Concept Alignment as a Prerequisite for Value Alignment

Author Names Omitted for Anonymous Review. Paper-ID [add your ID here]

Abstract—**Value alignment is essential for building AI systems that can safely and reliably interact with people. However, what a person values—and is even capable of valuing—depends on the concepts that they are currently using to understand and evaluate what happens in the world. The dependence of values on concepts means that *concept alignment* is a prerequisite for value alignment—agents need to align their representation of a situation with that of humans in order to successfully align their values. Here, we formally analyze the concept alignment problem in the inverse reinforcement learning setting, show how neglecting concept alignment can lead to systematic value misalignment, and describe an approach that helps minimize such failure modes by jointly reasoning about a person's concepts and values. Additionally, we report initial experimental results with human participants showing that humans reason about the concepts used by an agent when acting intentionally, in line with our joint reasoning model.**

I. Introduction

People's thoughts and actions are fundamentally shaped by the concepts they use to represent the world and formulate their goals. Imagine watching someone waiting to cross a busy intersection. Making sense of their behavior requires understanding their representation of things like "the crosswalk", "the road", "the bike lane", and "the right of way". For instance, it is important to take into account whether someone understands or is aware of the part of the street designated the "bike lane" while they wait since otherwise their intentions could be misinterpreted (e.g., a naïve observer might think someone standing in the bike lane is *trying* to get hit by a bicycle). Yet, current approaches to inferring human goals, rewards, and values (e.g., standard inverse reinforcement learning [1] and value alignment [3]) largely neglect the possibility that an observer and actor can have misaligned concepts. Our goal in this work is to formally state the problem of *concept alignment*, begin to explore algorithmic solutions, and compare these solutions to human judgments.

To formalize concept alignment, we draw on the recently proposed framework of *value-guided construal* [6], which provides a computational account of how humans form simplified representations of problems in order to solve them. A *construal* is a particular interpretation of a problem in terms of a set of concepts and related causal affordances: for example, if one understands the concept of the bike lane and includes it in their current construal, they are aware of the fact that bicycles are often on the bike lane, cars generally avoid the bike lane, you might get hit if you stand in the bike lane, etc. People often prefer simpler construals since they are less cognitively effortful [7], but this can affect the quality of one's actions—e.g., if you fail to distinguish the bike lane from the sidewalk, you might stand in a place where a bicycle will hit you! As we discuss later, our approach is to incorporate construals into a forward model of planning, which allows us to articulate the problem of conceptual misalignment as a form of misspecified *inverse* planning [2].

II. Related Work

Work on inferring human preferences and values is often done in the framework of inverse-reinforcement learning (IRL) [1, 3, 5] and inverse planning [2]. In the standard IRL setting, an agent is tasked with estimating or inferring the reward function that an expert is optimizing. An important benefit of IRL over other methods for learning from expert human behavior, such as behavioral cloning [10], is that it facilitates *generalization* to new scenarios outside of the data given. For instance, by inferring that a human has a dispreference for eating spinach after observing behavior at home, an agent could anticipate behavior in new scenarios in which spinach appears, such as in a restaurant. Over the past two decades, methods for IRL have been extended in various ways and even used as models for social cognition in cognitive science [4].

However, a key property of virtually all existing IRL methods is that they assume behavior emerges from a planning process that produces optimal or noisy-optimal policies [1, 9, 14]. This assumption is problematic because it is false [11, 12]. An alternative perspective that has been developed over the past few years is that people are *resource-rational*—that is, they think and act rationally, but are subject to cognitive limitations on time, memory, or attention [8]. A major research challenge for IRL, value alignment, and cognitive science is incorporating these ideas into estimating human preferences and values [4].

The work here builds on recent approaches to modeling resource-rational human planning in the *value-guided construal* framework, which provides an account of how humans rationally simplify problems and apply simplified concepts in order to plan [6, 7]. The key idea of value-guided construals is that people do not necessarily use all concepts available when representing a problem in order to make efficient use of limited attention (e.g., ignoring certain details of obstacles when navigating through a GridWorld). Applied to the IRL setting, this involves inverting the value-guided construal model of human decision-making and using it instead of the classical noisy-rational model. Our goal here is to provide an initial demonstration of the utility of incorporating concept simplification strategies into value alignment and IRL.

