Pragmatic-Pedagogic Value Alignment

Jaime F. Fisac, Monica A. Gates, Jessica B. Hamrick, Chang Liu, Dylan Hadfield-Menell, Malayandi Palaniappan, Dhruv Malik, S. Shankar Sastry, Thomas L. Griffiths, and Anca D. Dragan

Abstract For an autonomous system to provide value (e.g., to customers, designers, or society at large) it must have a reliable method to determine the intended goal. This is the essence of the value-alignment problem: ensuring that the objectives of an autonomous system match those of its human users. In robotics, value alignment is crucial to the design of collaborative robots that can integrate into human workflows, successfully learning and adapting to the objectives of their users as they go. We argue that a meaningful solution to the value-alignment problem will combine multi-agent decision theory with rich mathematical models of human cognition, enabling robots to tap into people’s natural collaborative capabilities. We present a solution to the cooperative inverse reinforcement learning (CIRL) dynamic game using well-established models of decision making and theory of mind from cognitive science. The solution accounts for two crucial aspects of collaborative value alignment: that the human will not plan her actions in isolation, but will reason pedagogically about how the robot might learn from them; and that the robot should anticipate this and interpret the human’s actions pragmatically. To our knowledge, this constitutes the first equilibrium analysis of value alignment grounded in an empirically validated cognitive model of the human.

Key words: Value Alignment, Human-Robot Interaction, Dynamic Game Theory
1 Introduction

The accelerating progress in artificial intelligence (AI) and robotics is bound to have a substantial impact in society, simultaneously unlocking new potential in augmenting and transcending human capabilities while also posing significant challenges to safe and effective human-robot interaction. In the short term, integrating robotic systems into human-dominated environments will require them to assess the intentions and preferences of their users in order to assist them effectively, while avoiding failures due to poor coordination. In the long term, ensuring that advanced and highly autonomous AI systems will be beneficial to individuals and society will hinge on their ability to correctly assimilate human values and objectives [1]. We envision the short- and long-term challenges as being inherently coupled, and predict that improving the ability of robots to understand and coordinate with their human users will inform solutions to the general AI value-alignment problem.

Achieving value alignment between humans and robots requires moving from typical single-agent AI formulations to robots that acknowledge the existence of a second agent—the human—who determines what the objective is. Thus, value alignment is fundamentally a multi-agent problem. One approach that formalizes this notion is Cooperative Inverse Reinforcement Learning (CIRL), which formulates value alignment as a two-player game between a human and a robot [2]. Both agents have the same reward function, which depends on both their actions, but only the human has knowledge of the parameters of this reward.

Solving a CIRL game requires multi-agent decision theory, but also the acknowledgement that we are not dealing with any multi-agent system: we have a human-robot system. This poses a unique challenge in that humans do not behave like idealized rational agents [3]. However, humans do excel at social interaction and are extremely perceptive of the mental states of others [4, 5]. As a result, they will naturally project mental states such as beliefs and intentions onto their robotic collaborators, thereby becoming invaluable allies in our robots’ quest for value alignment.

In the coming decades, tackling the value-alignment problem will be crucial to building collaborative robots that know what their human users want. In this paper, we show that value alignment is possible not just in theory, but also in practice. We introduce a solution for CIRL based on a model of the human agent that is grounded in cognitive science findings regarding human decision making [6] and pedagogical reasoning [7]. Our solution leverages two closely related insights to facilitate value alignment. First, to the extent that improving their collaborator’s understanding of their goals may be conducive to success, people will tend to behave pedagogically, deliberately choosing their actions to be informative about these goals. Second, the robot should anticipate this pedagogical reasoning in interpreting the actions of its human users, akin to how a pragmatic listener interprets a speaker’s utterance in natural language. Jointly, pedagogical actions and pragmatic interpretations enable stronger and faster inferences among people [7]. Our result suggests that it is possible for robots to partake in this naturally-emerging equilibrium, ultimately becoming more perceptive and competent collaborators.
2 Solving Value Alignment using Cognitive Models

2.1 Cooperative Inverse Reinforcement Learning (CIRL)

Cooperative Inverse Reinforcement Learning [2] formalizes value alignment as a two-player game, which we briefly present here. Consider two agents, a human $H$ and a robot $R$, engaged in a dynamic collaborative task involving a (possibly infinite) sequence of steps. The goal of both agents is to achieve the best possible outcome according to some objective $\theta \in \Theta$. However, this objective is only known to $H$. In order to contribute to the objective, $R$ will need to make inferences about $\theta$ from the actions of $H$ (an Inverse Reinforcement Learning (IRL) problem), and $H$ will have an incentive to behave informatively so that $R$ becomes more helpful, hence the term cooperative IRL.

