Goals vs. Rewards: Towards a Comparative Study of Objective Specification Mechanisms

Septia Rani
Department of Computer Science
Colorado State University
Fort Collins, Colorado, USA
septia.rani@colostate.edu

Serena Booth
Department of Computer Science
Brown University
Providence, Rhode Island, USA
serena_booth@brown.edu

Sarath Sreedharan
Department of Computer Science
Colorado State University
Fort Collins, Colorado, USA
sarath.sreedharan@colostate.edu

***Abstract*—In this late-breaking report, we look at two popular objective specification mechanisms for sequential decision-making problems, namely goals and rewards, and investigate how easy it would be for non-AI experts to use them effectively. Specifically, we propose a user study that allows us to test a user's ability to (a) use these mechanisms to direct a robot to generate some desired behavior and (b) predict the behavior resulting from a given objective specification. We conducted a small pilot study to test the study design and report some preliminary observations made regarding the two specification mechanisms.**

***Index Terms*—Objective Specification, Goals, Rewards**

I. Introduction

In this paper, we propose a study to examine the two most commonly used specification mechanisms in the sequential decision-making literature—rewards and goals—and see how well non-AI experts can work with them. Goals allow users to provide a partial specification of their desired end state. They are popularly used in classical planning [1] and have also received a lot of attention from recent work in using Large Language Models (LLMs) [2] for robot planning (cf. [3]). Rewards, on the other hand, are the underlying objective specification mechanisms used by reinforcement learning (RL) methods [4] and Markov Decision-making Processes (MDP) [5]. This allows one to associate numerical rewards with reaching some state or performing some action in a state.

We currently possess a rigorous understanding of these mechanisms' expressiveness and representational limitations (cf. [6]). Unfortunately, the ease with which users can express their underlying objectives in the expected forms has not, to our knowledge, been explicitly studied. While the development of LLMs has received attention as potentially intuitive interfaces to AI systems, they do not entirely solve the problem either. After all, the LLMs would need to translate the user utterances into the underlying objective specification, and it is unclear if these utterances would contain sufficient information needed for the translation.

In this late-breaking work, we discuss the design of a user study to examine the strengths and weaknesses of the two specification mechanisms when used by non-AI experts and present some initial results. In the proposed study, we expose participants to these objective specification mechanisms in intuitive tasks using simple interfaces and measure (a) how well the users are able to use the specific mechanism correctly and (b) how well they can understand an objective specified using each mechanism. While there have been some efforts at measuring the difficulty in specifying rewards [7], to the best of our knowledge, our work represents the first effort to perform such a comparative analysis of the two specification mechanisms among non-AI experts. Results from such studies could help us design instruction/objective specification interfaces that are intuitive and easy to use for everyday users. Such interfaces will allow them to instruct their robots more effectively, thus potentially avoiding objective misspecification and all related issues.

The rest of the paper is structured as follows: Section II will provide a brief discussion of rewards and goals as an objective specification mechanism and potential trade-offs. Section III will discuss the study design. We describe the specific hypotheses we focus on in Section IV and the preliminary results in Section V. Finally, the conclusion is described in Section VI.

II. Background

We will start by providing a brief sketch of the two specification mechanisms. To start with, goals as an objective specification mechanism is most commonly used in deterministic factored planning settings, also referred to as "*classical planning*" settings. In such cases, the states are represented by a set of boolean variables. A goal specification corresponds to a subset of these variables that the user would want to achieve, i.e., set true. A solution to such a classical planning problem corresponds to a plan, i.e., a sequence of actions whose execution in the initial state results in a state where the state variables listed in the goal specification are true. In the simplest formalism, an optimal plan corresponds to the shortest possible plan[1].

For reward functions, the formulation we will adopt is one where a reward is associated with a state action pair. We

Sarath Sreedharan's research is supported in part by NSF grant # 2303019 and other transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program.

[1]However, there are more expressive formalisms that allow one to associate non-unit costs with actions

will again assume a factored state representation, where a set of boolean variables represents each state. We will extend this factored representation to the reward and will assign a reward value to each state variable and action pair. The reward received for executing an action in a state equals the sum of individual rewards for each state variable true in the state. In an MDP planning or RL setting, the objective is to maximize the expected discounted sum of rewards. A solution here, named a policy, corresponds to a mapping from a state to an action to be executed in that state. An optimal policy here corresponds to one that returns the highest possible expected discounted sum of rewards, also referred to as the value of the policy.