III. Model

We begin by reviewing the basic formalism for sequential decision-making before turning to construals and the inverse planning problem.

A. Background

We represent sequential decision-making tasks as Markov decision-processes (MDPs) $M = \langle \mathcal{S}, \mathcal{A}, P_0, T, R, \gamma \rangle$, where $\mathcal{S}$ is a state space; $\mathcal{A}$ is an action space; $P_0 : \mathcal{S} \to [0, 1]$ is an initial state distribution; $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a transition function; $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a real-valued reward function; and $\gamma \in [0, 1)$ is a discount rate. A (stochastic) policy is a conditional probability distribution that maps states to distributions over actions, $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. We denote the Markov chain resulting from following policy π on an MDP with dynamics T as $T^\pi(s' \mid s) = \sum_a \pi(a \mid s)T(s' \mid s, a)$.

We consider standard (unregularized) and entropy-regularized solutions to MDPs. In the unregularized setting, the value function associated with a policy π on an MDP with dynamics T and reward function R maps each state to the expected cumulative, discounted reward that results from following π: $V^\pi_{(R,T)}(s) = \sum_a \pi(a \mid s)[R(s, a) + \gamma \sum_{s'} T(s' \mid s, a)V^\pi_{(R,T)}(s')]$. The state occupancy function (also known as the successor representation) associated with a policy π on an MDP with dynamics T is the expected discounted visitations to a state s^+ starting from a state s, $\rho^\pi_T(s; s^+) = \mathbf{1}[s^+ = s] + \gamma \sum_{s'} T^\pi(s' \mid s)\rho^\pi_T(s'; s^+)$. The optimal value function for an MDP M maximizes value at each state, $V^*_{(R,T)}(s) = \max_a\{R(s, a) + \gamma \sum_{s'} T(s' \mid s, a)V^*_{(R,T)}(s')\}$.

In the entropy-regularized setting, the value of a policy π on MDP M is modified to include an entropy term, which penalizes action distributions that are more deterministic: $H(\pi(\cdot \mid s)) = -\sum_a \pi(a \mid s) \ln\{\pi(a \mid s)\}$. When this penalty is parameterized by a weight β, we denote the optimal entropy-weighted value function as $V^\beta_{(R,T)}(s) = \max_\pi\{\sum_a \pi(a)[R(s, a) + \sum_{s'} T(s' \mid s, a)V^\beta_{(R,T}(s')] + \beta H(\pi)\}$.

B. Inverse Reinforcement Learning (IRL)

The standard IRL problem formulation involves an *observer* attempting to estimate the reward function of an expert *demonstrator* based on observed behavior. This can be formalized as Bayesian inference, where given a trajectory of expert acting in the task, $\zeta = \{\langle s_0, a_0\rangle, \langle s_1, a_1\rangle, ..., \langle s_T, a_T\rangle\}$, the observer infers the demonstrator's reward function, R:

$$P(R \mid \zeta) = \frac{P(\zeta \mid R)P(R)}{P(\zeta)}. \tag{1}$$

To calculate the likelihood of a trajectory ζ given a reward function R, it is typically assumed that the observer has knowledge of the dynamics of the demonstrator's task, T. Then, the likelihood is the probability of the trajectory being generated by the optimal policy under a candidate R:

$$P(\zeta \mid R) = \prod_{\langle s_t, a_t\rangle \in \zeta} \pi^\beta_{(R,T)}(a_t \mid s_t). \tag{2}$$

C. Inverse Construal

The inverse construal problem considers the possibility that although a resource-limited demonstrator is acting in a task with a particular dynamics T, they may not be *planning their actions* with respect to the fully-detailed dynamics. Rather, the demonstrator's behavior results from planning with respect to a *construed task dynamics*, $\tilde{T}$, that is simpler or easier to solve.