Formally, a CIRL game is a dynamic (Markov) game of two players ($H$ and $R$), described by a tuple $\langle S, \{A_H, A_R\}, T, \{\Theta, r\}, P_0, \gamma \rangle$, where $S$ is the set of possible states of the world; $A_H, A_R$ are the sets of actions available to $H$ and $R$ respectively; $T : S \times S \times A_H \times A_R \rightarrow [0, 1]$ a discrete transition measure\(^1\) over the next state, conditioned on the previous state and the actions of $H$ and $R$: $T(s', s, a_H, a_R); \Theta$ is the set of possible objectives; $r : S \times A_H \times A_R \times \Theta \rightarrow \mathbb{R}$ is a cumulative reward function assigning a real value to every tuple of state and actions for a given objective: $r(s, a_H, a_R; \theta); P_0 : S \times \Theta \rightarrow [0, 1]$ is a probability measure on the initial state and the objective; $\gamma \in [0, 1]$ is a geometric time discount factor.

2.2 Pragmatic Robots for Pedagogic Humans

Asymmetric information structures in games (even static ones) generally induce an infinite hierarchy of beliefs: our robot will need to maintain a Bayesian belief over the human’s objectives to decide on its actions. To reason about the robot’s decisions, the human would in principle need to maintain a belief on the robot’s belief, which will in turn inform her decisions, thereby requiring the robot to maintain a belief on the human’s belief about its own belief, and so on [8]. In [2], it was shown that an optimal pair of strategies can be found for any CIRL game by solving a partially observed Markov decision process (POMDP). This avoids this bottomless recursion as long as both agents are rational and can coordinate perfectly before the start of the game.

Unfortunately, rationality and prior coordination are nontrivial assumptions when considering human agents. Finding an equivalent tractability result for more realistic human models is therefore crucial in utilizing the CIRL formulation to solve real-world value-alignment problems involving people. We discover the key piece to the puzzle in recent cognitive studies of human pedagogical reasoning [7], in which

\(^1\) Note that the theoretical formulation is easily extended to arbitrary measurable sets; we limit our analysis to finite state and objective sets for computational tractability and clarity of exposition.
a teacher chooses actions or utterances to influence the beliefs of a learner who is aware of the teacher’s intention. The learner can interpret utterances pragmatically, a fact that the teacher can in turn exploit. The infinite recursion is averted by finding a fixed-point relation between the teacher’s best utterance and the learner’s best interpretation, exploiting a common modeling assumption in Bayesian theory of mind: the learner models the teacher as a noisily rational decision maker [9], who will be likelier to choose utterances that cause the learner to place a higher posterior belief on the correct hypothesis, given the learner’s current belief. While in reality, the teacher cannot exactly compute the learner’s belief, the model supposes that she estimates it (from the learner’s previous responses to her own utterances), and then introduces noise in her decisions to capture estimation inaccuracies. This modeling framework has been shown to predict complex behaviors observed in human teaching-learning interactions, in which pedagogical utterances and pragmatic interpretations often permit extremely efficient communication [7].

We adopt an analogous modeling framework to that in [7] for value alignment, with a critical difference: the ultimate objective of the human is not to explicitly improve the robot’s understanding of the true objective, but to optimize the team’s expected performance towards this objective. Pedagogic behavior thus emerges implicitly to the extent that a well-informed robot becomes a better collaborator in many situations.

2.3 Pragmatic-Pedagogic Fixed-Point Solution to CIRL

The robot does not have access to the true objective, $\theta$, but rather estimates a belief $b_R$ over $\theta$. We assume that this belief on $\theta$ can be expressed parametrically (this is always true if $\Theta$ is a finite set), and define $\triangle_\Theta$ to be the corresponding (finite-dimensional) parameter space, denoting $R$’s belief by $b_R \in \triangle_\Theta$. Let $Q : S \times \triangle_\Theta \times A_H \times A_R \times \Theta \to \mathbb{R}$ represent the state-action value function of the CIRL game for a given objective $\theta$, which we are seeking to compute: if $\theta \in \Theta$ is the true objective known to $H$, then $Q(s, b_R, a_H, a_R; \theta)$ represents the best performance the team can expect to achieve if $H$ chooses $a_H$ and $R$ chooses $a_R$ from state $s$, with $R$’s current belief being $b_R$. Note that in reality, the human does not know $b_R$, but for our purposes here we assume the human can compute it, following [7].

