Goal Alignment: Re-analyzing Value Alignment Problems Using Human-Aware AI [*]

Extended Abstract

Malek Mechergui
Colorado State University
Fort Collins, United States
Malek.Mechergui@colostate.edu

Sarath Sreedharan
Colorado State University
Fort Collins, United States
Sarath.Sreedharan@colostate.edu

ABSTRACT

Value alignment problems arise in scenarios where the specified objectives of an AI agent don't match the true underlying objectives of its users. While value alignment remains a popular topic within AI safety research, most existing works in this sphere tend to overlook one of the foundational causes for misalignment, namely the inherent asymmetry in human expectations about the agent's behavior and the behavior generated by the agent for the specified objective. To address this lacuna, we propose a novel formulation for the value alignment problem, named *Human-aware goal alignment* that highlights this central challenge related to value alignment. Additionally, we propose a first-of-its-kind interactive goal elicitation algorithm that is capable of using information generated under incorrect beliefs about the agent, to determine the true underlying goal of the user.

KEYWORDS

Value Alignment, Planning, Human-Aware AI

ACM Reference Format:
Malek Mechergui and Sarath Sreedharan. 2023. Goal Alignment: Re-analyzing Value Alignment Problems Using Human-Aware AI : Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Value alignment, as presented in [4], is the problem of ensuring that an AI agent's pursuit of its specified objectives will maximize or satisfy the true underlying objectives of its human user. This problem has been widely argued to be one of the most important problems related to AI safety [2, 8]. While there is a general consensus that the primary cause of the value misalignment problem is the user's failure to correctly anticipate the outcomes of their specification, current works related to value alignment tend to focus on addressing only some aspects of the overall problem. We argue that the extant literature on value alignment overlooks the fundamental problem that any information user provides to the system is going to be skewed by their beliefs about the agent model, which may be different from the agent's own model. This in turn means that the user's expectations about the behavior the agent would exhibit in response to a particular goal specification could be drastically different from what might actually be followed. As such, we argue that any complete solution to value alignment must account for such an asymmetry in expectations. In fact, for a system to correctly use any information provided by the user it must try to re-interpret it in the light of this inherent difference between the user and the agent. Thus in this paper, we will present a new formalization of the value alignment problem that accounts for this asymmetry between the user and the AI agent. We will ground this formulation in one of the most basic sequential decision-making settings, namely deterministic goal-directed planning. This setting will transform the value alignment problem to a *goal alignment problem*, where we will focus on a scenario where the user's belief could be different from the agent model. This formulation will build on and generalize a framework called Human-Aware AI [9], which was originally introduced to generate explainable behavior. Under this new formulation, we will see how agents can use their knowledge about the existing differences between the user's beliefs and the agent model to reason about human's true underlying objectives.

2 PROBLEM FORMULATION

We will focus on deterministic goal-directed planning problems represented by a tuple: $\mathcal{M} = \langle D, I, G \rangle$ [3]. Under this notation, $D = \langle F, A \rangle$ is the domain model of the planning problem, where F is a set of propositional fluents and A provides the set of actions the agent can execute. Finally, I corresponds to the start state and G captures the partial goal specification, such that any state $s \supseteq G$ is considered a valid goal state. We will use $\mathcal{T}$ to denote the transition function. A solution to a planning problem takes the form of a plan, i.e. a sequence of actions whose execution in the initial state would result in a goal state, i.e., $\pi = \langle a_1, ..., a_k \rangle$ is a plan if $\mathcal{T}(\pi, I^{\mathcal{M}}, D^{\mathcal{M}}) \supseteq G^{\mathcal{M}}$. Additionally, each action has a unit cost thus, $C(\pi) = |\pi|$. We will start by denoting the domain model and the initial state captured by the robot as $D^R = \langle F, A^R \rangle$ and I^R. The human's beliefs about the robot model, initial state, and their specified goal are respectively denoted as $\mathcal{M}^H = \langle D^H, I^H, G^H \rangle$, where $D^H = \langle F, A^H \rangle$. Here the value misalignment problem translates to the possibility that a plan that achieves the specified goal need not achieve the underlying human goal.

