^H \, A_\theta(s_t, b_t; a^\dagger, \tilde{a})$$

for any two actions $a^\dagger, \tilde{a} \in \mathcal{A}^H$ and goal θ.

Importantly, A_θ depends on β^R through R1's softmax response model, which is used to define Q^{H2}. If A_θ vanishes in some regime of β^R, then driving π^{H2} close to 1 (for $A_\theta > 0$) or 0 (for $A_\theta < 0$) requires $\beta^H |A_\theta| \to \infty$, i.e., β^H must scale at the reciprocal rate.

Running example: We now derive scaling laws in the two-goal Tetris example.

Exponential regime ($b^-(\blacksquare) < \frac{1}{3}$, $\beta^R \to \infty$). In the ceiling-saturation regime, $b^+_{R1} \approx b^\infty_{R1} = 2b^-(\blacksquare)/(1+b^-(\blacksquare))$ (5.1), keeping $2-4b^+_{R1}$ bounded below by $\kappa(b^-(\blacksquare)) := 2(1-3b^-(\blacksquare))/(1+b^-(\blacksquare)) > 0$. Hence, R1's softmax response to $a^H_t = \leftarrow$ decays as $\pi^{R1}(\leftarrow \mid s_t, \leftarrow, b^\infty_{R1}) \approx e^{-\beta^R \kappa(b^-(\blacksquare))}$ as $\beta^R \to \infty$, while

$\pi^{R1}(\uparrow|\, s_t, \uparrow, \cdot) = (1 + e^{2\beta^R})^{-1}$ is belief-independent. All other R1 responses give zero reward under and drop out of Q^{H2}. Substituting π^{R1} yields

$$A_{\blacksquare\blacksquare}(s_t, b_t; \leftarrow, \uparrow) = \Theta\big(e^{-\beta^R \kappa(b^-(\blacksquare\blacksquare))}\big), \quad -A_{\blacksquare\blacksquare}(s_t, b_t; \leftarrow, \uparrow) = \Theta\big(e^{-\beta^R \kappa(b^-(\blacksquare\blacksquare))}\big).$$

Since $\mathcal{A}^H = \{\leftarrow, \uparrow\}$, obtaining a sharp sign-flip cue requires $\beta^H A_{\blacksquare\blacksquare} \to \infty$ and $\beta^H(-A_{\blacksquare\blacksquare}) \to \infty$. Because *both* vanish like $e^{-\beta^R \kappa}$, a sufficient scaling condition is

$$\frac{1}{\beta^H} = o\Big(e^{-\beta^R \kappa(b^-(\blacksquare\blacksquare))}\Big), \qquad (\beta^R \to \infty;\ b^-(\blacksquare\blacksquare) < \tfrac{1}{3}),$$

i.e., β^H must grow faster than $e^{\beta^R \kappa(b^-(\blacksquare\blacksquare))}$ as β^R increases.

Benign regime ($b^-(\blacksquare\blacksquare) \geq \frac{1}{3}$, $\beta^R \to \infty$). The IOC ceiling lies at or above the decision boundary $b = \frac{1}{2}$, so increasing β^R *amplifies* the helpful response and no exponential scaling of β^H is required.

Hyperbolic regime ($\beta^R \to 0$). At $\beta^R = 0$, R1's policy is uniform. A first-order expansion of R1's softmax yields $A_{\blacksquare\blacksquare}(s_t, b_t; \leftarrow, \uparrow) = 2 + O(\beta^R)$ and $-A_{\blacksquare\blacksquare}(s_t, b_t; \leftarrow, \uparrow) = \Theta(\beta^R)$. Thus, as $\beta^R \to 0$, $\pi^{H2}(\leftarrow|\, \blacksquare\blacksquare)$ can be driven close to 1 without any blow-up, but suppressing $\pi^{H2}(\leftarrow|\, \blacksquare\blacksquare)$ requires $\beta^H(-A_{\blacksquare\blacksquare}) \to \infty$. It suffices that

$$\frac{1}{\beta^H} = o(\beta^R), \qquad (\beta^R \to 0),$$

or, equivalently, β^H must grow faster than $1/\beta^R$ as β^R decreases to 0.

Together, these regimes explain the exponential and hyperbolic β^H scaling observed empirically in Figure 2.

5.5 One-Round Pragmatic–Pedagogic Equilibrium

We now prove that one round of pragmatic–pedagogic reasoning suffices to reach the fixed-point, infinite-horizon equilibrium of an action-separable assistance game (Definition 4).

