THESIS

GOAL ALIGNMENT: RE-ANALYZING VALUE ALIGNMENT PROBLEMS USING HUMAN-AWARE AI

Submitted by

Malek Mechergui

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2024

Master's Committee:

Advisor: Sarath Sreedharan

Nathaniel Blanchard
Ali Pezeshki

ABSTRACT

GOAL ALIGNMENT: RE-ANALYZING VALUE ALIGNMENT PROBLEMS USING HUMAN-AWARE AI

While the question of misspecified objectives has gotten much attention in recent years, most works in this area primarily focus on the challenges related to the complexity of the objective specification mechanism, for example, the use of reward functions. However, the complexity of the objective specification mechanism is just one of many reasons why the user may have misspecified their objective. A foundational cause for misspecification that is being overlooked by the previous works is 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, we propose a novel formulation for the objective misspecification problem that builds on the human-aware planning literature, which was originally introduced to support explanation and explicable behavioral generation. Additionally, we propose a first-of-its-kind interactive algorithm that is capable of using information generated under incorrect beliefs about the agent to determine the true underlying goal of the user.

DEDICATION

I dedicate this work to my beloved parents and to the cherished memory of my aunt, who was not only my family but also my best friend.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The Value alignment problem, as presented in [Hadfield-Menell et al.(2016)], is the problem of ensuring that an AI agent's pursuit of its specified objectives will maximize or satisfy the true underlying objective of its human user. The issue is usually studied and discussed in the context or scenarios where such misalignments could have catastrophic consequences. It's important to note that the Value alignment problem has been widely argued to be one of the most important problems related to AI safety [Christian(2020), Russell(2019)].

While there is a general consensus that the primary cause of the value misalignment problem is the user's (the human) failure to correctly anticipate and precisely predict the outcomes of their specification, current works tend to focus on addressing only some aspects of the problem.

In particular, most works within the value alignment problem tend to focus on decision-theoretic settings, where the objectives are specified as reward functions and try to address the problems that are closely connected to the nature of this representation scheme (cf. [Hadfield-Menell et al.(2016), Leike et al.(2018), Hadfield-Menell et al.(2017)]).

We argue that the extant literature on value alignment overlooks the fundamental problem that,

any information the user provides to the system is going to be skewed by not only their inherent limitations in inferential capabilities but also their beliefs about the agent model, which may be different from the agent's own model.

This in turn means that the user's expectation about the behavior the agent would exhibit in response to their particular goal specification could be drastically different from what might actually be followed by the agent. Arguably, this asymmetry between the user's expectations about agent behavior and the agent's true behavior is one of the main factors that gives rise to the misalignment in the first place. As such, 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.

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 do so by first removing many of the extraneous parts of the problem that are artifacts of the setting rather than the true nature of the value misalignment problem. In fact, we will focus on 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*, which will be specifically grounded in a scenario where the user's belief could be different from the agent model.

Please note that *all problems studied and formalized in this work are equally present in more complex settings*, and we hope that our initial framework can act as the foundation for building

solutions for such special settings.

To achieve this, we will build on and generalize a framework called Human-Aware AI [Sreedharan et al.(2022a)], which was originally introduced to generate explainable behavior.

The framework uses psychological concepts of mental models [Premack and Woodruff(1978)] to model and understand human-AI interaction. Figure 1.1 shows an example of how we could build on the human-aware AI framework to understand how goal misspecification may arise.

As clearly illustrated, the human (the user) is specifying a goal to an agent to elicit a behavior they would deem desirable. However, if their beliefs about the agent model are different from the true agent model or if their reasoning process is faulty, it could lead to the human providing goals that may result in completely unexpected behaviors/outcomes. This also means that if the agent hopes to identify and try to satisfy the true objectives of the user, it must identify the existing differences between the user's beliefs and the agent model and use this difference to reason about the intended behavior.

We introduce an approach to performing such reasoning that *only uses assumptions made in either value-alignment or human-aware AI literature*.