Definition 1. *A goal specification G^H is said to be misaligned with the human goal G^* for a robot domain model D^R and initial state I^R, if there exists an action sequence $\pi = \langle a_1, ..., a_k \rangle$ such that $\mathcal{T}(\pi, I^R, D^R) \supseteq G^H$, but $\mathcal{T}(\pi, I^R, D^R) \not\supseteq G^*$*

[*] Full paper can be found in https://arxiv.org/abs/2302.00813

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

Keeping with the existing works in value-alignment, we assume that the human can provide a plan π^H that they believe can achieve the true goal (per $\mathcal{M}^H$). However, just because π^H is executable in $\mathcal{M}^H$, there is no guarantee that the robot can execute it, or that executing it will result in the same goal state. As a starting point, we will assume that the human only cares about the final outcome of a plan, thus, only the goal state matters. Therefore, instead of following the specified plan, the robot will try to identify a plan that will satisfy the final state expected by the human. Note that this is compatible with cases where the human may have trajectory level constraints, as they can be compiled down into goal state fluents (cf. [1]).
Now let us assume that the human goal specification is a partial specification of the true goal , i.e., $G^H \subseteq G^*$. Now the central challenge for the system is to determine if it can achieve G^*, and, if so, to come up with a plan that satisfies the goal G^*. However, the fact that the human provided the robot with a plan gives us information about what G^*. For one, we can assert that G^* must be a subset of what the human believes would have resulted from executing the plan ($\mathcal{T}(\pi^H, I^H, D^H)$). The problem of course is how one identifies the exact subset. Besides, queries to directly get G^* (say by asking, *'are you sure you only need me to achieve G^H?'*) are bound to fail. In fact, the difference between G^H and G^*, is not just a result of them forgetting some fluents, but a reflection of their beliefs about the task. The robot could on the other hand ask the human whether they care about any given fluent. Thus we will introduce a function $O^{G^*} : F \rightarrow [0, 1]$ that will return 1 if a given fluent is part of G^*. This will become the central interaction mechanism through which we will solve our underlying goal misalignment problem.

Definition 2. *A **human-aware goal alignment (HAGL)** is specified by the tuple $\mathcal{H} = \langle D^R, I^R, G^H, D^H, I^H, \pi^H, O^{G^*} \rangle$, where there exists an unknown goal G^*, such that $\mathcal{T}(\pi^H, I^H, D^H) \supseteq G^*$ and $G^H \subseteq G^*$ and $\forall, f \in F, O^{G^*}(f) = 1$, if and only if $f \in G^*$. Now the goal of the robot is to find π^R such that $\mathcal{T}(\pi^R, I^R, D^R) \supseteq G^*$, if one exists, while minimizing the queries to O^{G^*}*

As with many of the human-aware planning works, we will assume access to D^H and I^H. In terms of computational complexity, we can compare HAGL against planning and see that it is at the very least as hard as solving classical planning problems, i.e., it is at least PSPACE-Hard.

3 SOLVING HAGL

To solve HAGL, we will approximate the value of information related to querying each fluent and then query the ones with the highest value. We will only use this procedure if G^H is achievable, but the robot can't achieve all the fluents that were made true by the human plan in the human model ($\mathcal{T}(\pi^H, I^H, D^H)$). We will calculate the value associated with querying about each fluent, as

$$\mathcal{V}^Q(f) = p(f \in G^*) \times V(f \in G^*) + (1 - p(f \in G^*)) \times V(f \notin G^*)$$

Where $p(f \in G^*)$ is the probability that fluent is part of the goal and $V(f \in G^*)$, respective values of knowing whether f is part of the goal or not. For simplification purposes, the achievement of each fluent can be done independently of the other. Let $S^H_{G^*} = \mathcal{T}(\pi^H, I^H, D^H)$ and let $\hat{F} \subseteq S^H_{G^*}$ be the set of fluents in the goal state that the robot cannot achieve in its model. Now to calculate the probability, we will employ a strategy similar to the ones used in goal recognition [7] and keeping with the conventions used by [7], we can formalize this as

$$p(f \in G^*) \propto e^{-1\times\beta\times|C(\pi^H)-C(\hat{\pi}^*_f)|}$$

Where $\hat{\pi}^*_f$ is the optimal plan in the human model for the goal $G^H \cup f$, β is a rationality parameter that also controls the randomness of the decision-maker given that we assume the human to be a noisy rational decision-maker[6]. The value function reflects the certainty the robot has regarding the achievability of the goal state. Now we can find a lower bound on this true value by just using the probability that the goal is unachievable.