Theorem 1 (Pragmatic–pedagogic equilibrium in one best response). *Fix a state s_t and prior b_t over Θ. Suppose the assistance game is action-separable at (s_t, b_t) (Definition 4). Then, the limiting rational ($\beta^H, \beta^R \to \infty$) R3–H2 policies ($\pi^{R3\star}, \pi^{H2\star}$) constitute a pragmatic–pedagogic equilibrium of the assistance game at (s_t, b_t).*

Proof. It suffices to show that further pragmatic–pedagogic iterations will return the same strategies already found by H2 and R3. We begin with H4, who evaluates each a^H by the expected team value under the R3 response:

$$Q^{H4}(s_t, a^H; \theta, b_t) := \mathop{\mathbb{E}}_{a^R \sim \pi^{R3}(\cdot|s_t, a^H, b_t)} \big[Q^{\text{team}}(s_t, a^H, a^R; \theta)\big].$$

Let $\theta^\dagger$ be H4's true goal. By action-separability (Definition 4), $\mathcal{L}(\theta^\dagger) \neq \emptyset$; let $a^\dagger \in \mathcal{L}(\theta^\dagger)$ be a leveraging action for $\theta^\dagger$. By Proposition 1, $a^\dagger$ concentrates R3's

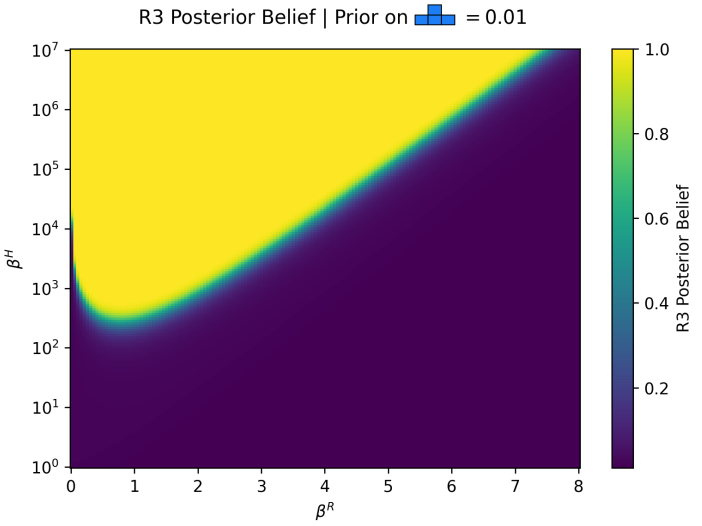


Fig. 2: One-step pragmatic posterior $b^{+}_{R3}(\blacksquare \mid s_t = \square, a^H_t = \leftarrow)$ as a function of β^R, β^H with prior $b^{-}(\blacksquare) = 0.01$. The phase boundary from low to near-certain belief is consistent with the derived scaling regimes: hyperbolic growth as $\beta^R \to 0$ and exponential growth as $\beta^R \to \infty$.

posterior belief around $\theta^\dagger$ (with arbitrary certainty for sufficiently large β^H). Since $a^\dagger$ has leverage for $\theta^\dagger$, there must be at least one robot response (produced by R1 with nonzero probability) that yields optimal expected value under $\theta^\dagger$ and strictly suboptimal under all alternative $\tilde{\theta}$; given that R3 is arbitrarily confident in $\theta^\dagger$, it chooses such an optimal response $a^{R3\star}(a^\dagger)$ *with probability 1* (or, as modeled by H4, with an arbitrarily high probability). Since this is true for each possible "true" goal, the R3-response-induced outcomes yield strict *pedagogic advantages* for H4:

$$Q^{H4}(s_t, a^\dagger; \theta^\dagger, b_t) > Q^{H4}(s_t, \tilde{a}; \theta^\dagger, b_t), \quad Q^{H4}(s_t, a^\dagger; \tilde{\theta}, b_t) < Q^{H4}(s_t, \tilde{a}; \tilde{\theta}, b_t).$$