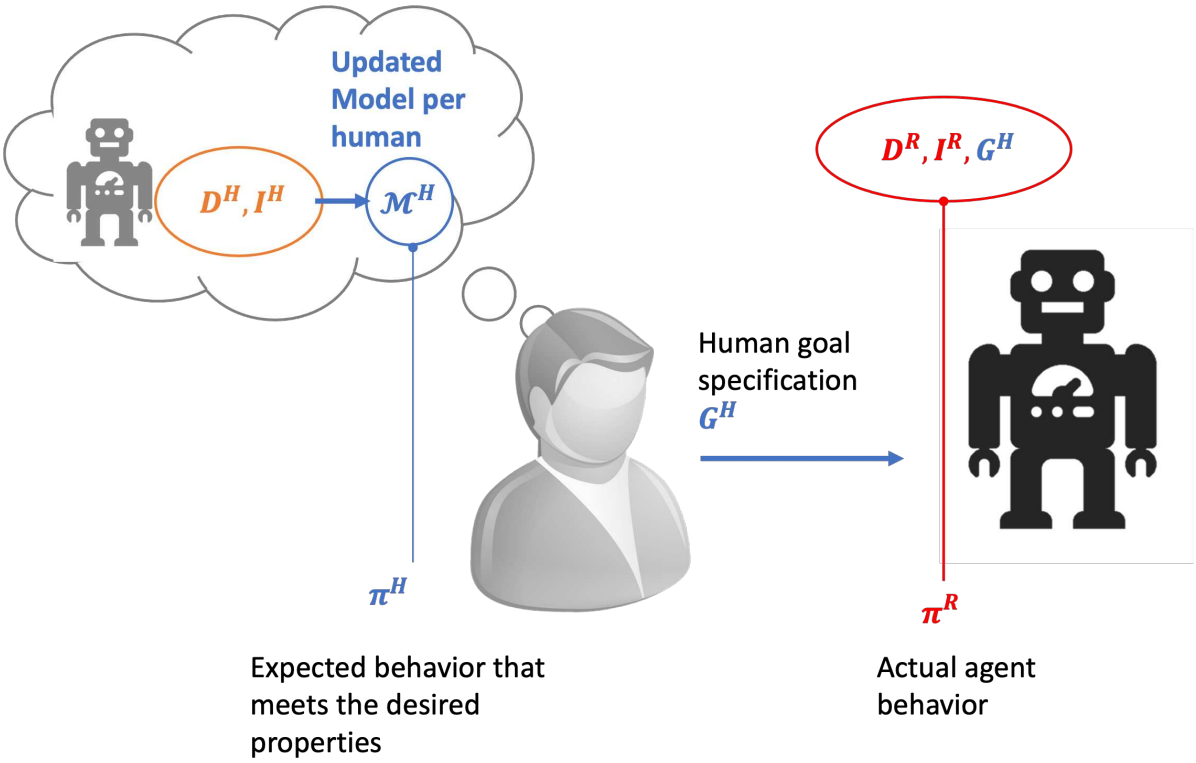

Figure 1.1: An overview of the objective specification process as contextualized in a generalized Human-aware AI framework. Humans ascribe a domain model and initial state to the agent, which may differ from the true model. Now the human identifies a goal specification whose inclusion in the agent's model they believe will result in plans they would prefer. Note that the human is generating the model updates based on a potentially incorrect understanding of the system's model and using possibly faulty reasoning. The resulting outcomes from pursuing that goal using the robot model could differ greatly from what the human expected.

In summary, the primary contributions of this paper are as follows:

- We formalize and define the problem of *Human-aware Goal Alignment*; a formulation of the Value Alignment problem that explicitly accounts for the asymmetry between the user's expectations and the agent's decisions.

- We establish the lower bound complexity of the human-aware Goal Alignment problem.

- We introduce a first-of-its-kind interactive goal elicitation algorithm that can use information generated from incorrect model beliefs.

- We provide an empirical evaluation demonstrating the computational characteristics of our algorithm.

Chapter 2

Background

2.1 PDDL: Planning Domain Definition Language

The Planning Domain Definition Language (PDDL) was originally made for the first-ever International Planning Competition (IPC) [Aeronautiques et al.(1998)]. The planning language was introduced to serve in the definition, standardization, and automation of Artificial Intelligence classical planning problems. PDDL is a formal language designed to support and extend one of the early formal languages for planning systems such as the Stanford Research Institute Problem Solver (STRIPS) [Fikes and Nilsson(1971)] and other syntactic features.

PDDL encloses a description of the world that is easily readable and especially interpretable by humans. A PDDL planning task is essentially composed of two separate files:

```
(define (problem <problem name>)
   (:domain <domain name>)
   <PDDL code for objects>
   <PDDL code for initial state>
   <PDDL code for goal specification>
)
```

Figure 2.1: Example of a Problem file structure.

- A problem file, as shown in Figure 2.1, specifies:
  - **The initial state:** Starting conditions or baselines (predicates that are true at the beginning) of the planning task (ultimately the world) from which the subject (human or agent) will perform actions.
  - **The goal state:** Final or desired conditions that must be achieved (true) after the execution of a plan (sequence of actions).
  - **The objects:** Specific entities or objects in the world that we are interested in and that belong to the defined types.

```
(define (domain <domain name>)
   <PDDL code for predicates>
   <PDDL code for first action>
   [...]
   <PDDL code for last action>
)
```

Figure 2.2: Example of a Domain file structure.

- A domain file, as shown in Figure 2.2, defines:
  - **The types:** Classes of objects that are used in the planning domain and involved in the different actions and predicates.

- **The predicates:** Logical statements (can either be True or False) that describe the state or relationship between the objects at any given time.
- **The available actions:** Possible operations that can be performed to transition from one state to another. Each action has a set of **preconditions** that need to be satisfied (true) for the action to be executed, and a set of **effects**, predicates that will become true upon executing the action.

2.2 Deterministic goal-directed planning problems

We will focus on deterministic goal-directed planning problems. Such problems can be represented and modeled using a tuple of the form $\mathcal{M} = \langle D, I, G \rangle$ [Geffner and Bonet(2013)]. Under this notation, D corresponds to the domain model of the planning problem, which is further defined by the tuple, $D = \langle F, A \rangle$, where F is a set of propositional fluents that are used to define the state space of the planning problem and A provides the set of actions that can be executed by the agent. Each state possible under the given planning problem can be uniquely identified by the set of fluents that are true in that state, thus the total number of possible states is equal to $2^{|F|}$. 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.

Now, each action $a \in A$ is further defined by the tuple, $a = \langle pre_{+}(a), add(a), del(a) \rangle$, where pre_{+}

are the preconditions that need to be satisfied to execute a, while add and del denote the add(s) and delete(s) effects related to the action.

In a planning problem, we will use $\mathcal{T}$ to capture the effects of executing an action a at a given state s, in other words, applying an action a to a state s transitions the system to another state s'.

$\mathcal{T}(a, s, D)$ is defined as:

$$\mathcal{T}(a, s, D) = \begin{cases} (s \setminus del(a)) \cup add(a), & \text{if } pre_{+}(a) \subseteq s \\ \textit{undefined}, & \text{otherwise} \end{cases}$$

Overloading the notation a little bit, we will also use $\mathcal{T}$ to capture the consequence of executing a sequence of actions $< a_1, a_2, \ldots, a_k >$, i.e.,

$$\mathcal{T}(< a_1, a_2, \ldots, a_k >, s, D) =$$

$$\mathcal{T}(a_1, \mathcal{T}(< a_2, a_3, \ldots, a_k >, s, D), D).$$

A solution to a planning problem takes the form of a plan, denoted by π, where a plan is a sequence of actions whose execution from the initial state would result in a goal state, i.e., $\pi = \langle a_1, \ldots, a_k \rangle$ is a plan if for a model $\mathcal{M}$, we have $\mathcal{T}(\pi, I^{\mathcal{M}}, D^{\mathcal{M}}) \supseteq G^{\mathcal{M}}$.

