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 As intelligent systems gain autonomy and capability, it becomes vital to ensure that their objectives match those of their human users; this is known as the value-alignment problem. In robotics, value alignment is key to the design of collaborative robots that can integrate into human workflows, successfully inferring and adapting to their users' objectives as they go. We argue that a meaningful solution to value alignment must 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 based on well-established cognitive models of decision making and theory of mind. The solution captures a key reciprocity relation: the human will not plan her actions in isolation, but rather reason pedagogically about how the robot might learn from them; the robot, in turn, can anticipate this and interpret the human's actions pragmatically. To our knowledge, this work constitutes the first formal analysis of value alignment grounded in empirically validated cognitive models.

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

All authors are with the University of California, Berkeley.
e-mail: {jfisac,mgates,jhamrick,changliu,dhm,malayandi,dhruvmalik,
shankar_sastry,tom_griffiths,anca}@berkeley.edu

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*.

Successful value alignment requires moving from typical single-agent AI formulations to robots that account for a second agent—the human—who determines what the objective is. In other words, value alignment is fundamentally a multi-agent problem. Cooperative Inverse Reinforcement Learning (CIRL) formulates value alignment as a two-player game in which a human and a robot share a common reward function, but *only the human* has knowledge of this reward [2]. In practice, solving a CIRL game requires more than multi-agent decision theory: we are not dealing with *any* multi-agent system, but with 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]. They will naturally project mental states such as beliefs and intentions onto their robotic collaborators, 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 (CIRL) [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 θ 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; $T : S \times S \times A_H \times A_R \to [0,1]$ a discrete transition measure[1] over the next state, conditioned on the previous state and the actions of H and R: $T(s'|s, a_H, a_R)$; Θ is the set of possible objectives; $r : S \times A_H \times A_R \times \Theta \to \mathbb{R}$ is a cumulative reward function assigning a real value to every tuple of state and actions for a given objective: $r(s, a_H, a_R; \theta)$; $P_0 : S \times \Theta \to [0,1]$ is a probability measure on the initial state and the objective; $\gamma \in [0,1]$ is a geometric time discount factor making future rewards gradually less valuable.

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, when dealing with human agents, rationality and prior coordination are nontrivial assumptions. Finding an equivalent tractability result for more realistic human models is therefore crucial in using the CIRL formulation to solve real-world value-alignment problems. We discover the key insight in cognitive studies of human *pedagogical reasoning* [7], in which a teacher chooses actions or utterances to influence the beliefs of a learner who is aware of the teacher's intention. The teacher can then exploit the fact that the learner can interpret utterances pragmatically. 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

[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.

causing the learner to place a high 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 utterances), then introduces noise in her decisions to capture estimation inaccuracies. This framework can predict complex behaviors observed in human teaching-learning interactions, in which pedagogical utterances and pragmatic interpretations permit 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.

2.3 Pragmatic-Pedagogic Equilibrium Solution to CIRL

The robot does not have access to the true objective θ, but rather estimates a belief b_R over θ. We assume that this belief on θ can be expressed parametrically (this is always true if Θ is a finite set), and define $\triangle_\Theta$ to be the corresponding (finite-dimensional) parameter space, denoting R's belief by $b_R \in \triangle_\Theta$. While in reality the human cannot directly observe b_R, we assume, as in [7], that she can compute it or infer it from the robot's behavior (and model estimation inaccuracies as noise in her policy). We can then 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 θ, 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.

In order to solve for Q, we seek to establish an appropriate dynamic programming relation for the game, given a well-defined information structure and a model of the human's decision making. Since it is typically possible for people to predict a robot's next action if they see its beginning [10], we assume that H can observe a_R at each turn before committing to a_H. A well-established model of human decision making in psychology and econometrics is the Luce choice rule, which models people's decisions probabilistically, making high-utility choices more likely than those with lower utility [9]. In particular, we employ a common case of the Luce choice rule, the Boltzmann (or *soft-max*) noisy rationality model [6], in which the probability of a choice decays exponentially as its utility decreases in comparison to competing options. The relevant utility metric in our case is the sought Q (which captures H's best expected outcome for each of her available actions a_H). Therefore the probability that H will choose action a_H has the form

$$\pi_H^{\odot}(a_H | s, b_R, a_R; \theta) \propto \exp\big(\beta Q(s, b_R, a_H, a_R; \theta)\big) \ , \tag{1}$$

where $\beta > 0$ is termed the *rationality coefficient* of H and quantifies the concentration of H's choices around the optimum; as $\beta \to \infty$, H becomes a perfect rational agent, while, as $\beta \to 0$, H becomes indifferent to Q. The above expression can be interpreted by R as the *likelihood* of action a_H given a particular θ. The evolution of R's belief b_R is then given (deterministically) by the Bayesian update

$$b'_R(\theta|s,b_R,a_R,a_H) \propto \pi_H^{\odot}(a_H|s,b_R,a_R;\theta)b_R(\theta) \ , \tag{2}$$

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, however, the fixed-point relation derived here involves Q itself, which we have yet to compute.

