When Humans Aren't Optimal: Agents that Collaborate with Risk-Aware Humans

Minae Kwon* **Erdem Biyik*** **Aditi Talati*** **Karan Bhasin‡** **Dylan P. Losey†**

Dorsa Sadigh*

Abstract

In order to collaborate safely and efficiently, AI agents need to anticipate how their human partners will behave. Some of today's agents model humans as if they were also agents, and assume users are always *optimal*. Other agents account for human limitations, and relax this assumption so that the human is *noisily* rational. Both of these models make sense when the human receives deterministic rewards: i.e., gaining either \$100 or \$130 with certainty. But in real-world scenarios, rewards are rarely deterministic. Instead, we must make choices subject to *risk* and *uncertainty*—and in these settings, evidence suggests humans exhibit a cognitive bias towards *suboptimal* behavior [1]. For example, when deciding between gaining \$100 with certainty or \$130 only 80% of the time, people tend to make the risk-averse choice—even though it leads to a lower expected gain! In this paper, we adopt a well-known *Risk-Aware* human model from behavioral economics called Cumulative Prospect Theory and enable agents to leverage this model during human-agent interaction. In our user studies, we offer supporting evidence that the Risk-Aware model more accurately predicts suboptimal human behavior. We find that this increased modeling accuracy results in safer and more efficient human-agent collaboration. Overall, we extend existing rational human models so that collaborative agents can anticipate and plan around suboptimal human behavior during human-agent interaction.

1 Introduction

When agents [2] collaborate with humans, they must *anticipate* how the human will behave for seamless and safe interaction. Consider a scenario where an autonomous car is waiting at an intersection (see top of Fig. 1). The autonomous car wants to make an unprotected left turn, but a human driven car is approaching in the oncoming lane. The human's traffic light is yellow, and will soon turn red. Should the autonomous car predict that this human will stop—so that the autonomous car can safely turn left—or anticipate that the human will try and make the light—where turning left leads to a collision?

Previous agents anticipated that humans acted like agents, and made *rational* decisions to maximize their reward [2, 3, 4, 5, 6, 7]. However, assuming humans are always rational fails to account for the limited time, computational resources, and noise that affect human decision making, and so today's agents anticipate that humans make *noisily* rational choices [8, 9, 10, 11, 12]. Under this model, the human is always most likely to choose the action leading to the highest reward, but the agent also recognizes that the human may behave suboptimally. This makes sense when humans are faced with

** Stanford University, † Virginia Tech, ‡ The Harker School [mnkwon, ebiyik, atalati, dorsa]@stanford.edu, losey@vt.edu, karanbhasin03@gmail.com

[2] While we use the term *agent* to refer to all AI agents, we consider an autonomous car and a robot as examples of agents in this work.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

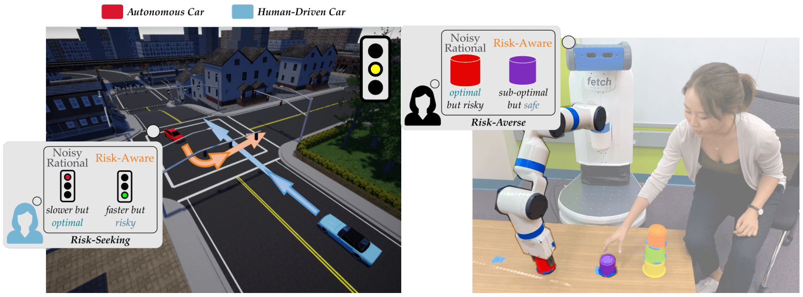


Figure 1: Agents collaborating with humans that must deal with risk. (Left) Autonomous car predicting if the human will try to make the light. (Right) Robot arm anticipating which cup the human will grab. In real world scenarios, people exhibit a cognitive bias towards irrational but Risk-Aware behavior.

deterministic rewards: e.g., the light will definitely turn red in 5 seconds. Here, the human knows whether or not they will make the light, and can accelerate or decelerate accordingly. But in real world settings, we usually do not have access to deterministic rewards. Instead, we need to deal with uncertainty and estimate *risk* in every scenario. Returning to our example, imagine that the human has a 95% chance of making the light if they accelerate: success saves some time during their commute, while failure could result in a ticket or even a collision. It is still rational for the human to decelerate; however, a risk-seeking user will attempt to make the light. How the agent models the human affects the *safety* and *efficiency* of this interaction: a Noisy Rational agent believes it should turn left, while a *Risk-Aware* agent realizes that the human is likely to run the light, and waits to prevent a collision.