Thus, an observer that takes into account the resource limitations faced by human planners should instead be aiming to solve an inference problem that incorporates the possibility of alternative task construals. Formally, this is the problem:

$$P(R, \tilde{T} \mid \zeta) = \frac{P(\zeta \mid R, \tilde{T})P(R, \tilde{T})}{P(\zeta)}, \tag{3}$$

where the likelihood is given by

$$P(\zeta \mid R, \tilde{T}) = \prod_{\langle s_t, a_t\rangle \in \zeta} \pi^\beta_{(R,\tilde{T})}(a_t \mid s_t). \tag{4}$$

D. Consequences of not considering construals

How bad can the estimate of R be when assuming the true dynamics T versus attempting to estimate the demonstrator's construal $\tilde{T}$? If we use a maximum causal entropy formulation of IRL to get an estimated policy $\hat{\pi}^{\text{InvRL}}$ and compare this to the estimated policy assuming the demonstrator is using a construal, $\hat{\pi}^{\text{InvCon}}$, then the learner's performance gap on the true task is [13]:

$$|v^{\hat{\pi}^{\text{InvCon}}}_{(R,T)} - v^{\hat{\pi}^{\text{InvRL}}}_{(R,T)}| \leq \frac{\gamma \cdot |R|^{\max}}{(1-\gamma)^2} \cdot \max_{s,a} ||T(\cdot \mid s, a) - \tilde{T}(\cdot \mid s, a)||_1$$

where $|R|^{\max} = \max_{s,a} |R(s, a)|$. In other words, if the observer has an inaccurate estimate of the transition function the actor uses to plan, they may drastically mis-estimate the reward function that motivated behavior. This provides a formal expression of our introductory example, in which failing to consider that a person does not know about or is unaware of a bike lane might lead one to interpret standing in the bike lane as indicating a *desire* to be hit by a bicycle.

IV. A Simple Example of Concept Misalignment

To investigate the impact of modeling (or not modeling) a construal on value alignment between a human subject and a machine IRL agent, we use *blocks and notches* maze tasks similar to those developed by Ho, Cohen & Griffiths [7] to study rigidity in people's construals (Figure 1). Each block and notches maze consists of a start state, a nearby goal (pink), a faraway goal (green), black walls, and blue 3x3 blocks. The blocks generally prevent movement, except for smaller notches (light blue) that permit movement through the blocks.

A. Notches

In our simulations, notches (represented by light blue squares within the 3x3 blue blocks) are shortcuts through the maze. All human subjects are shown the same view, but only some human subjects notice and learn how to use the notches; others ignore the light blue vs dark blue distinction and treat the entire 3x3 blue block as an obstacle, which they navigate around. In other words, the human subjects with different construals of the same ground truth grid learn different paths [7].

A standard IRL agent is misaligned at the concept level because it assumes an optimal policy (and therefore has no notion that a human might not understand notches or how to use them). Humans, as we have discussed before, often act in ways that are not conventionally considered optimal or even rational. The IRL agent, without an understanding of the human subjects' different construals, draws incorrect conclusions about the human subjects' values (rewards).

Of the four trajectories (Figure 1) used in our experiments, the two on the right are routes taken by humans who did not pay attention to the notches. The near (pink) goal is unreachable without using notches; think of it as a doughnut shop enclosed on all sides by traversable construction areas (3x3 dark blue blocks) through which there are shortcuts (light blue notches). The grids on the left show the trajectory of a human subject who has learned that a notch is a shortcut, and has used the notch to form a more efficient path to their preferred goal. On the right are the trajectories of a human subject who only knows the blue 3x3 blocks are obstacles, without paying attention to the fact that some sub-blocks (notches) are not obstacles at all. Looking at these trajectories on the right, the IRL agent which does not have any notion of construals and assumes an optimal policy would naturally assume the human subject has a value-related reason for avoiding the pink goal, and would thus assume that the green goal has a higher reward. Thus we see value misalignment emerge as a consequence of concept misalignment between the human subject and the IRL agent.