In order to solve for $Q$ under a noisily rational pedagogical human, we need to establish an appropriate dynamic programming relation for the game given a well-defined information structure. Since it is typically possible for people to predict the robot’s action if they see its beginning [10], we model the game’s information structure as feedback Stackelberg [11], in which $H$ can observe $a_R$ at each turn before committing to $a_H$. $H$ then follows a noisily rational policy for choosing $a_H$. Here, we choose the well-established Boltzmann (or soft-max) model of noisy rationality [6, 9] with the sought $Q$ as the utility metric:

$$
\pi_H^\beta(a_H|s, b_R, a_R; \theta) \propto \exp \left( \beta Q(s, b_R, a_H, a_R; \theta) \right),
$$

(1)
where \( \beta > 0 \) is termed the \textit{rationality coefficient} of \( H \). The above expression gives the likelihood of action \( a_H \) given a particular \( \theta \). The evolution of \( R \)’s belief \( b_R \) is then given (deterministically) by the Bayesian update

\[
 b'_R(\theta | s, b_R, a_R, a_H) = \pi^\theta_H(a_H | s, b_R, a_R; \theta) b_R(\theta)
\]  

(2)

Jointly, (1) and (2) define a fixed-point equation analogous to the one in [7], which states how \( R \) should pragmatically update \( b_R \) based on a noisily rational pedagogic \( a_H \). This amounts to a deterministic transition function for \( R \)’s belief, \( b'_R = f(b(s, b_R, a_H, a_R)) \). Crucially, this relation is dependent on \( Q \) itself, which we have yet to compute.

Unlike \( H, R \) is modeled as a rational agent; however, not knowing the true \( \theta \), the best \( R \) can do is to maximize \(^2\) the expectation of \( Q \) based on its current belief \(^3\) \( b_R \):

\[
\pi^\theta_R(s, b_R) := \arg\max_{a_R} \sum_{a_H, \theta} Q(s, b_R, a_H, a_R; \theta) \cdot \pi^\theta_H(a_H | s, b_R, a_R; \theta) b_R(\theta)
\]  

(3)

Combining (2) with the state transition measure \( T(s' | s, a_H, a_R) \), we can define the Bellman equation for \( H \) under the noisily rational policy \( \pi^\theta_H \) for any given \( \theta \in \Theta \):

\[
Q(s, b_R, a_H, a_R; \theta) = r(s, a_H, a_R; \theta) + \mathbb{E}_{s', a'_R} \left[ \gamma \cdot Q(s', b'_R, a'_H, \pi^\theta_R(s', b'_R); \theta) \right]
\]  

(4)

where \( s' \sim T(s' | s, a_H, a_R) \); \( b'_R = f(s, b_R, a_H, a_R) \); \( a'_R \sim \pi^\theta_R(a_H | s', b'_R, \pi^\theta_R(s', b'_R); \theta) \). Note that \( H \)’s next action \( a'_H \) implicitly depends on \( R \)’s action at the next turn.

Substituting (1-3) into (4), we obtain the sough dynamic programming relation for the CIRL problem under a noisily rational-pedagogic human and a pragmatic robot. The human is pedagogic because she takes actions according to (1), which takes into account how her actions will influence the robot’s belief about the objective. The robot is pragmatic because it assumes the human is actively aware of how her actions convey the objective, and interprets them accordingly.

The above Bellman relation for \( Q(s, b_R, a_H, a_R; \theta) \), coupled in \( \theta \), can be solved (for \( \gamma < 1 \)) via value iteration. Similar to [2], the resulting problem is a POMDP. It is important to point out that, although they are computationally simpler than the general multi-agent planning, POMDPs are still PSPACE-complete [12], so reducing pragmatic-pedagogic equilibrium computation to solving a POMDP falls short of rendering the problem tractable in general. However, it does allow algorithms to leverage and benefit from the extensive research on practical algorithms for approximate planning in large POMDPs [13].

\(^2\) We assume for simplicity that the optimum is unique or a well-defined disambiguation rule exists.

\(^3\) Note that this does not imply \textit{certainty equivalence}, nor do we assume separation of estimation and control: \( R \) is fully reasoning about how its actions and those of \( H \) may affect its future beliefs.
Fig. 1 Simple collaborative scenario with two possible objectives. The human $H$ wants soup but the robot $R$ is initially confident that her goal is to make salad. Even under a full POMDP formulation, if $R$ reasons “literally” about $H$’s actions using standard IRL (assuming that $H$ behaves as if $R$ knows the true objective), it fails to infer the correct objective. Conversely, under the pragmatic-pedagogic CIRL equilibrium, $R$ views $H$ as having an incentive to pick actions pedagogically to correct $R$’s belief when needed; under the pragmatic interpretation, $H$’s wait action in the second turn (instead of adding spinach, which would be strongly preferred by a pedagogic $H$ wanting salad) becomes a strong indicator that $H$ wants soup. Even though $H$’s actions are the same under both solutions, only the pragmatic $R$ achieves value alignment and completes the recipe.