We can additionally associate a cost with each action; however, to keep the formulation simple, we

will simply assume that each action has a unit cost and $C(\pi) = |\pi|$.

We will refer to a plan π as being optimal if there exist no other valid plans with cost $\leq C(\pi)$.

Chapter 3

Related Work

The recognition of potential dangers of misspecification of agent objectives has a long history within AI [Turing(1996), Wiener(1960)], and builds on ideas from even earlier philosophers. However, the modern form of the problem was effectively established by [Hadfield-Menell et al.(2016)], where they formalize the notion of assistive games to help optimize for the human's unspecified objective. Apart from the formalization, one of the core technical contributions of the work was the development of an algorithm to help generate more informative traces.

However, as we will see, such information would be influenced not only by their inability to perform correct introspection (commonly acknowledged in the literature) but also by their misunderstandings about the agent itself.

Other prominent works in this direction include research on reward design [Hadfield-Menell et al.(2017)], works that query the human about preferred behavior [Leike et al.(2018)], and other studies on generating informative traces [Fisac et al.(2017)].

Additionally, there are works that investigate the moral aspects of value alignment [Peterson(2019), Leike(2022)]; however, we will treat the problem of developing moral agents as being

orthogonal to the problem of aligning objectives. None of these works explicitly try to model the role played by the human and agent asymmetries in causing this misalignment in the first place.

3.1 Human-aware AI

Human-aware AI [Sreedharan et al.(2022a)] is a framework originally developed to generate explainable behavior, built on earlier efforts to use theory-of-mind in the context of human-AI interaction [Devin and Alami(2016)].

The framework hypothesizes that potential asymmetries between the human and the AI agent can cause a mismatch between the decisions chosen by the system and what the human would have expected. Such mismatches can confuse the human regarding why the agent may be following a particular action, which in turn would require the agent to explain its current decisions to the user. In general, these works identify three broad classes of asymmetries between the user and the agent [Sreedharan(2022)]: asymmetry in knowledge about the task, asymmetry in inferential capabilities, and asymmetry in vocabulary.

The explanation methods developed under the aegis of human-aware AI (cf. [Sreedharan et al.(2021a), Sreedharan et al.(2021b), Sreedharan et al.(2022b)]) tend to focus on identifying and addressing these asymmetries so that the agent and the user can reconcile their differences in expectations about the right course of action for a given problem.

In many ways, the goal of this work is to invert the process. We are trying to identify and leverage asymmetries to reconstruct and then try to meet the original expectations the human had, based on the information they provide.

3.2 Explicable Planning

In this sense, our work is also closely related to a method called explicable planning [Zhang et al.(2017b)], where the system tries to generate behavior that matches user expectations. However, in explicable planning, the final goal is usually provided, and the objective of the planning process is to generate plans that closely match behaviors that the human expects. In our case, we will not try to match the generated behavior with what the human expects, but rather focus solely on ensuring that the outcomes we generate satisfy what the user expected (the behavior that generates that outcome may look nothing like what the user expected).

3.3 Psychological and Social Sciences Values

A parallel thread of work in value alignment, that is orthogonal to our efforts, is the formulation of the set of values that the agent needs to be imbued with (cf. [Lera-Leri et al.(2022), Serramia et al.(2021), Montes and Sierra(2022)]). These works build on notions of values as determined in the wider psychological and social sciences literature [Schwartz(2012), Gouldner(1975)].

Our method is completely compatible with these efforts, as our objective is to ensure how these values, once identified, can be enforced in the agent. Our framework currently makes no commitments regarding what goals or objectives are specified by the user.

3.4 Model Elicitation

Another closely related set of works is that of model elicitation [Grover et al.(2020), Aineto et al.(2019)], preference elicitation [Mantik et al.(2022), Chen and Pu(2004)], resolving reward uncertainty [Zhang et al.(2017a), Wilson et al.(2012)], goal refinement [Mohajeriparizi et al.(2022)], and the technique of knowledge tracing [Corbett and Anderson(1994)], as applied in the context of intelligent tutoring systems. All these works are trying to solve a closely related problem: acquiring or eliciting some model information from a user or another agent.

However, such works are fundamentally incompatible with our setting, as none of the works in these areas currently allow the system to leverage information generated by users under potentially incorrect beliefs about the system.

Chapter 4

Motivating Example

4.1 Example Settings

Consider an intelligent robotic assistant that is used to help with daily household chores. The robot is expected to take task specifications along with any optional guidance from its users and fulfill their requirements.

We assume that the robot is aware that the goals specified by the user may be incomplete. As a specific example, consider a scenario where the user asks the robot to prepare a cup of tea. If the robot were to simply opt for the optimal plan, it might reach for the tea leaves closest to it and make tea with them. In this case, those leaves turn out to be low-quality tea stored at the bottom of the kitchen cupboard.

However, if the robot followed this plan, the prepared tea would not satisfy the user's expectations. When asking for a cup of tea, the user likely hoped for tea made with high-quality tea leaves but may have just forgotten to specify the quality or overlooked the possibility that the tea could have been made using poor-quality tea leaves.

4.2 The Problem

The robot cannot independently come up with what the human may have really wanted, and querying them about all other possibilities might be extremely difficult and resource-demanding. Fortunately, in this case, the human may be willing to provide additional instructions to guide the task. We will consider the simplest case where the human provides an entire plan (step by step) for making the tea. We will assume that the plan involves the robot fetching a ladder, placing it next to the cupboard, climbing the ladder to fetch good-quality tea leaves, and then making the tea. Unfortunately, this is not a plan the robot can execute on its own since, unbeknownst to the user, it cannot climb ladders. However, assuming this plan, at least in the human model, captures the human's true intent, it could provide the robot with clues about the actual human goal.