$$V(f \in G^*) \approxeq \sum_{\bar{G}} P(G^* = \bar{G}) \times \mathbb{1}(\bar{G} \text{ not solvable})$$

Where $G^H \subseteq \bar{G} \subseteq S^H_{G^*}$ and $f \in \bar{G}$. We can similarly define $V(f \notin G^*)$, but now we will only consider subsets of goal state that don't contain f. However, if we additionally assume that if a fluent is achievable in isolation in the robot model, it can also be achieved as part of any goal state, we only need to care about the fluents that are part of $\hat{F}$ So we will define

$$\tilde{V}(f \in G^*) = \begin{cases} 1 \text{ if } f \text{ is not achievable} \\ \prod_{\hat{f} \in \hat{F}} p(\hat{f} \in G^*) \text{ Otherwise} \end{cases}$$

In the case of $\tilde{V}(f \notin G^*)$ the value is always given as $\tilde{V}(f \notin G^*) = \prod_{\hat{f} \in \hat{F} \setminus \{f\}} p(\hat{f} \in G^*)$. Now the important point of this approximation is the assumption that each fluent's independent achievability reflects its overall achievability. Obviously, there are many cases where this assumption may not hold, but we can show that the value $\tilde{V}$ is guaranteed to be an under approximation of V.

Now that we have a value associated with each fluent. We will start by querying them in the order of their value. We will end the query process under one of the three conditions: **1)** The human says yes to a fluent that cannot be achieved, **2)** The current subset of fluents the human has said yes to cannot be achieved along with the goal or **3)** There exists a plan that can achieve the current subset of fluents the human has said yes to can be achieved along with G^H and any unqueried fluent. The first two conditions correspond to cases where the robot can't achieve the expected goal and the latter where the robot can achieve a superset of G^* and thus that plan would be acceptable to the human. This procedure is guaranteed to be complete. This result follows from the fact that in the worst case, it would ask about every fluent that is part of $S^H_{G^*}$ and will be able to determine if a plan exists or not.

4 EVALUATION AND CONCLUSION

For evaluating our proposed algorithm, we ran our method on a set of problems selected from standard IPC benchmark problems [5]. The initial empirical evaluation shows that our algorithm helps reduce the number of queries to the human before the system can come up with a plan guaranteed to satisfy the true human goal. As next step, we hope to extend the framework to support more complex decision-making settings including decision-theoretic ones, or by looking at the use of more realistic decision-making models for humans while relaxing assumptions about the access to the human mental model of the robot.

ACKNOWLEDGMENT

The authors would like to thank Dr. Subbarao Kambhampati and Serena Booth for helpful discussions on the topic.

REFERENCES

[1] Jorge A. Baier, Fahiem Bacchus, and Sheila A. McIlraith. 2009. A heuristic search approach to planning with temporally extended preferences. *Artif. Intell.* 173, 5-6 (2009), 593–618.

[2] Brian Christian. 2020. *The alignment problem: Machine learning and human values.* WW Norton & Company.

[3] Hector Geffner and Blai Bonet. 2013. *A concise introduction to models and methods for automated planning.* Synthesis Lectures on Artificial Intelligence and Machine Learning, Vol. 7. Morgan & Claypool Publishers, Kentfield, CA, USA. 1–141 pages. https://doi.org/10.2200/S00513ED1V01Y201306AIM022

[4] Dylan Hadfield-Menell, Stuart Russell, Pieter Abbeel, and Anca D. Dragan. 2016. Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain.* Curran Associates, Inc., Barcelona, Spain, 3909–3917.

[5] International Planning Competition. 2011. IPC Competition Domains. https://goo.gl/i35bxc.

[6] Hong Jun Jeon, Smitha Milli, and Anca D. Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.* Curran Associates, Inc., Virtual, 4415–4426.

[7] Miquel Ramírez and Hector Geffner. 2010. Probabilistic Plan Recognition Using Off-the-Shelf Classical Planners. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010,* Maria Fox and David Poole (Eds.). AAAI Press, Atlanta, Georgia, USA.

[8] Stuart Russell. 2019. *Human compatible: Artificial intelligence and the problem of control.* Penguin.

[9] Sarath Sreedharan, Anagha Kulkarni, and Subbarao Kambhampati. 2022. Explainable Human–AI Interaction: A Planning Perspective. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 16, 1 (2022), 1–184.