At first glance, it is easy to see that a goal specification only provides information about the end state, while a reward function allows us to specify signals for desirable intermediate signals. However, this is not to say that reward functions subsume goals. It is worth noting that a goal naturally corresponds to an end state, whose achievement corresponds to the robot completing the current task. On the other hand, rewards cannot easily capture such considerations; rather, the most common way to encode such requirements is to turn some states into absorber states. In an absorber state, the transition dynamics are modified so that once the agent enters that state, it can no longer transition to a different state. Transition functions usually capture how the world or the environment state evolves in response to a robot's action. As such, it would be strange to modify the transition function every time the agent is assigned a new task. In this paper, we will instead make use of an exit action that the robot can perform to stop the task execution. Once such an exit action is available to the robot, one can easily translate a goal to a reward function by simply assigning a high reward to goal states. For discounted infinite horizon MDPs [5], such a reward function would automatically lead to the robot trying to reach the goal state as soon as it can.

It is worth noting that goals and rewards are not the only objective specification mechanisms that have been considered. Current literature has considered other mechanisms like reward machines [8], program snippets [9], and even fragments of temporal logic [10]. The reason we choose goals and rewards is because they are more foundational and widely used than the other mechanisms.

III. Study Design

To compare the two mechanisms, we designed three intuitive but diverse domains in which two primary tasks related to each mechanism can be tested: (1) the user's ability to provide an objective specification that will result in a given behavior and (2) their ability to predict the behavior from a given specification. We chose domains that non-AI experts could understand without considerable training but corresponded to potential real-world robotics applications. Specifically, the domains included (1) a robot navigation task, (2) a tabletop pick and place task, and (3) a task with a self-driving vehicle. We chose deterministic versions of the tasks to avoid potential confounders that may arise from the stochasticity of the environment dynamics. The environment setting for each domain can be seen in Figure 1.