Once the robot understands the human's belief about its capabilities, it can try to simulate the plan in the human model to identify the expected outcome. It can then analyze which fluents that are true in the goal state may also contribute to achieving the true human goal. In our example, this could involve the fluent regarding the use of high-quality tea leaves, as well as the relative position of the ladder and whether the robot was expected to use it (the ladder).

One central challenge in this motivating example is to develop a method whereby the robot can find a plan that is guaranteed to satisfy the unspecified human goal while minimizing the number of times it needs to query the human for more information. This involves organizing the queries

effectively and determining how to ask them.

To address this challenge, we could explore techniques that leverage the information inferred from the provided plan, allowing the robot to refine its understanding of the user's expectations and fill in the gaps in the task specification without overburdening the user with questions.

Chapter 5

Goal Alignment Problem

5.1 Problem Settings

Our setting involves a robot—here used as a stand-in for any autonomous agent—that is expected to perform tasks assigned by a human. We will consider both a model of the robot and a model of the human's beliefs about that robot.

The domain model utilized by the robot (R) will be denoted as $D^R = \langle F, A^R \rangle$, with the initial state represented as I^R. In line with conventions from human-aware AI, we recognize that the human assigning the task may possess differing beliefs about the robot's model and/or the current state. These discrepancies can reflect the human's biases, their limited understanding of the task, or their restricted inferential capabilities.

The model representing the human's (H) beliefs about the robot will be denoted as $\mathcal{M}^H = \langle D^H, I^H, G^H \rangle$, where $D^H = \langle F, A^H \rangle$ is the domain model the human ascribes to the robot, I^H is the human's belief about the initial state, and G^H is the goal specified by the human for the robot. This goal specification is formulated based on the human's beliefs about the robot's capabilities and their own preferences regarding the expected outcome.

In the motivating example, G^H simply states that tea needs to be made. The assumption that both the human and robot share fluents (F) is common in human-aware planning problems (cf. [Sreedharan et al.(2022a)]), and methods like [Sreedharan et al.(2022b)] can help relax this assumption. The value alignment problem arises when optimizing the robot's specific objectives does not necessarily maximize the underlying human reward. In our setting, this means that a plan achieving the specified goal may not fulfill the true human goal. For example, the goal of making tea is misaligned if the robot follows a plan that neglects the need for high-quality tea leaves.

More formally, we define the goal misalignment problem as follows:

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 $\mathcal{T}(\pi, I^R, D^R) \not\supseteq G^*$.*

5.2 Addressing the Goal Misalignment Problem

Traditionally, one of the primary sources of information for addressing value alignment problems (cf. [Hadfield-Menell et al.(2016)]) comes from potential traces provided by humans that satisfy their underlying objectives. Such information assumes that, while a human may struggle to specify their objectives accurately, they can recognize when a state satisfying those objectives is achieved and may reason about how to reach such states.

In our case, this information is encapsulated in the human-specified plan π^H, which the human believes the robot can follow to achieve the goal. In our example, this corresponds to the user's plan involving the use of ladders.

The simplicity of this setting mitigates many traditional challenges identified in value alignment problems. Goals are a more straightforward structure for specifying objectives compared to rewards, which complicates many approaches, such as [Hadfield-Menell et al.(2017)] and [Leike et al.(2018)]. Empirical evidence shows that people struggle to specify effective reward functions [Booth et al.(2023)]. Conversely, psychological evidence indicates that people plan in terms of goals and subgoals [Simon(1977)], suggesting that specifying goals is easier than specifying rewards.

For deterministic tasks, a single plan suffices to reach the goal. Unlike [Hadfield-Menell et al.(2016)], we need not employ inverse-reinforcement learning algorithms to infer a more general reward function implied by the trace.

5.3 Foundational Problems and Formal Definition

The clarity of this setting allows us to examine foundational problems often obscured by complexity. Even in this straightforward context, the human's ability to specify objectives effectively hinges on their accurate understanding of the robot's capabilities and their ability to anticipate the

plans the robot may generate in response to the goal. Limitations in the human's inferential capabilities can hinder this anticipation, which is central to all value alignment problems.

Returning to the plan π^H, even if it could achieve the true goal in the human's mental model, there is no guarantee that the robot can execute it, nor that executing it will lead to the same goal state. In our example, the robot cannot execute the specified plan because it cannot perform the action of climbing the ladder.

To effectively achieve the human's expected goal state, the robot should focus on recreating the final state anticipated by the human rather than strictly following the provided plan. However, this presents another challenge, as the robot may not precisely replicate the final state resulting from executing the plan in the human's mental model.

Assuming there are fluents corresponding to the tools the robot uses, it will struggle to replicate the final state since it cannot climb the ladder and thus cannot turn the fluent related to the ladder being used true. This scenario aligns with cases where the human may have trajectory-level constraints, as these can be expressed as goal state fluents (cf. [Baier et al.(2009)]). Let the unknown goal the human has be G^*, with $G^H \subseteq G^*$. The central challenge is to determine whether the agent can achieve G^* and, if so, to devise a plan that satisfies it.

The human-specified plan offers insights into G^*. We can assert that G^* must be a subset of what the human believes would result from executing the plan ($\mathcal{T}(\pi^H, I^H, D^H)$). The challenge is iden-

tifying the exact subset. Although goals are intuitive for humans, directly querying them about G^* (e.g., asking, "Are you sure you only need me to achieve G^H?") is unlikely to succeed. The difference between G^H and G^* reflects their beliefs about the task, not merely forgotten fluents. For instance, in the tea-making task, a human might not specify the need for water, as they cannot envision making tea without it.