When agents treat nearby humans as noisily rational, they *miss out* on how risk biases human decisions. Instead, we assert: *to ensure safe and efficient interaction, agents must recognize that people behave suboptimally when risk is involved.*

Our approach is inspired by behavioral economics, where results indicate that users maintain a nonlinear transformation between *actual* and *perceived* rewards and probabilities [1, 13]. Here, the human over- or under-weights differences between rewards, resulting in a cognitive bias (a systematic error in judgment) that leads to risk-averse or risk-seeking behavior. We equip agents with this cognitive model, enabling them to anticipate risk-affected human behavior and better collaborate with humans in everyday scenarios (see Fig. 1).

Overall, we make the following contributions:

Incorporating Risk in Models of Humans. We propose using Cumulative Prospect Theory as a Risk-Aware model. We formalize a theory-of-mind (ToM) where the agent models the human as reacting to their decisions or environmental conditions. We integrate Cumulative Prospect Theory into this formalism so that the agent can model suboptimal human actions under risk. In a simulated autonomous driving environment, our user studies demonstrate that the Risk-Aware agent more accurately predicts the human's behavior than a Noisy Rational baseline.

Determining when to Reason about Risk. We identify the types of scenarios where reasoning about risk is important. Our results suggest that scenarios with close expected rewards is the most important in determining whether humans will act suboptimally.

Safe and Efficient Collaboration when Humans face Uncertainty. We develop planning algorithms so that agents can leverage our Risk-Aware human model to improve collaboration. In a collaborative cup stacking task, shown on the bottom in Fig. 1, we use a robot arm as an example of a Risk-Aware agent. The Risk-Aware robot arm anticipated that participants would choose suboptimal but risk-averse actions, and planned trajectories to avoid interfering with the human's motions. Users completed the task more efficiently with the Risk-Aware agent, and also subjectively preferred working with the Risk-Aware agent over the Noisy Rational baseline.

2 Related Work

Previous work has shown that agents that model humans' behavior exhibit improved performance in many applications, such as assistive robotics [14, 15, 16, 17], motion planning [18, 19], collaborative games [20], and autonomous driving [21, 22, 23]. One reason behind this success is that human modeling equips agents with a theory of mind (ToM), or the ability to attribute a mind to oneself and others [24, 25]. [26] showed ToM can improve performance in human-agent collaboration.

For this purpose, researchers have developed various human models. In robotics, the Noisy Rational choice model has remained extremely popular due to its simplicity. Several works in reward learning [9, 27, 28, 29, 12, 30], reinforcement learning [11], inverse reinforcement learning [31, 10, 32], inverse planning [33], and human-robot collaboration [34] employed the noisy rational model for human decision-making.

In cognitive science, psychology and behavioral economics, researchers have developed other decision-making models. For example, [35] investigated decision making under time constraints; [36] developed a model based on stochastic processes to model humans' process of making a selection between two options, again under a time constraint. Among all of these works, Cumulative Prospect Theory (CPT) [13] remains prominent as it successfully models suboptimal human decision making under risk. Later works studied how Cumulative Prospect Theory can be employed for time-constrained decision making [37, 38].

In this paper, we adopt Cumulative Prospect Theory as an example of a Risk-Aware model. We show that it not only leads to more accurate predictions of human actions, but also increases the performance of the agent and the human-agent team.

3 Formalism

We outline and compare two methods: Noisy Rational and Cumulative Prospect Theory (CPT) [13]. CPT is a prominent model of human decision-making under risk [37, 38] and we use it as an example of a Risk-Aware model. For more details on the CPT model and how to integrate both models into a partially observable Markov decision process (POMDP), please see the Appendix.