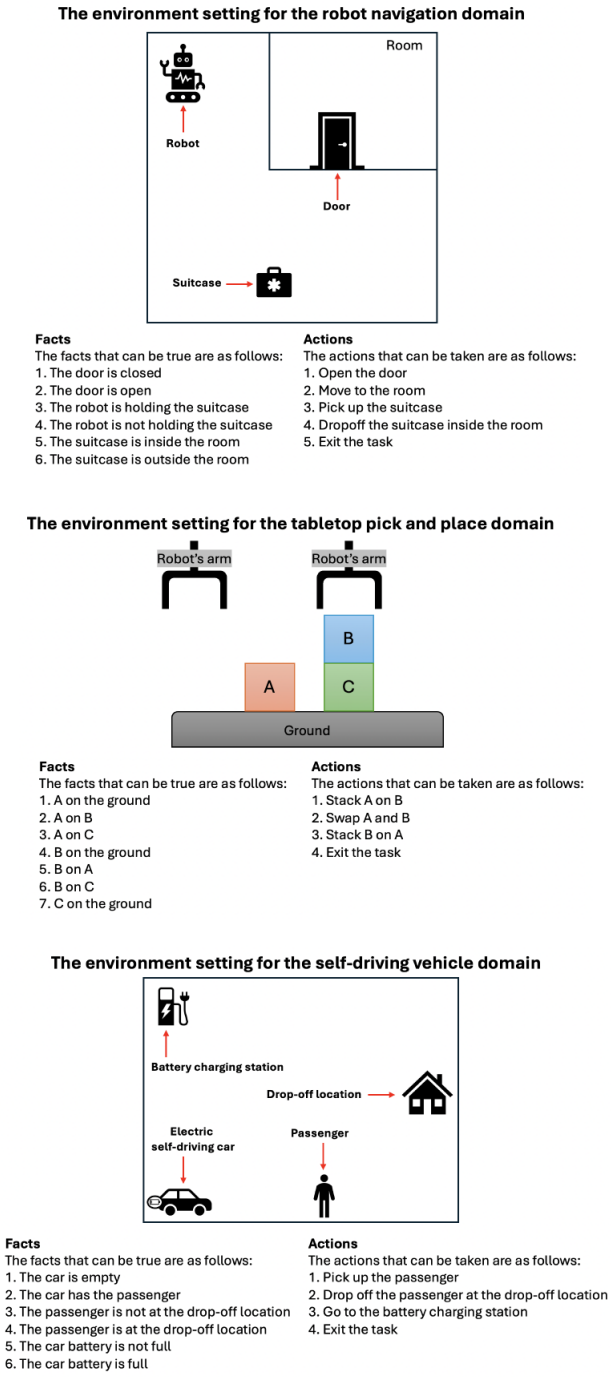

Fig. 1. A visualization of each domain used in the study.

The navigation task involves robots navigating through a workspace. In this case, we have a robot that needs to pick up and drop off a suitcase in different locations within a small workspace. The pick and place domain contains a set of blocks that can be stacked on top of one another. The objective is usually to achieve a specific configuration of the blocks. For

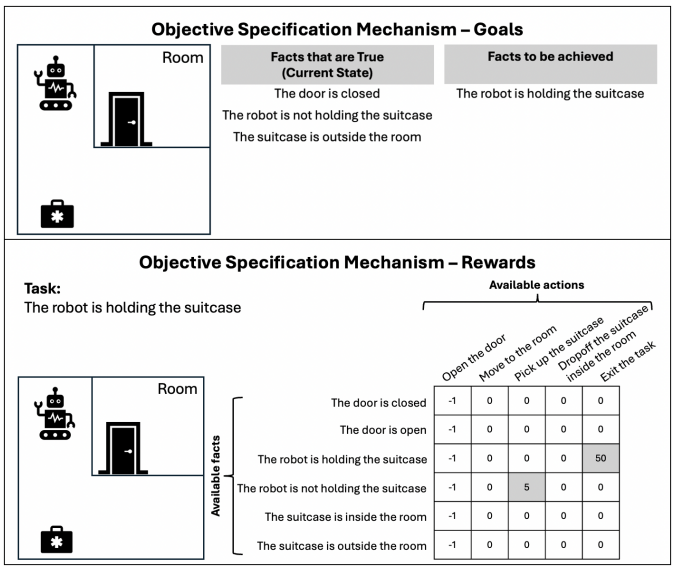

Fig. 2. Illustrations for the sample specifications that could be shown to the participants.

the self-driving vehicle domain, we have a self-driving car powered by a battery that needs to pick up and drop off a passenger in different locations. It also needs to charge the battery to make sure that the battery is enough to perform its task. In each environment setting, the current state is defined by a set of binary variables, henceforth referred to as facts. There is also a set of actions that can be taken by the robot, including an exit action that will allow the robot to end the task. Each domain had about 6-7 facts and 4-5 actions. We choose to keep the facts and action counts similar so as to balance the workload between domains.

We will use these domains to create surveys that will test the participants' ability to specify an objective that will result in some provided behavior or their ability to predict what behavior will result from a given objective. The survey we propose to build around these scenarios will use a mixed study design, combining both between-subjects and within-subjects study designs. The participants will be shown either the specification task or prediction task (between subjects), chosen from three different problem domains as mentioned above. Given the problem domain, the participants will be tested on how well they are able to complete the specified task across the two objective specification mechanisms (within subjects). We will use a counterbalancing technique to vary the order in which participants will be shown the different specification mechanisms. This is to ensure that no single order influences the results of the study.

For each objective specification mechanism, there are two sections in the survey: demo and test. The demo section is basically a learning phase, where participants are familiarized and introduced to the concepts of goal and reward specifications. In the demo section, participants will be shown a video that demonstrates a simple behavior along with the corresponding goal or reward (see the example illustration in Figure 2). For goals, the video will show the "facts to be achieved (goal state)" and how the "facts that are true (current state)" change during the duration of robot behavior until it reaches the goal state. On the other hand, for rewards, the video shows the rewards matrix and how individual rewards from the matrix will be added to the total when the agent performs specific actions. For example, based on the illustration in Figure 2, the agent will get 50 points if it takes an "exit the task" action while the fact that "the robot is holding the suitcase" is true.