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

$$\pi_R^*(s,b_R) := \arg\max_{a_R} \sum_{a_H,\theta} Q(s,b_R,a_H,a_R;\theta)\cdot\pi_H^{\odot}(a_H|s,b_R,a_R;\theta)b_R(\theta) \ . \tag{3}$$

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 $\pi_H^{\odot}$ for any given $\theta \in \Theta$:

$$Q(s,b_R,a_H,a_R;\theta) = r(s,a_H,a_R;\theta) + \mathbb{E}_{s',a'_H}\left[\gamma\cdot Q'\left(s',b'_R,a'_H,\pi_R^*(s',b'_R);\theta\right)\right] \ , \tag{4}$$

where $s' \sim T(s'|s,a_H,a_R)$; $b'_R = f_b(s,b_R,a_H,a_R)$; $a'_H \sim \pi_H^{\odot}(a_H|s',b'_R,\pi_R^*(s',b'_R);\theta)$. 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 sought 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 resulting problem is similar to a POMDP (in this case formulated in belief-state MDP form), with the important difference that the belief transition depends on the value function itself. In spite of this complication, the problem can be solved in backward time through dynamic programming: each Bellman update will be based on a pragmatic-pedagogic fixed point that encodes an equilibrium between the Q function (and therefore H's policy for choosing her action) and the belief transition (that is, R's rule for interpreting H's actions). Evidence in [7] suggests that people are proficient at finding such equilibria, even though uniqueness is not guaranteed in general; study of disambiguation is an open research direction.

[2] We assume for simplicity that the optimum is unique or a well-defined disambiguation rule exists.

[3] Note that this does not imply *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.

3 A Proof-of-Concept

We introduce the benchmark domain ChefWorld, 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). The team obtains a reward only if H's intended recipe is successfully cooked. 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.

Our problem has 3 ingredients, each with 2 or 3 states: spinach (absent, chopped), tomatoes (absent, chopped, puréed), and bread (absent, sliced, toasted). Recipes correspond to (joint) target states for the food. Soup requires the tomatoes to be chopped then puréed, the bread to be sliced then toasted, and no spinach. Salad requires the spinach and tomatoes to be chopped, and the bread to be sliced then toasted. H and R can slice or chop any of the foods, while only R can purée tomatoes or toast bread.

A simple scenario with the above two recipes is solved using discretized belief-state value iteration and presented as an illustrative example in Fig 1. R has a wrong initial belief about H's intended recipe. Under standard IRL, H fails to communicate her recipe. But if R is pragmatic and H is pedagogic, H is able to change R's belief and they successfully collaborate to make the meal.

In addition, we computed the solution to games with 4 recipes through a modification of POMDP value iteration (Table 1). In the pragmatic-pedagogic CIRL equilibrium with $\beta = 5$, H and R successfully cook the correct recipe 97% of the time, whereas under the standard IRL framework (with H acting as an expert disregarding R's inferences) they only succeed 46% of the time—less than half as often.

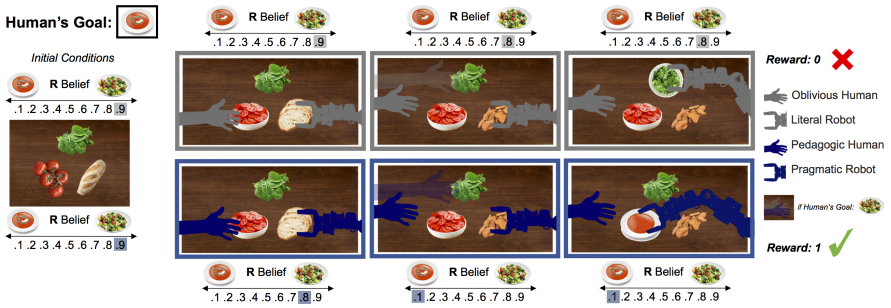

Fig. 1 Simple collaborative scenario with 2 possible objectives. The human H wants soup but the robot R initially believes her goal is salad. Even under a full POMDP formulation, if R reasons "literally" about H's actions using standard IRL (assuming H behaves as if R *knew* the true objective), it fails to infer the correct objective. Conversely, under the pragmatic-pedagogic CIRL equilibrium, R views H as incentivized to choose pedagogic actions that will fix R's belief when needed. Under the pragmatic interpretation, H's *wait* action in turn 2 (instead of adding spinach, which would be preferred by a pedagogic H wanting salad) indicates H wants soup. While H's actions are the same under both solutions, only the pragmatic R achieves value alignment and completes the recipe.