B. Value misalignment

In our reinforcement learning framework, we use rewards as a proxy for values. To demonstrate how concept misalignment can lead to value misalignment between humans and machines, we employ an inverse reinforcement learning agent to infer the human's values (reward function). Without knowledge of the construals (different understanding of notches), the agent might misattribute the path to a higher reward value for the chosen goal, not realizing the other goal may in fact have a higher reward, but may be impossible to reach without using/paying attention to notches.

As a measure of how alignment at the construal level can improve value alignment, we compare the posterior probability $P(reward, construal|traj)$ when jointly modeling the reward and the construal, to $P'(reward|traj)$, the standard IRL posterior which assumes that the trajectory is coming from a policy optimal with respect to the true transition function.

V. Human Experiments

To demonstrate that humans use their knowledge of construals when making inferences about others' paths, we ran a human participant study. We showed 100 participants the same four trajectories given to the two IRL agents (Figure 1) and asked them to make the same inferences.

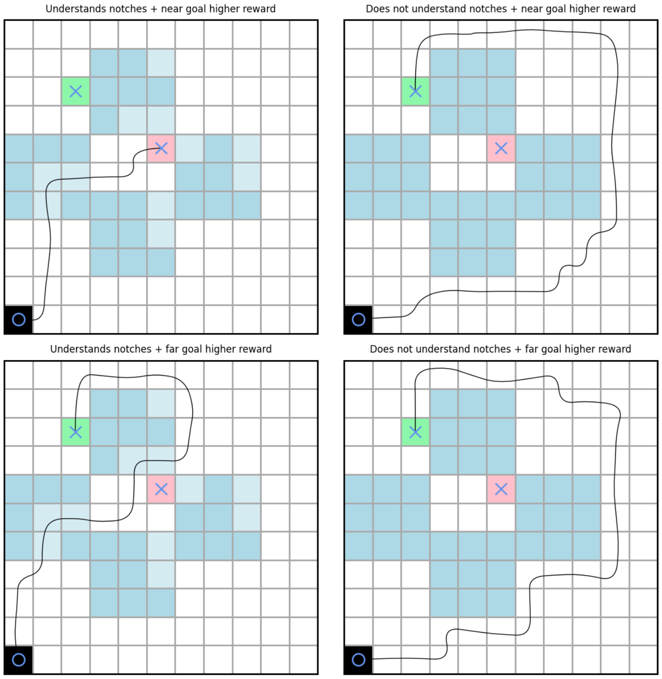

Understands notches + near goal higher reward
Does not understand notches + near goal higher reward
Understands notches + far goal higher reward
Does not understand notches + far goal higher reward

Fig. 1. Four trajectories produced by different combinations of rewards and construals. The two trajectories on the right with the construal ”Does not understand notches” look similar, because the near (pink) goal is impossible to reach when not construing notches.

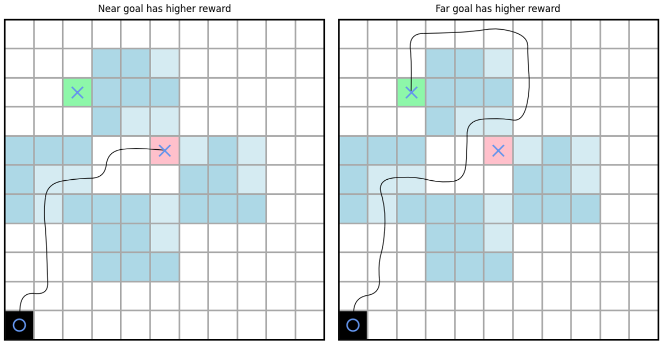
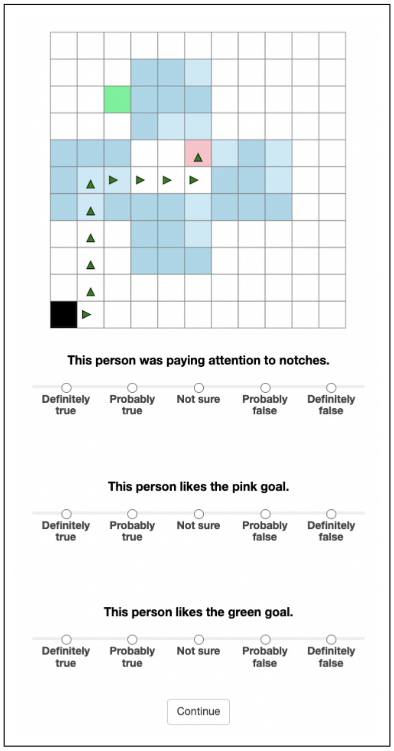