For the first task, i.e., ease of objective specification, the test section will show a sample behavior to the user. Then, participants are asked to come up with goals and/or rewards for that scenario. We refer to goals as facts and rewards as scores to simplify the description to non-AI expert participants. From the participants' answers, we can determine whether their specifications are correct or incorrect. For the incorrect one, the potential sources of errors can be analyzed, including over-specification and under-specification.

On the other hand, to test how easily non-AI experts can understand goals and rewards, instead of showing the demonstration, we show the correct goal (list of facts to be achieved) or the rewards specification (in the form of scores). Then, we ask the participants to predict or interpret the behavior of the agent based on that. Specifically, we provide three video options and ask them to choose one that most aligns with the given goals or rewards.

Additionally, at the end of the survey, we ask the participants to directly compare the two specification mechanisms in terms of their easiness, intuitiveness, likeability, and challenge. We also ask for qualitative feedback on why they think that particular objective specification mechanism is easier or harder than the other. Finally, we collect demographic information such as age, gender, highest level of education, and familiarity with computer science and AI subjects.

IV. Hypotheses

Our study is primarily designed to measure how the choice of specification mechanism can affect the user's ability to specify objectives and predict agent behavior. The primary hypotheses we plan to test here are as follows:

H1-a Participants are more likely to provide accurate goal specifications than accurate reward specifications.

H1-b Participants are more likely to correctly interpret goal specifications than reward specifications.

The next question we would like to ask is in regard to the workload, in particular cognitive load, imposed by the two mechanisms.

H2-a Reward specifications will result in a higher workload than goal specifications.

H2-b Trying to interpret reward functions will result in a higher workload than goal specifications.

Now, we also wanted to use this as an opportunity to understand ways in which the users may incorrectly specify their objectives, which brings us to the hypothesis:

H3 Participants are more likely to underspecify objectives than overspecify them.

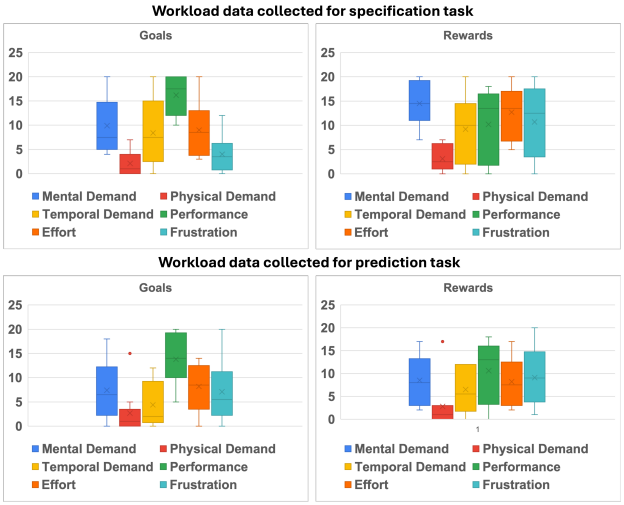

Fig. 3. The results collected from the NASA-TLX questionnaire for the two mechanisms across the two tasks.

We will test the above hypothesis for both reward and goal specification cases.

To assess the H2, we will measure the participants' workload for each objective specification mechanism and task in the survey. NASA Task Load Index (TLX) is used to measure the perceived workload. NASA TLX has six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration level [11]. Each dimension is measured using the rating scale.

V. Preliminary Results

We conducted an initial pilot study on a small participant pool, not to assess the hypotheses but to assess the study design and process. We recruited a total of 20 participants from Prolific: 10 participants for the specification task and 10 participants for the prediction task. They were paid six dollars for twenty minutes, and they identified their native language as English. For each task, there were five men and five women. The majority of them reported having never taken an AI course. This study was IRB-approved. Participants were provided with informed consent before they started the survey. Multiple attention check questions were included throughout the study. Each participant was randomly shown one of the three domains, and the order in which the specification mechanism was shown was randomized to ensure the results were counterbalanced.