We assume a setting where a human needs to select from a set of actions $\mathcal{A}_H$. Each action $a_H \in \mathcal{A}_H$ may have several possible consequences, where, without loss of generality, we denote the number of consequences as K. For a given human action a_H, we express the probabilities of each consequence and their corresponding rewards as a set of pairs:

$$C(a_H) = \left\{ \left(p^{(1)}, R_H^{(1)}(a_H)\right), \left(p^{(2)}, R_H^{(2)}(a_H)\right), \ldots, \left(p^{(K)}, R_H^{(K)}(a_H)\right) \right\}$$

Noisy Rational Model. According to the noisy rational model, humans are more likely to choose actions with the highest expected reward, and are less likely to choose suboptimal actions (i.e., they are optimal with some noise). The noise comes from constraints such as limited time or computational resources. For instance, in the autonomous driving example, Noisy Rational model would predict the human will most likely choose the optimal action and decelerate. Denoting the expected reward of the human for action a_H as

$$R_H(a_H) = p^{(1)} R_H^{(1)}(a_H) + p^{(2)} R_H^{(2)}(a_H) + \ldots, p^{(K)} R_H^{(K)}(a_H),$$

the noisy rational model asserts

$$P(a_H) = \frac{\exp\left(\theta \cdot R_H(a_H)\right)}{\sum_{a \in \mathcal{A}_H} \exp\left(\theta \cdot R_H(a)\right)}, \tag{1}$$

where $\theta \in [0, \infty)$ is a temperature parameter, commonly referred to as the *rationality coefficient*, which controls how noisy the human is. While larger θ models the human as a better reward maximizer, setting $\theta = 0$ means the human chooses actions uniformly at random.

Hence, the Noisy Rational model is simply a linear transformation of the reward with the rationality coefficient $\theta \geq 0$, which makes the transformation monotonically non-decreasing. As the model does not transform the probability values, it becomes impossible to model suboptimal humans using this approach. The closest Noisy Rational can get to modeling suboptimal humans is to assign a uniform probability to all actions.

Risk-Aware Model. We adopt Cumulative Prospect Theory (CPT) [13] as an example of a Risk-Aware model. According to this model, humans are not simply Noisy Rational. They may, for example, be suboptimally risk-seeking or risk-averse. The Risk-Aware model captures suboptimal decision-making by transforming both the probabilities and the rewards. These transformations aim to represent what humans actually perceive. The reward transformation is a pairwise function:

$$v(R) = \begin{cases} R^{\alpha} \text{ if } R \geq 0 \\ -\lambda(-R)^{\beta} \text{ if } R < 0 \end{cases} .$$

The parameters $\alpha, \beta \in [0, 1]$ represent how differences among rewards are perceived. For instance, when $\alpha, \beta \in (0, 1)$, the model predicts that humans will perceive differences between large positive (or negative) rewards as relatively lower than the differences between smaller positive (resp. negative) rewards, even though the true differences are equal. $\lambda \in [0, \infty)$ characterizes how much more (or less) important negative rewards are compared to positive rewards. When $\lambda > 1$, humans are modeled as loss-averse, assigning more importance to losses compared to gains. The reverse is true when $\lambda \in [0, 1)$.

The Risk-Aware model also implements a transformation over the probabilities. The probabilities $(p^{(1)}, p^{(2)}, \dots)$ are divided into two groups based on whether their corresponding true rewards are positive or negative. The probability transformations corresponding to positive and negative rewards (w^+, w^-) are as follows:

$$w^+(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{1/\gamma}},\ w^-(p) = \frac{p^\delta}{(p^\delta + (1-p)^\delta)^{1/\delta}},$$

where $\gamma, \delta \in [0, 1]$ [3]. For details on how to construct the probability of an action using these transformations, please see the Appendix.

4 Autonomous Driving

In our first user study, we focus on the autonomous driving scenario from the left of Fig. 1. Here the autonomous car—which wants to make an unprotected left turn—needs to determine whether the human-driven car is going to try to make the light. We asked human drivers whether they would accelerate or stop in this scenario. Specifically, we adjusted the *information* and *time* available for the human driver to make their decision. We also varied the level of *risk* by changing the probability that the light would turn red. Based on the participant's choices in each of these cases, we learned *Noisy Rational* and *Risk-Aware* human models. Our results demonstrate that autonomous cars that model humans as *Risk-Aware* are better able to explain and anticipate the behavior of human drivers, particularly when drivers make suboptimal choices.

Experimental Setup. We used the driving example shown in Fig. 1. Human drivers were told that they are returning a rental car, and are approaching a light that is currently yellow. If they run the red light, they have to pay a \$500 ticket. But stopping at the light will prevent the human from returning their rental car on time, which also has an associated fine! Accordingly, the human drivers had to decide between *accelerating* (and potentially running the light) or *stopping* (and returning the rental car with a fine).