This is equivalent to $a^\dagger$ providing *pedagogic leverage* for $\theta^\dagger$ under H4 as it did under H2 (typically stronger, due to R3's higher concentration around $a^{R3\star}$). Analogously to the H2–R3 case, the strict leverage implies that the likelihood ratio $\pi^{H4}(a^\dagger \mid \theta^\dagger)/\pi^{H4}(a^\dagger \mid \tilde{\theta})$ grows without bound (for $\tilde{\theta} \neq \theta^\dagger$) as $\beta^H \to \infty$, driving $b^{+}_{R5}(\theta^\dagger \mid a^\dagger) \to 1$, mirroring Proposition 1 for R5–H4. Hence:

$$\pi^{H4\star}(\cdot \mid s_t, b_t; \theta) \equiv \pi^{H2\star}(\cdot \mid s_t, b_t; \theta) \equiv \mathbb{1}[L(\cdot, \theta; s_t, b_t)],$$
$$\pi^{R5\star}(\cdot \mid s_t, b_t, a^\dagger) \equiv \pi^{R3\star}(\cdot \mid s_t, b_t, a^\dagger) \equiv \text{Uniform}\big[\arg\max_{a^R} Q^{\text{team}}\big(s_t, a^\dagger, a^R; \theta^\dagger\big)\big]. \quad \square$$

Algorithm 1: One-Round PPBR

Input: state s_t, prior b_t
Output: action-separable (Boolean), injective selection $\mathcal{L}' : \Theta \to \mathcal{A}^H(s_t)$ from the leverage map $\mathcal{L}$ (Definition 3)

Compute the limiting solipsistic human policy π_∞^{H0} at (s_t, b_t);
foreach *candidate human action* a^H **do**
  Compute the limiting literal posterior $b_\infty^{R1}(\cdot \mid a^H)$;
  Compute the corresponding robot response policy $\pi^{R1}(\cdot \mid a^H)$;
Compute the induced pedagogic human values Q^{H2};
foreach *goal* θ *in any chosen traversal order over* Θ **do**
  Select a candidate leveraging action $a_\theta^\dagger \leftarrow \arg\max_{a \in \mathcal{A}^H(s_t)} Q^{H2}(s_t, a; \theta, b_t)$;
  if $a_\theta^\dagger$ *has already been assigned to a prior goal* **then**
    return (False, $\emptyset$);
  else
    $\mathcal{L}'(\theta) \leftarrow a_\theta^\dagger$;
return (True, $\mathcal{L}'$);

6 Algorithm

Given a finite $\beta^R > 0$, Pragmatic–Pedagogic Best Response (PPBR, Algorithm 1) computes the full equilibrium of an action-separable assistance game (Definition 4) by returning an injective selection $\mathcal{L}' : \Theta \to \mathcal{A}^H(s_t)$ from the leverage map $\mathcal{L}$ (Definition 3), i.e., $\mathcal{L}'(\theta) \in \mathcal{L}(\theta)$ for every $\theta \in \Theta$. Define the *leveraged action set* $\mathcal{A}^\dagger := \{\mathcal{L}'(\theta) : \theta \in \Theta\}$. Since the sets $\{\mathcal{L}(\theta)\}_{\theta \in \Theta}$ are mutually exclusive, $\mathcal{L}'$ is injective and admits a well-defined inverse $(\mathcal{L}')^{-1} : \mathcal{A}^\dagger \to \Theta$. The limiting H2 policy, R3 belief, and R3 response are then readily obtained as

$$\begin{aligned} \pi_\infty^{H2}(a^H \mid s_t, b_t; \theta) &= \mathbb{1}\big[a^H = \mathcal{L}'(\theta)\big], \\ b_\infty^{R3}(\theta \mid a^H) &= \mathbb{1}\big[\theta = (\mathcal{L}')^{-1}(a^H)\big], \\ \pi_\infty^{R3}(a^R \mid s_t, b_t, a^H) &= \pi^{R\star}\big(a^R \mid s_t, a^H; (\mathcal{L}')^{-1}(a^H)\big). \end{aligned}$$

Together, these limiting strategies achieve the full optimal value of the assistance game, $V^{\mathcal{G}}(s_t, b_t; \theta)$. While Algorithm 1 directly computes the limiting behavior as $\beta^H, \beta^R \to \infty$, setting $\beta^R = 1$ and plugging in finite values of β^H yields the curves depicted in Figure 3.

7 Limitations and Future Work

A natural extension is *split leverage*, in which a single action is H2-optimal for a strict subset of goals (ruling out the rest without uniquely identifying one). Characterizing the corresponding class of assistance games and lifting Algorithm 1 to a multi-step procedure is left to future work. In addition, PPBR evaluates