a) Testing User's Ability To Provide Specification.: Out of the ten participants, six participants were able to provide a correct goal specification. On the other hand, only three participants provided correct reward specifications. This seems to provide some preliminary evidence to support H1-a. More interestingly, four of the six correct goal specifications were over-specified, while we only saw one under-specified goal. This seems to be less aligned with our hypothesis H3. Note that we do not count under-specified objectives as correct. While underspecified objective specifications could support the demonstrated behavior, they could also potentially support other behaviors. Looking at the overspecified objectives, we see that the participants were trying to directly encode the full behavior into the specification. This behavior has been noted in reward specification (cf. [7]), but it is interesting to see that such behavior carries over to goal specification. Figure 3 plots the results from the NASA-TLX questionnaire. We see an increase across all dimensions, except effort, as we move from goal to reward settings. The increase in load is particularly apparent for mental demand. It is interesting that the effort[2] did not show a noticeable difference, which could be attributed to the lower sample count. In general, the results seem to suggest that H2-a may hold as well.

b) Testing User's Ability To Predict Behavior From Specifications.: Moving on to the second task, we see that seven out of ten participants could predict the correct behavior from goal specification. Similarly, there are also seven participants who could correctly predict the behavior from the given reward specification. As such, the data is not sufficient to make any claims about H1-b. Figure 3 also plots the results from the NASA-TLX questionnaire for this task. We see that the results for most dimensions are quite comparable, except for goals the participants seem to have self-reported higher performance. Both outlier points in Figure 3 came from the same participant.

In our subjective questionnaire, most participants marked the goal mechanism as being the most intuitive, easiest, and the one they liked the most. This preference for goals over rewards is also reflected in much of the free text feedback we collected. For goals, we saw comments like how it was "just seems easier to get the goals accomplished" and "very straightforward." On the other hand, for rewards, we saw comments like "values in the tables are confusing" and "hard to remember the score numbers". This was true for both tasks.

VI. Conclusion

This paper presents a potential user study comparing two widely used objective specification mechanisms in terms of their usability by everyday users. We ran an initial study on the setup to collect some preliminary results related to such a comparison. While we see some results one would expect, like goals being easier to specify than rewards, we also see surprising ones. For example, the fact that even though people seem to be able to do as well on predicting behavior from rewards as much as goals, they still found goals easier and more intuitive. Based on the current results of the pilot study, no changes are needed in the study design. In future work, we hope to run this study on a larger participant pool and perform statistical analysis to verify the hypotheses. We also hope that this work spurs more interest in the question of the advantages and disadvantages of different objective specifications.

Acknowledgment

We would like to thank Malek Mechergui for the help in preparing the initial draft of the survey.

[2]for the effort the user is asked to rate on a scale of one to 20 "*How hard did you have to work to accomplish your level of performance?*"

REFERENCES

[1] M. T. Cox, "A model of planning, action, and interpretation with goal reasoning," in *Proceedings of the 4th Annual Conference on Advances in Cognitive Systems*, 2016, pp. 48–63.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[3] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on robot learning*. PMLR, 2023, pp. 287–318.

[4] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[5] M. L. Puterman, "Markov decision processes," *Handbooks in operations research and management science*, vol. 2, pp. 331–434, 1990.

[6] D. Abel, W. Dabney, A. Harutyunyan, M. K. Ho, M. L. Littman, D. Precup, and S. Singh, "On the expressivity of markov reward," in *NeurIPS*, 2021, pp. 7799–7812.

[7] S. Booth, W. B. Knox, J. Shah, S. Niekum, P. Stone, and A. Allievi, "The perils of trial-and-error reward design: Misdesign through overfitting and invalid task specifications," in *AAAI*. AAAI Press, 2023, pp. 5920–5929.

[8] R. T. Icarte, T. Q. Klassen, R. A. Valenzano, and S. A. McIlraith, "Using reward machines for high-level task specification and decomposition in reinforcement learning," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 2112–2121.

[9] J. Andreas, D. Klein, and S. Levine, "Modular multitask reinforcement learning with policy sketches," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 166–175.

[10] G. D. Giacomo, L. Iocchi, M. Favorito, and F. Patrizi, "Foundations for restraining bolts: Reinforcement learning with ltlf/ldlf restraining specifications," in *ICAPS*. AAAI Press, 2019, pp. 128–136.

[11] S. G. Hart, "Nasa task load index (tlx)," 1986.