Independent Variables. We varied the amount of *information* and *time* that the human drivers had to make their decision. We also tested two different *risk* levels: one where accelerating was optimal, and one where stopping was optimal. Our parameters for information, time, and risk are provided in Table 1 in the Appendix.
Information. We varied the amount of information that the driver was given on three levels: None, Explicit, and Implicit. Under None, the driver must rely on their own prior to assess the probability that the light will turn red. By contrast, in Explicit we inform the driver of the exact probability. Because probabilities are rarely given to us in practice, we also tested Implicit, where drivers observed other peoples' experiences to estimate the probability of a red light.
Time. We compared two levels for time: a Timed setting, where drivers had to make their choice in under 8 seconds, and a Not Timed setting, where drivers could deliberate as long as necessary.
Risk. We varied risk along two levels: High and Low. When the risk was High, the light turned red 95% of the time, and when risk was Low, the light turned red only 5% of the time.

Dependent Measures. We aggregated the user responses into *action distributions*. These distributions report the percentage of human drivers who chose to accelerate and stop under each treatment level. Next, we learned *Noisy Rational* and *Risk-Aware* models of human drivers for the autonomous car to leverage[4]. To measure the accuracy of these models, we compared the Kullback-Leibler (KL)

[3] For details on the probability and reward transformations please consult page 309 of [13]

[4] To learn the models, we used the Metropolis-Hastings algorithm [39] and obtained 30 independent samples of the model parameters.

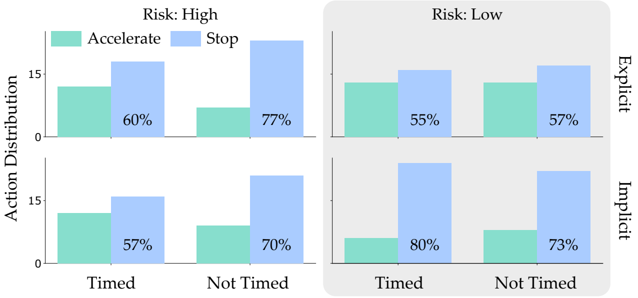


Figure 2: Action distributions for human drivers. Stopping was the *suboptimal* choice when the light rarely turns red (Low).

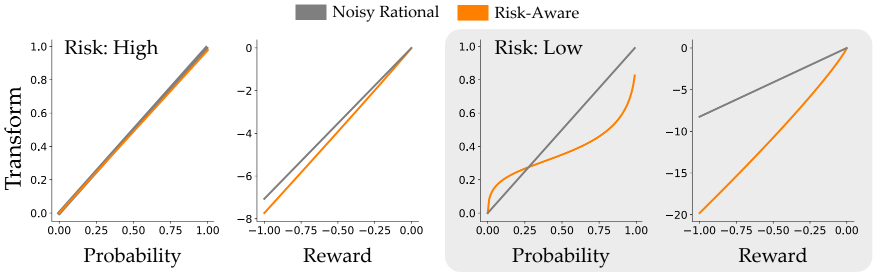


Figure 3: Averaged probability and reward transformations for human drivers that are modeled as *Noisy Rational* or *Risk-Aware*.

divergence between the *true* action distribution and the model's *predicted* action distribution. We report the *log KL divergence* for both *Noisy Rational* and *Risk-Aware* models in Fig. 4.

Hypothesis 1. Autonomous cars which use *Risk-Aware* models of human drivers will more accurately predict human action distributions than autonomous cars who treat humans as *noisily rational* agents.

Baseline. In order to confirm that our users were trying to make optimal choices, we also queried the human drivers for their preferred actions in settings where the expected rewards were *far apart* (e.g., where the expected reward for accelerating was much higher than the expected reward for stopping). In these baseline trials, users overwhelmingly chose the *optimal* action (93% of trials).

Results. The results from our autonomous driving user study are summarized in Figs. 2, 3, and 4. In each of the tested situations, most users elected to stop at the light (see Fig. 2). Although stopping at the light is the optimal action in the High risk case—where the light turns red 95% of the time—stopping was actually *suboptimal* in the Low risk case—where the light only turns red 5% of the time. Because humans chose optimal actions in some cases (High risk) and suboptimal actions in other situations (Low risk), the autonomous car interacting with these human drivers must be able to anticipate *both* optimal and suboptimal behavior.