Fig. 2. Trajectories produced by modeling only rewards, without accounting for the fact that some people pay attention to notches and others don't.

Each participant was shown a live replay of each trajectory, and then asked to infer (Figure 3) whether the person who took this route:

1) Was paying attention to the notches
2) Liked the near goal
3) Liked the far goal

The latter two questions we equate to the posterior of the IRL algorithm's reward inferences about each goal. The answer choices for each question are as follows, with the number in parentheses showing the mapped numerical value in our Results section:



This person was paying attention to notches.

○ Definitely true ○ Probably true ○ Not sure ○ Probably false ○ Definitely false

This person likes the pink goal.

○ Definitely true ○ Probably true ○ Not sure ○ Probably false ○ Definitely false

This person likes the green goal.

○ Definitely true ○ Probably true ○ Not sure ○ Probably false ○ Definitely false

Fig. 3. One frame of the data collection process where we collected human judgements on the IRL task given the four trajectories in Figure 1.

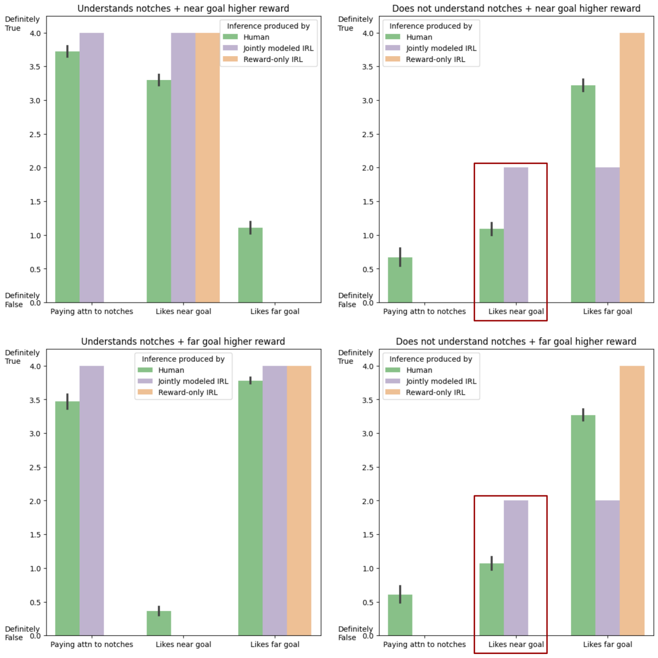


Fig. 4. Inferences produced by humans and the two models. In the two ”Does not understand notches” scenarios, it is impossible to know whether the unreachable goal has a higher reward; jointly modeling construals and rewards allows the IRL algorithm to successfully model this uncertainty. Human subjects display similar uncertainty when making this inference. The reward-only IRL agent answers (incorrectly) with full certainty.

1) Definitely True (4)
2) Probably True (3)
3) Not sure (2)
4) Probably False (1)
5) Definitely False (0)

A full walkthrough of instructions, visuals, and questions shown to the human participants is included in the supplementary materials. We also scale the IRL posterior inferences to this 0-4 scale for direct comparison with the human judgments.