Instead, the robot could ask the human whether they care about a given fluent (e.g., "Would you mind if the tea was not made with water?"). We introduce a function $\mathcal{O}^{G^*} : F \rightarrow [0, 1]$ that returns 1 if a fluent is part of G^*.

The central computational challenge is to find plans that achieve the goal while minimizing queries to humans. With all components specified, we can formally define the central problem.

Definition 2. *A **human-aware goal alignment (HAGL)** is defined by the tuple $\mathcal{H} = \langle D^R, I^R, G^H, D^H, I^H, \pi^H, \mathcal{O}^{G^*} \rangle$, where there exists an unknown goal G^* such that $\mathcal{T}(\pi^H, I^H, D^H) \supseteq G^*$, $G^H \subseteq G^*$, and $\forall f \in F, \mathcal{O}^{G^*}(f) = 1$ if and only if $f \in G^*$. The robot's goal is to find a plan π^R such that $\mathcal{T}(\pi^R, I^R, D^R) \supseteq G^*$, if one exists, while minimizing queries to $\mathcal{O}^{G^*}$.*

As with many human-aware planning works, we assume access to D^H and I^H. Notably, the solution we propose to find a plan resulting in a superset of G^* aligns with cases where the human wishes to avoid undesirable side effects. This can be achieved by incorporating new fluents corre-

sponding to the negations of existing ones. Our current formulation can capture scenarios where a fluent corresponds to an undesirable side effect by adding the fluent for the negation of the undesirable fluent into the goal specification G^*.

To assess the complexity of the specified problem, we can compare it to classical planning, showing that it is at least as hard as solving classical planning problems, i.e., it is at least PSPACE-Hard.

Theorem 1. *The decision version of HAGL, i.e., the problem of determining whether there exists a plan for a given HAGL problem $\mathcal{H}$ that satisfies G^* with just K queries to $\mathcal{O}^{G^*}$, is at least PSPACE-Hard.*

Proof Sketch. We can establish this by demonstrating that a plan existence problem for a model $\mathcal{M} = \langle D, I, G \rangle$ (known to be PSPACE-Complete [Bylander(1994)]) can be compiled into a HAGL problem. Specifically, we set $G^* = G$, the robot domain model and initial state are the same as those in the original planning problem, and the human model includes an action a for each action in the original model. The human-specified plan will merely be a applied sequentially. The fluents $\{f_1, ..., f_n\}$ correspond directly to the original fluents from $\mathcal{M}$, and we construct $\mathcal{O}^{G^*}$ such that it indicates the fluents necessary for reaching the goal. □

This further highlights our argument that even when one removes many of the traditional complexities associated with value alignment, we still find a complex and challenging computational problem at the heart of the goal-alignment problem—one that could have clear implications for

everyday interactions humans have with AI systems.

One of the big advantages that this formulation has over the traditional ones is the fact that $\mathcal{T}(\pi^H, I^H, D^H)$ already gives an upper bound on the possible contents of the human goal. In fact, if the robot can achieve a state that is a superset of $\mathcal{T}(\pi^H, I^H, D^H)$, then that plan is guaranteed to satisfy the true human goal. This is only possible because the robot is maintaining an explicit model of the human's beliefs about the robot model.

However, this is just one way in which modeling human beliefs can help the robot in finding plans that satisfy the true human goal. As we will see in the next section, we can further leverage the human model to obtain better estimates on which of these goal fluents the human may have actually intended to achieve (as opposed to mere unintended side effects).

Chapter 6

A Solution for Goal Alignment Problem

In addition to introducing a new version of the value alignment problem, we also proposed a solution for the goal alignment problem, as described earlier. In particular, we approximated the value of information related to querying each fluent and then iteratively queried the ones with the highest value of information. We only used this procedure when G^H is achievable. However, the robot cannot necessarily 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 calculated the value associated with querying about each fluent as follows:

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

$$(1 - p(f \in G^*)) \times V(f \notin G^*)$$

where $p(f \in G^*)$ is the probability that fluent f is part of the goal, and $V(f \in G^*)$ represents the respective values of knowing whether f is part of the goal or not.

Let $S^H_{G^*}$ represent the state that results from executing the plan π^H in the human model (i.e.,

$S^H_{G^*} = \mathcal{T}(\pi^H, I^H, D^H)$), and let $\hat{F} \subseteq S^H_{G^*}$ be the set of fluents in the goal state that the robot cannot achieve in its true model.

To calculate the probability, we will employ a strategy similar to those used in goal recognition [Ramírez and Geffner(2010)]. Specifically, we will detect whether the suboptimality of the plan specified by the human may be explained by a given fluent f. If the inclusion of fluent f in the goal set (i.e., $G^H \cup \{f\}$) makes the optimal plan for the new goal in the human model closer to the cost of the specified plan, then we assign a higher probability to that fluent.

Following the conventions used by [Ramírez and Geffner(2010)], 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 a plan that is optimal in the human model for the goal $G^H \cup \{f\}$, and β is referred to as a rationality parameter that controls the randomness of the decision-maker. Note that this approach assumes the human follows a noisy rational decision-making process, an assumption that has psychological validity [Jeon et al.(2020)].

The value function we are interested in should reflect the certainty the robot has regarding the achievability of the goal state. If the robot knows for certain that the goal state can be achieved or cannot be achieved, then it will be set to 1. More formally, the value will be equal to the sum of the

probability that G^* is unachievable and the probability that there exists a single plan that achieves G^* (these two terms are mutually exclusive).