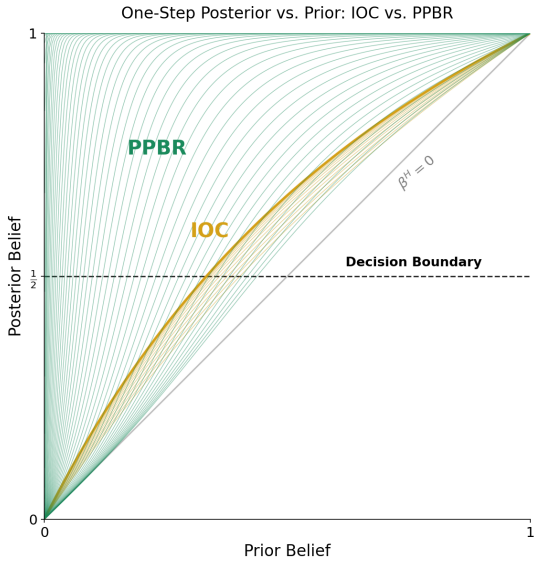


Fig. 3: One-step posterior $b^+(\blacksquare \mid a_t^H = \leftarrow)$ vs. prior $b^-(\blacksquare)$ at $s_t = \blacksquare$ with $\beta^R = 1$. IOC (R1, orange) saturates at the ceiling $2b^-/(1+b^-)$ as β^H increases. PPBR (R3, green) instead breaks the ceiling and approaches $b_\infty^{R3} = 1$.

nested softmax likelihoods and belief updates, which becomes increasingly expensive as the goal and action spaces grow. Leveraging problem structure (e.g., action abstraction) or learned amortization may preserve online tractability.

8 Conclusion

In this paper, we identified a nontrivial class of *action-separable* assistance games in which pragmatic–pedagogic reasoning resolves goal uncertainty in one time step. We formalized the IOC inference ceiling that leads to short-term incorrigibility from certain priors, and showed that positive pedagogic leverage breaks this ceiling, enabling one-step corrigibility and empowerment. We then proved that a single round of pragmatic–pedagogic reasoning reaches the fixed-point, infinite-horizon equilibrium of action-separable assistance games. We proposed Pragmatic–Pedagogic Best Response (PPBR), a practical algorithm that tests for action-separability and returns the corresponding pedagogic action assignment. Finally, in a collaborative Tetris building task, we empirically validated our results, demonstrating that PPBR outperforms IOC in the assistance setting.

Acknowledgments. We extend a special thank you to Donggeon Oh and Tom Silver for insightful discussions and feedback.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] P. F. Christiano, J. Leike, T. Brown, et al. "Deep Reinforcement Learning from Human Preferences". *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, et al. Vol. 30. Curran Associates, Inc., 2017.

[2] S. Casper, X. Davies, C. Shi, et al. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback". *Transactions on Machine Learning Research* (2023).

[3] L. Lang, D. Foote, S. Russell, et al. "When Your AIs Deceive You: Challenges of Partial Observability in Reinforcement Learning from Human Feedback". *Advances in Neural Information Processing Systems*. Ed. by A. Globerson, L. Mackey, D. Belgrave, et al. Vol. 37. Curran Associates, Inc., 2024, pp. 93240–93299.

[4] M. Williams, M. Carroll, A. Narang, et al. "On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback". *International Conference on Learning Representations (ICLR)*. 2025. eprint: `Poster`.

[5] A. Bestick, R. Bajcsy, and A. D. Dragan. "Implicitly Assisting Humans to Choose Good Grasps in Robot to Human Handovers". *International Symposium on Experimental Robotics*. Vol. 1. Springer Proceedings in Advanced Robotics. Cham: Springer, 2017, pp. 341–354.

[6] C. Liu, J. B. Hamrick, J. F. Fisac, et al. "Goal Inference Improves Objective and Perceived Performance in Human-Robot Collaboration". *International Conference on Autonomous Agents & Multiagent Systems*. AAMAS '16. International Foundation for Autonomous Agents and Multiagent Systems, 2016, pp. 940–948.

[7] A. Bobu, A. Bajcsy, J. F. Fisac, and A. D. Dragan. "Learning under Misspecified Objective Spaces". *The 2nd Conference on Robot Learning*. Ed. by A. Billard, A. Dragan, J. Peters, and J. Morimoto. Vol. 87. Proceedings of Machine Learning Research. PMLR, 2018, pp. 796–805.

[8] A. Bobu, A. Peng, P. Agrawal, et al. "Aligning Human and Robot Representations". *ACM/IEEE International Conference on Human-Robot Interaction*. HRI '24. Association for Computing Machinery, 2024, pp. 42–54.

[9] A. Fern, S. Natarajan, K. Judah, and P. Tadepalli. "A Decision-Theoretic Model of Assistance". *Journal of Artificial Intelligence Research* 50 (2014), pp. 71–104.