3 A Proof-of-Concept

We consider a household collaboration setting in which a human $H$ seeks to prepare a meal with the help of an intelligent robotic manipulator $R$. There are multiple possible meals that $H$ may want to prepare using the available ingredients, and $R$ does not know beforehand which one she has chosen (we assume $H$ cannot or will not tell $R$ explicitly). If $H$ is aware of $R$’s uncertainty, she should take actions that give $R$ actionable information, particularly the information that she expects will allow $R$ to be as helpful as possible as the task progresses.

Concretely, our problem has three food types: spinach, bread, and tomatoes. Each food can be in two or three states: spinach can be raw or chopped; tomatoes can be raw, chopped, or liquefied; and bread can be raw, sliced, or toasted. The different recipes correspond to different (joint) target states for the food. For example, soup requires liquefied tomatoes, toasted bread, and no spinach. $H$ and $R$ can slice or chop any of the foods, while only $R$ can puree tomatoes or toast bread. If $H$ and $R$ work on the same food they incur a high cost (e.g., injury) and the game ends.

A simple scenario with two recipes is solved for infinite horizon using discretized belief-state value iteration and presented as an illustrative example in Fig 1. $R$ initially has the wrong belief about $H$’s intended recipe. If $R$ is pragmatic and $H$ is pedagogic, then $H$ is able to substantively change $R$’s belief and together they successfully collaborate on making the human’s desired meal. If $R$ interprets $H$’s actions literally, then $H$ is unable to effectively communicate her desired recipe. We also computed the solution to games with 4 recipes through a modification of POMDP value iteration. In the pedagogic-pragmatic equilibrium, $H$ and $R$ successfully cook the correct recipe 86% of the time. However, if $R$ is literal and $H$ is oblivious (as in standard IRL), they only successfully cook $H$’s desired meal 15% of the time.
References

[1] Dario Amodei, Jacob Steinhardt, Dan Man, and Paul Christiano. “Concrete Problems in AI Safety”. arXiv preprint (2017), pp. 1–29.

[2] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. “Cooperative Inverse Reinforcement Learning”. Advances in Neural Information Processing Systems (NIPS) (2016).

[3] Amos Tversky and Daniel Kahneman. “Judgment under Uncertainty: Heuristics and Biases”. Science 185.4157 (1974), pp. 1124–1131.

[4] Fritz Heider and Marianne Simmel. “An Experimental Study of Apparent Behavior”. The American Journal of Psychology 57.2 (1944), pp. 243–259.

[5] Andrew N. Meltzoff. “Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children.” Developmental Psychology 31.5 (1995), pp. 838–850.

[6] Chris L. Baker and Joshua B. Tenenbaum. “Modeling Human Plan Recognition Using Bayesian Theory of Mind”. Plan, Activity, and Intent Recognition. Ed. by Gita Sukthankar, Christopher Geib, Hung Hai Bui, et al. 2014, pp. 177–204.

[7] Patrick Shafto, Noah D. Goodman, and Thomas L. Griffiths. “A rational account of pedagogical reasoning: Teaching by, and learning from, examples”. Cognitive Psychology 71 (2014), pp. 55–89.

[8] Shmuel Zamir. “Bayesian games: Games with incomplete information”. Computational Complexity: Theory, Techniques, and Applications 9781461418 (2012), pp. 238–253.

[9] R. Ducan Luce. Individual Choice Behavior: a Theoretical Analysis. John Wiley and sons, 1959.

[10] Anca. D. Dragan and Siddhartha Srinivasa. “Integrating human observer inferences into robot motion planning”. Autonomous Robots (2014).

[11] Tamer Basar and Alain Haurie. “Feedback Equilibria in Differential Games with Structural and Modal Uncertainties”. Advances in Large Scale Systems 1 (1984), pp. 163–201.

[12] Martin Mundhenk, Judy Goldsmith, Christopher Lusena, and Eric Allender. “Complexity of Finite-horizon Markov Decision Process Problems”. Journal of the ACM (JACM) 47.4 (2000), pp. 681–720.

[13] David Silver and Joel Veness. “Monte-Carlo Planning in Large POMDPs”. Advances in neural information processing systems. 2010, pp. 2164–2172.