We can find a lower bound on this true value by just using the probability that the goal is unachievable:

$$V(f \in G^*) \geq \sum_{\bar{G}_f} P(G^* = \bar{G}_f) \times \delta(\bar{G} \text{ not solvable})$$

where $\bar{G}_f$ is any subset of $S^H_{G^*}$ containing G^H that satisfies $f \in G^*$ (i.e., $G^H \subseteq \bar{G} \subseteq S^H_{G^*}$ and $f \in \bar{G}_f$), $P(G^* = \bar{G})$ is the probability that the true goal is the same as $\bar{G}$, and $\delta(\bar{G}$ not solvable$)$ is an indicator function that evaluates to true if $\bar{G}$ is unsolvable. We can similarly define $V(f \notin G^*)$, but we will only consider subsets of the goal state that do not contain f.

Calculating this lower bound on the true value can still be computationally expensive, as it would require testing the achievability of every subset that satisfies the conditions discussed above (and calculating the probabilities as well). However, we can further find a lower bound for this lower bound by setting the value to be the probability of all the remaining fluents in $\hat{F}$ being part of the goal (which we approximate by multiplying the individual probabilities). This is a lower bound of the above equation because the set of all $\hat{G}_f$ is a superset of all possible goal candidates where $\hat{F}$ is present.

Specifically, we set the approximation as:

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

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 we can show that this formulation results in a lower bound when the remaining fluents are independent given the goal specification:

Proposition 1. *For a given HAGL problem for an $f \in S^H_{G^*}$, we will have $V(f \in G^*) \geq \tilde{V}(f \in G^*)$ and $V(f \notin G^*) \geq \tilde{V}(f \notin G^*)$, provided the probabilities $P(f_i \in G^*)$ and $P(f_i \notin G^*)$ are independent of other fluents in $\hat{F}$.*

Proof Sketch. This follows from two facts: (a) $\sum_{\bar{G}_f} P(G^* = \bar{G}_f) = P(f \in G^*)$ for any $f \in \hat{F}$, and $\bar{G}_f$ includes all sets that satisfy the specified condition, and (b) the set of $\bar{G}_f$ contains all subsets that satisfy $\hat{F}$. When $\tilde{V}(f \in G^*) = 1$, then $V(f \in G^*)$ must also equal one since all possible goals with f are unachievable. For the second case, we know that we cannot achieve any state that includes $\hat{F}$. These terms are part of the set $\bar{G}_f$, hence summing the probabilities over all unreachable $\bar{G}$ must be greater than the probability of $G^* = \hat{F}$. For cases where they are independent, the probability $G^* = \hat{F}$ will be equal to $\prod_{\hat{f} \in \hat{F} \setminus \{f\}} p(\hat{f} \in G^*)$. This proves the

first part; we can use similar reasoning to show the relation also exists between $V(f \notin G^*)$ and $\tilde{V}(f \notin G^*)$. □

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 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.

3. There exists a plan that can achieve the current subset of fluents the human has said yes to, along with G^H and any unqueried fluent.

The first two conditions correspond to cases where the robot cannot achieve the expected goal, while the latter indicates that the robot can achieve a superset of G^* and thus that plan would be acceptable to the human. Algorithm 1 presents the pseudocode for the overall procedure.

Proposition 2. *Algorithm 1 is complete for any given HAGL problem, i.e., it will always find a solution if one exists.*

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. In the case of the running motivating example, the set $\hat{F}$ only consists of the fluent corresponding to the use of the ladder. The fluents

Algorithm 1 An approximation-based algorithm to find a solution to a HAGL

Input: $\mathcal{H} = \langle D^R, I^R, G^H, \pi^H, \mathcal{O}^{G^*} \rangle$
$S^H_{G^*} = \mathcal{T}(\pi^H, I^H, D^H)$
if $\langle D^R, I^R, G^H \rangle$ not solvable **then**
  return No plan exists
end if
if $\langle D^R, I^R, S^H_{G^*} \rangle$ is solvable **then**
  return Return a valid plan for $\langle D^R, I^R, S^H_{G^*} \rangle$
end if
$Q \leftarrow$ A queue of fluents from the set $S^H_{G^*} \setminus G^H$ ordered by $\mathcal{V}^Q$
$\mathbb{C} \leftarrow \emptyset$
while Q is not empty **do**
  $f \leftarrow Q.pop()$
  if $\mathcal{O}^{G^*}$(f) == 1 **then**
    $\mathbb{C} = \mathbb{C} \cup \{f\}$
    if $\langle D^R, I^R, G^H \cup \mathbb{C} \rangle$ not solvable **then**
      return No plan exists
    end if
  else
    $\hat{G} = G^H \cup \mathbb{C} \cup Q$
    if $\langle D^R, I^R, \hat{G} \rangle$ is solvable **then**
      return Return a valid plan for $\langle D^R, I^R, \hat{G} \rangle$
    end if
  end if
end while
if $\langle D^R, I^R, G^H \cup \mathbb{C} \rangle$ not solvable **then**
  return No plan exists
else
  return Return a valid plan for $\langle D^R, I^R, G^H \cup \mathbb{C} \rangle$
end if

corresponding to the use of the ladder and the use of the high-quality tea leaves will be assigned the highest probability. In this case, the proposed algorithm generates a plan that achieves the remaining goal fluents once the human is queried about whether the ladder is part of the goal. Averaged across ten runs, we found that for the running example, our algorithm queries an average of 4.2 times (with the maximum number of queries being 8).

Chapter 7

Empirical Evaluation

For evaluation, we ran our method on a set of problems selected from standard IPC benchmark problems [International Planning Competition(2011)]. Our primary motivation was to test the effectiveness of our method in reducing the number of times the user would need to be queried before the true goal is found. Since we are unaware of any existing methods we can directly apply in this setting, we will compare the number of queries generated against a simple baseline that would query the user about all potential goal predicates.

Specifically, the hypothesis we will test is:

Hypothesis 1. *The average number of queries generated by our algorithm will be lower than the naive upper bound on the number of queries, which is equal to* $|S^H_{G^*} \setminus G^H|$.