VI. Results

There are three components to our results: human data, IRL inference when jointly modeling rewards and construals, and IRL inference when modeling only reward. These results are shown side-by-side in Figure 4. The posteriors of the IRL inference are scaled to match the 0-4 scale of the human data. Error bars for human data are one standard error from the mean over all 100 participants, for each question of each trajectory.

VII. Discussion

In this work, we formulate the problem of conceptual alignment within the framework of value-guided construals. When people are faced with a task, they often do not represent it in full detail and instead engage in simplification strategies to make more efficient use of limited cognitive resources [6]. As a result, people may use simplified concepts that lead to different behaviors than if they had represented the task in complete detail. Our main goal here has been to formalize the inverse problem of estimating what simplified concepts people are using and show how such an approach is needed for successful value alignment and IRL in a simple setting.

In both scenarios of the ”Does not understand notches” construal, it is impossible to know whether the unreachable goal has a higher reward (see Figure 1); jointly modeling construals and rewards allows one IRL algorithm to successfully model this uncertainty. Human subjects display similar uncertainty when making this inference. The reward-only IRL agent answers (incorrectly) with full certainty.

Modeling construals and allowing for alignment at a conceptual level enables the IRL algorithm to correctly infer uncertainty around values instead of confidently making an incorrect inference. Modeling construals also brings the IRL behavior closer to the human participants' behavior, because both recognize the uncertainty when inferring another human's values (rewards).

VIII. Future work

In future work, we intend to test this approach in a wider variety of settings with human experts as well as develop inference algorithms that can scale to larger reward and construal spaces. More broadly, we hope that demonstrating the critical importance of concept alignment to the larger goal of value alignment will open the door to future work characterizing concept and value alignment in real-world settings.

References

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[2] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113 (3):329–349, 2009.

[3] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

[4] Mark K. Ho and Thomas L. Griffiths. Cognitive science as a source of forward and inverse models of human decisions for robotics and control. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):33–53, 2022. doi: 10.1146/annurev-control-042920-015547. URL https://doi.org/10.1146/annurev-control-042920-015547.

[5] Mark K Ho, Michael Littman, James MacGlashan, Fiery Cushman, and Joseph L Austerweil. Showing versus doing: Teaching by demonstration. *Advances in neural information processing systems*, 29, 2016.

[6] Mark K Ho, David Abel, Carlos G Correa, Michael L Littman, Jonathan D Cohen, and Thomas L Griffiths. People construct simplified mental representations to plan. *Nature*, 606(7912):129–136, 2022.

[7] Mark K Ho, Jonathan D Cohen, and Tom Griffiths. Rational simplification and rigidity in human planning, Mar 2023. URL psyarxiv.com/aqxws.

[8] Falk Lieder and Thomas L Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43, 2020.

[9] Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous agents and multi-agent systems*, 30(1):30–59, 2016.

[10] Paul Munro, Hannu Toivonen, Geoffrey I. Webb, Wray Buntine, Peter Orbanz, Yee Whye Teh, Pascal Poupart, Claude Sammut, Caude Sammut, Hendrik Blockeel, Dev Rajnarayan, David Wolpert, Wulfram Gerstner, C. David Page, Sriraam Natarajan, and Geoffrey Hinton. Behavioral cloning. In *Encyclopedia of Machine Learning*, pages 93–97. Springer US, 2011. doi: 10.1007/978-0-387-30164-8_69. URL https://doi.org/10.1007/978-0-387-30164-8_69.

[11] Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99, February 1955. doi: 10.2307/1884852. URL https://doi.org/10.2307/1884852.

[12] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185 (4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124. URL https://www.science.org/doi/abs/10.1126/science.185.4157.1124.

[13] Luca Viano, Yu-Ting Huang, Parameswaran Kamalaruban, Adrian Weller, and Volkan Cevher. Robust inverse reinforcement learning under transition dynamics mismatch. *Advances in Neural Information Processing Systems*, 34: 25917–25931, 2021.

[14] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1433–1438. AAAI Press, 2008. ISBN 9781577353683.