	Boltzmann ($\beta = 1$)	Boltzmann ($\beta = 2.5$)	Boltzmann ($\beta = 5$)	Rational
IRL	0.2351	0.3783	0.4555	0.7083
CIRL	0.2916	0.7026	0.9727	1.0000

Table 1 A comparison of the expected value (or equivalently here, the probability of success) achieved by CIRL and IRL on the *ChefWorld* domain with four recipes when the robot begins with a uniform belief over the set of recipes. We ran each algorithm across different models of the human's behavior, namely a rational model and a Boltzmann-rational model with various values of β (a higher β corresponds to a more rational human). When the human is highly irrational ($\beta = 1$), both CIRL and IRL unsurprisingly perform rather poorly. However, as the human becomes less noisy ($\beta = 2.5$, $\beta = 5$), CIRL outperforms IRL by a significant margin; in fact, the pragmatic-pedagogic CIRL strategy with a Boltzmann-rational human performs comparably ($\beta = 2.5$) or even substantially outperforms ($\beta = 5$) the IRL result when the human is perfectly rational.

4 Discussion

We have presented here an analysis of the AI value alignment problem that incorporates a well-established model of human decision making and theory of mind into the game-theoretic framework of cooperative inverse reinforcement learning (CIRL). Using this analysis, we derive a Bellman backup that allows solving the dynamic game through dynamic programming. At every instant, the backup rule is based on a pragmatic-pedagogic equilibrium between the robot and the human: the robot is uncertain about the objective and therefore incentivized to learn it from the human, whereas the human has an incentive to help the robot infer the objective so that it can become more helpful.

We note that this type of pragmatic-pedagogic equilibrium, recently studied in the cognitive science literature for human teaching and learning [7], may not be unique in general: there may exist two actions for H and two corresponding interpretations for R leading to different fixed points. For example, H could press a blue or a red button which R could then interpret as asking it to pick up a blue or a red object. Although we might feel that blue-blue/red-red is a more intuitive pairing, blue-red/red-blue is valid as well: that is, if H thinks that R will interpret pressing the blue button as asking for the red object then she will certainly be incentivized to press blue when she wants red; and in this case R's policy should consistently be to pick up the red object upon H's press of the blue button. When multiple conventions are possible, human beings tend to naturally disambiguate between them, converging on salient equilibria or "focal points" [11]. Accounting for this phenomenon is likely to be instrumental for developing competent human-centered robots.

On the other hand, it is important to point out that, although they are computationally simpler than more general multi-agent planning problems, POMDPs are still PSPACE-complete [12], so reducing pragmatic-pedagogic equilibrium computation to solving a modified POMDP falls short of rendering the problem tractable in general. However, finding a POMDP-like Bellman backup does open the door to efficient CIRL solution methods that leverage and benefit from the extensive research on practical algorithms for approximate planning in large POMDPs [13].

We find the results in this work promising for two reasons. First, they provide insight into how CIRL games can be not only theoretically formulated but also practically solved. Second, they demonstrate, for the first time, formal solutions to value alignment that depart from the ideal assumption of a rational human agent and instead benefit from modern studies of human cognition. We predict that developing efficient solution approaches and incorporating more realistic human models will constitute important and fruitful research directions for value alignment.

Acknowledgements This work is supported by ONR under the Embedded Humans MURI (N00014-13-1-0341), by AFOSR under Implicit Communication (16RT0676), and by the Center for Human-Compatible AI.

References

[1] D. Amodei, J. Steinhardt, D. Man, and P. Christiano. "Concrete Problems in AI Safety". *arXiv preprint* (2017).

[2] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. "Cooperative Inverse Reinforcement Learning". *NIPS* (2016).

[3] A. Tversky and D. Kahneman. "Judgment under Uncertainty: Heuristics and Biases". *Science* 185.4157 (1974).

[4] F. Heider and M. Simmel. "An Experimental Study of Apparent Behavior". *The American Journal of Psychology* 57.2 (1944).

[5] A. N. Meltzoff. "Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children." *Dev. Psych.* 31.5 (1995).

[6] C. L. Baker and J. B. Tenenbaum. "Modeling Human Plan Recognition Using Bayesian Theory of Mind". *Plan, Activity, and Intent Recognition*. 2014.

[7] P. Shafto, N. D. Goodman, and T. L. Griffiths. "A rational account of pedagogical reasoning: Teaching by, and learning from, examples". *Cog. Psych.* 71 (2014).

[8] S. Zamir. "Bayesian games: Games with incomplete information". *Computational Complexity: Theory, Techniques, and Applications* (2012).

[9] R. D. Luce. *Individual Choice Behavior: a Theoretical Analysis*. John Wiley and Sons, 1959.

[10] A. D. Dragan and S. Srinivasa. "Integrating human observer inferences into robot motion planning". *Autonomous Robots* (2014).

[11] T. C. Schelling. *The strategy of conflict*. Harvard University Press, 1960.

[12] M. Mundhenk, J. Goldsmith, C. Lusena, and E. Allender. "Complexity of Finite-horizon Markov Decision Process Problems". *J. ACM* 47.4 (2000).

[13] D. Silver and J. Veness. "Monte-Carlo Planning in Large POMDPs". *NIPS*. 2010.