7.1 Experimental Setup Design

We considered five domains: Blocksworld, Driverlog, Elevators, Rover, and Logistics. For each domain, we selected five instances used in previous competitions. The true goal in this case consisted of the goal specified as part of the original problem, while we created the goal specification

provided to the robot by randomly deleting a predicate from the goal specification.

The human model was formed by randomly deleting preconditions and deletes from the original domain description, while we used the original domain description as the robot model. All plans were generated using the FastDownward planner [Helmert(2006)], employing A-star search with the LM-cut heuristic [Helmert and Domshlak(2009)] and setting β to one for probability calculation.

All experiments were run on an AlmaLinux 8.9 machine with 32GB RAM and 16 Intel(R) Xeon(R) 2.60GHz CPUs.

7.2 Evaluation Results and Analysis

We ran our algorithm on each problem instance ten times, and the results from our evaluation are provided in Table 7.1. The second column in Table 1 presents the baseline upper bound on the number of queries, while the second and third columns list the average number of queries generated and the average time taken by our algorithm (along with their standard deviations).

The most striking result is that, apart from the Blocksworld domain, we observe a significant drop in the number of queries across almost all domains. In fact, for many problems, the algorithm doesn't even need to generate a single query to identify a plan guaranteed to satisfy the user's hidden goal. This indicates that for these problems, our method was able to find a plan that could

achieve a superset of the goal state expected by the user without requiring any queries.

In cases where the gains are less pronounced, particularly in Blocksworld, it seems to correspond to situations where the number of fluents in the goal states is small. This suggests that our method will be most effective in problems with a larger fluent set, and by extension, a larger state space. This property is particularly useful, as a naive querying strategy will not be viable in such problems.

Additionally, note that the time taken to complete the whole interaction is short and remains within acceptable bounds for real-time interaction with users.

The code for the experiments can be found at: https://github.com/HAPILab/GoalAlignment.

Table 7.1: A summary of the number of queries generated and the time taken by our method on standard IPC problems.

Problem Instance	$\lvert S^H_{G^*} \setminus G^H \rvert$	No of Queries		Time (secs)	
		Mean	Std	Mean	Std
Blocks	7	6.4	1.1	5.08	0.37
	3	2.6	0.52	2.72	0.2
	7	5.9	1.1	4.9	0.37
	4	3.8	0	3.37	0.1
	8	7.3	1.1	5.6	0.24
Driverlog	21	0	0	0.81	0.03
	24	0	0	1	0.02
	26	0	0	0.83	0.01
	23	0	0	0.9	0.01
	23	14.1	4.8	20.32	1.17
Elevator	25	0	0	0.71	0.02
	24	0	0	0.73	0.04
	25	14	4.16	13.30	1.04
	25	0	0	0.70	0.03
	24	6.7	4.35	11.07	1.05
Logistics	12	10.8	1.4	8.7	0.55
	13	0	0	0.78	0.03
	13	0	0	0.78	0.03
	12	9.8	2.2	8.63	0.48
	12	10.3	1.34	8.5	0.33
Rover	46	0	0	1.1	0.08
	42	0	0	1.07	0.05
	55	0	0	1.13	0.05
	55	29.3	11.88	34.72	3.4
	69	0	0	4.74	0.07

Chapter 8

Conclusion and Discussion

In this paper, we present a reformulation of the value alignment problem, which explicitly accounts for an often overlooked aspect of the problem, namely the asymmetry between the human's belief and the agent's true model. Even in this setting, we see that the goal alignment problem remains a challenging one.

We also explore how we could leverage human mental models to potentially generate better ways to query individuals for more information about their underlying objectives.

Our initial empirical evaluation shows that even this approximate algorithm helps reduce the number of queries we need to ask the human before the system can formulate a plan that is guaranteed to satisfy the true human goal.

There are multiple ways this work could be extended. One possibility would be to broaden the scope to support more complex decision-making settings, including decision-theoretic ones. Another avenue could involve investigating the use of alternative decision-making models for humans and relaxing assumptions about access to the human mental model of the robot.

While the value alignment problem is generally discussed in the context of AI safety, such mis-

specification and misalignment could affect every possible interaction between a human and an AI agent.

As such, we hope that more researchers working in the area of human-AI interaction will consider such misalignment issues when designing their systems.

Bibliography

[Aeronautiques et al.(1998)] Constructions Aeronautiques, Adele Howe, Craig Knoblock, ISI Drew McDermott, Ashwin Ram, Manuela Veloso, Daniel Weld, David Wilkins SRI, Anthony Barrett, Dave Christianson, et al. 1998. PDDL| The Planning Domain Definition Language. *Technical Report, Tech. Rep.* (1998).

[Aineto et al.(2019)] Diego Aineto, Sergio Jiménez, Eva Onaindia, and Miquel Ramírez. 2019. Model recognition as planning. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 29. 13–21.

[Baier et al.(2009)] 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.

[Booth et al.(2023)] Serena Booth, W Bradley Knox, Julie Shah, Scott Niekum, Peter Stone, and Alessandro Allievi. 2023. The perils of trial-and-error reward design: misdesign through overfitting and invalid task specifications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 5920–5929.

[Bylander(1994)] Tom Bylander. 1994. The Computational Complexity of Propositional STRIPS Planning. *Artif. Intell.* 69, 1-2 (1994), 165–204.

[Chen and Pu(2004)] Li Chen and Pearl Pu. 2004. *Survey of preference elicitation methods*. Technical Report. Ecole Politechnique Federale de Lausanne (EPFL).

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

[Corbett and Anderson(1994)] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[Devin and Alami(2016)] Sandra Devin and Rachid Alami. 2016. An implemented theory of mind to improve human-robot shared plans execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 319–326. https://doi.org/10.1109/HRI.2016.7451768

[Fikes and Nilsson(1971)] Richard E. Fikes and Nils J. Nilsson. 1971. Strips: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence* 2, 3 (1971), 189–208. https://doi.org/10.1016/0004-3702(71)90010-5

[Fisac et al.(2017)] 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. 2017. Pragmatic-Pedagogic Value Alignment. In *Robotics Research, The 18th International Symposium, ISRR 2017, Puerto Varas, Chile, December 11-14, 2017 (Springer Proceedings in Advanced Robotics, Vol. 10)*. Springer, Puerto Varas, Chile, 49–57.