[10] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. "Cooperative Inverse Reinforcement Learning". *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, et al. Vol. 29. Curran Associates, Inc., 2016.

[11] J. F. Fisac, M. A. Gates, J. B. Hamrick, et al. "Pragmatic–Pedagogic Value Alignment". *International Symposium on Robotics Research (ISRR 2017)*. Vol. 10. Springer International Publishing, 2017, pp. 49–57.

[12] D. Malik, M. Palaniappan, J. Fisac, et al. "An Efficient, Generalized Bellman Update for Cooperative Inverse Reinforcement Learning". *International Conference on Machine Learning*. PMLR, 2018, pp. 3394–3402.

[13] P. Shafto, N. D. Goodman, and T. L. Griffiths. "A Rational Account of Pedagogical Reasoning: Teaching by, and Learning from, Examples". *Cognitive Psychology* 71 (2014), pp. 55–89.

[14] C. Laidlaw, E. Bronstein, T. Guo, et al. "AssistanceZero: Scalably Solving Assistance Games". *International Conference on Machine Learning*. Ed. by A. Singh, M. Fazel, D. Hsu, et al. Vol. 267. Proceedings of Machine Learning Research. PMLR, 2025, pp. 32278–32305.

[15] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling. "Learning Policies for Partially Observable Environments: Scaling Up". *Machine Learning Proceedings 1995*. Ed. by A. Prieditis and S. Russell. San Francisco (CA): Morgan Kaufmann, 1995, pp. 362–370.

[16] N. Wiener. "Some Moral and Technical Consequences of Automation". *Science* 131.3410 (1960), pp. 1355–1358.

[17] A. Y. Ng and S. J. Russell. "Algorithms for Inverse Reinforcement Learning". *Seventeenth International Conference on Machine Learning*. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 663–670.

[18] D. Ramachandran and E. Amir. "Bayesian Inverse Reinforcement Learning". *International Joint Conference on Artifical Intelligence*. IJCAI'07. Morgan Kaufmann Publishers Inc., 2007, pp. 2586–2591.

[19] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. "Maximum Entropy Inverse Reinforcement Learning". *AAAI Conference on Artificial Intelligence*. AAAI, 2008, p. 6.

[20] D. O. Stahl and P. W. Wilson. "On Players Models of Other Players: Theory and Experimental Evidence". *Games and Economic Behavior* 10.1 (1995), pp. 218–254.

[21] M. A. Costa-Gomes, V. P. Crawford, and B. Broseta. "Cognition and Behavior in Normal-Form Games: An Experimental Study". *Econometrica* 69.5 (2001), pp. 1193–1235.

[22] C. F. Camerer, T.-H. Ho, and J.-K. Chong. "A Cognitive Hierarchy Model of Games". *The Quarterly Journal of Economics* 119.3 (2004), pp. 861–898.

[23] S. Milli and A. D. Dragan. "Literal or Pedagogic Human? Analyzing Human Model Misspecification in Objective Learning". *The 35th Uncertainty in Artificial Intelligence Conference*. Ed. by R. P. Adams and V. Gogate. Vol. 115. Proceedings of Machine Learning Research. PMLR, 2020, pp. 925–934.

[24] S. Milli, D. Hadfield-Menell, A. Dragan, and S. Russell. "Should Robots be Obedient?" *Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 2017, pp. 4754–4760.

[25] D. Hadfield-Menell, A. D. Dragan, P. Abbeel, and S. Russell. "The Off-Switch Game". *Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*. 2017, pp. 220–227.

[26] A. Bobu, D. R. R. Scobee, J. F. Fisac, et al. "LESS is More: Rethinking Probabilistic Models of Human Behavior". *ACM/IEEE International Conference on Human-Robot Interaction*. HRI '20. Association for Computing Machinery, 2020, pp. 429–437.

[27] A. S. Klyubin, D. Polani, and C. L. Nehaniv. "Empowerment: A Universal Agent-Centric Measure of Control". *IEEE Congress on Evolutionary Computation (CEC)*. Vol. 1. 2005, pp. 128–135.

[28] V. Myers, E. Ellis, S. Levine, et al. "Learning to Assist Humans without Inferring Rewards". *Advances in Neural Information Processing Systems*. Ed. by A. Globerson, L. Mackey, D. Belgrave, et al. Vol. 37. Curran Associates, Inc., 2024, pp. 71540–71567.

[29] T. C. Schelling. "Bargaining, Communication, and Limited War". *Conflict Resolution* 1.1 (1957), pp. 19–36.

[30] T. C. Schelling. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press, 1980.