[Geffner and Bonet(2013)] 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

[Gouldner(1975)] Helen Gouldner. 1975. THE NATURE OF HUMAN VALUES. By Milton Rokeach. New York: Free Press, 1973. 438 pp. *Social Forces* 53, 4 (06 1975), 659–660. https://doi.org/10.1093/sf/53.4.659

[Grover et al.(2020)] Sachin Grover, David E. Smith, and Subbarao Kambhampati. 2020. Model Elicitation through Direct Questioning. *CoRR* abs/2011.12262 (2020).

[Hadfield-Menell et al.(2017)] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J. Russell, and Anca D. Dragan. 2017. Inverse Reward Design. In *Advances in Neural Information*

Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. Curran Associates, Inc., Long Beach, CA, USA, 6765–6774.

[Hadfield-Menell et al.(2016)] 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.

[Helmert(2006)] Malte Helmert. 2006. The Fast Downward Planning System. *J. Artif. Intell. Res.* 26 (2006), 191–246.

[Helmert and Domshlak(2009)] Malte Helmert and Carmel Domshlak. 2009. Landmarks, Critical Paths and Abstractions: What's the Difference Anyway?. In *Proceedings of the 19th International Conference on Automated Planning and Scheduling, ICAPS 2009, Thessaloniki, Greece, September 19-23, 2009*. AAAI, Thessaloniki, Greece, 162–169.

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

[Jeon et al.(2020)] 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.

[Leike(2022)] Jan Leike. 2022. Our approach to alignment research. https://openai.com/blog/our-approach-to-alignment-research/

[Leike et al.(2018)] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *CoRR* abs/1811.07871 (2018).

[Lera-Leri et al.(2022)] Roger Lera-Leri, Filippo Bistaffa, Marc Serramia, Maite López-Sánchez, and Juan A. Rodríguez-Aguilar. 2022. Towards Pluralistic Value Alignment: Aggregating Value Systems Through l_p-Regression. In *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 780–788.

[Mantik et al.(2022)] Sheryl Mantik, Minyi Li, and Julie Porteous. 2022. A preference elicitation framework for automated planning. *Expert Systems with Applications* 208 (2022), 118014.

[Mohajeriparizi et al.(2022)] Mostafa Mohajeriparizi, Giovanni Sileno, and Tom van Engers. 2022. Preference-Based Goal Refinement in BDI Agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 917–925.

[Montes and Sierra(2022)] Nieves Montes and Carles Sierra. 2022. Synthesis and Properties of Optimally Value-Aligned Normative Systems. *J. Artif. Intell. Res.* 74 (2022), 1739–1774.

[Peterson(2019)] Martin Peterson. 2019. The value alignment problem: a geometric approach. *Ethics Inf. Technol.* 21, 1 (2019), 19–28.

[Premack and Woodruff(1978)] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.

[Ramírez and Geffner(2010)] 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.

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

[Schwartz(2012)] Shalom H Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online Readings Psychol. Cult.* 2, 1 (Dec. 2012).

[Serramia et al.(2021)] Marc Serramia, Maite López-Sánchez, Stefano Moretti, and Juan Antonio Rodríguez-Aguilar. 2021. On the dominant set selection problem and its application to value alignment. *Autonomous Agents and Multi-Agent Systems* 35, 2 (30 Jul 2021), 42. https://doi.org/10.1007/s10458-021-09519-5

[Simon(1977)] Herbert A Simon. 1977. The logic of heuristic decision making. In *Models of discovery*. Springer, New York, 154–175.

[Sreedharan(2022)] Sarath Sreedharan. 2022. *Foundations of Human-Aware Explanations for Sequential Decision-Making Problems*. Ph. D. Dissertation. Arizona State University.

[Sreedharan et al.(2021a)] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. 2021a. Foundations of explanations as model reconciliation. *Artif. Intell.* 301 (2021), 103558.

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

[Sreedharan et al.(2022b)] Sarath Sreedharan, Utkarsh Soni, Mudit Verma, Siddharth Srivastava, and Subbarao Kambhampati. 2022b. Bridging the Gap: Providing Post-Hoc Symbolic Explanations for Sequential Decision-Making Problems with Inscrutable Representations. In

The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, Virtual, o–1v9hdSult. https://openreview.net/forum?id=o-1v9hdSult

[Sreedharan et al.(2021b)] Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. 2021b. Using state abstractions to compute personalized contrastive explanations for AI agent behavior. *Artif. Intell.* 301 (2021), 103570.

[Turing(1996)] Alan M Turing. 1996. Intelligent machinery, a heretical theory. *Philosophia Mathematica* 4, 3 (1996), 256–260.

[Wiener(1960)] Norbert Wiener. 1960. Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science* 131, 3410 (1960), 1355–1358.

[Wilson et al.(2012)] Aaron Wilson, Alan Fern, and Prasad Tadepalli. 2012. A bayesian approach for policy learning from trajectory preference queries. *Advances in neural information processing systems* 25 (2012).

[Zhang et al.(2017a)] Shun Zhang, Edmund Durfee, and Satinder Singh. 2017a. Approximately-optimal queries for planning in reward-uncertain Markov decision processes. In *Twenty-Seventh International Conference on Automated Planning and Scheduling*.

[Zhang et al.(2017b)] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. 2017b. Plan explicability and predictability for robot task planning. In *2017 IEEE International Conference on Robotics and Automation, ICRA 2017, Singapore, Singapore, May 29 - June 3, 2017*. IEEE, Singapore, 1313–1320.