In cases where the human was rational, autonomous cars learned similar *Noisy Rational* and *Risk-Aware* models (see Fig. 3). However, the *Risk-Aware* model was noticeably different in situations where the human was suboptimal. Here autonomous cars using our formalism learned that human drivers *overestimated* the likelihood that the light would turn red, and *underestimated* the reward of running the light. Viewed together, the *Risk-Aware* model suggests that human drivers were risk-averse when the light rarely turned red, and risk-neutral when the light frequently turned red.

Autonomous cars using our *Risk-Averse* model of human drivers were better able to predict how humans would behave (see Fig. 4). Across all treatment levels, *Risk-Averse* attained a log KL divergence of -5.7 ± 3.3, while *Noisy Rational* only reached -3.3 ± 1.3. This difference was statistically significant ($t(239) = -11.5$, $p < .001$). Breaking our results down by risk, in the High case both models were similarly accurate, and any differences were insignificant ($t(119) = .42$, $p = .67$). But in the Low case—where human drivers were suboptimal—the *Risk-Averse* model significantly outperformed the *Noisy Rational* baseline ($t(119) = -17.3$, $p < .001$).

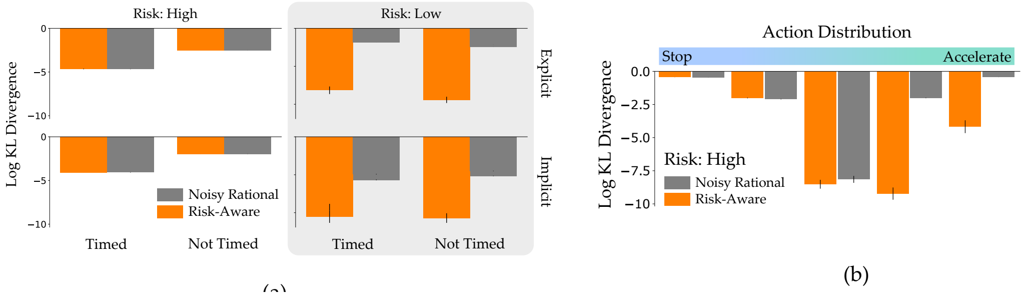


Figure 4: Model accuracy (lower is better) on (a) human and (b) simulated data.

Overall, the results from our autonomous driving user study support hypothesis H1. Autonomous cars leveraging a *Risk-Aware* model were able to understand and anticipate human drivers both in situations where the human is optimal *or* suboptimal, while the *Noisy Rational* model could not explain why the participants preferred to take a safer (but suboptimal) action.

Follow-up: Disentangling Risk and Suboptimal Decisions. After completing our user study, we performed a simulated experiment within the autonomous driving domain. Within this experiment, we *fixed* the probability that the light would turn red, and then *varied* the human driver's action distribution. When fixing the probability, we used the High risk scenario where the optimal decision was to *stop*. The purpose of this follow-up experiment was to make sure that our model can also explain suboptimally *aggressive* drivers, and to ensure that our results are not tied to the Low risk scenario. Our simulated results are displayed in Fig. 4. As before, when the human driver chose the optimal action, both *Noisy Rational* and *Risk-Aware* models were equally accurate. But when the human behaved aggressively—and tried to make the light—only the *Risk-Aware* autonomous car could anticipate their suboptimal behavior. These results suggest that the improved accuracy of the *Risk-Aware* model is tied to user *suboptimality*, and not to the particular type of risk.

5 Collaborative Cup Stacking

Within the autonomous driving user studies, we demonstrated that our *Risk-Aware* model enables agents to *accurately* anticipate their human partners. Next, we want to explore how our formalism leverages this accuracy to improve *safety* and *efficiency* during human-agent interaction. To test the usefulness of our model, we performed two user studies with a 7-DoF robotic arm (Fetch, Fetch Robotics) as our AI agent. In an **online** user study, we verify that the *Risk-Aware* model can accurately model humans in a collaborative setting. In an **in-person** user study, the agent leverages *Risk-Aware* and *Noisy Rational* models to anticipate human choices and plan trajectories that avoid interfering with the participant.

Experimental Setup. The collaborative cup stacking task is shown in Fig. 1. We placed five cups on the table between the person and agent. The agent knew the location and size of the cups *a priori*, and had learned motions to pick up and place these cups into a tower. However, the agent did not know which cups its human partner would pick up.

The human chooses their cups with two potential towers in mind: an *efficient but unstable* tower, which was more likely to fall, or a *inefficient but stable* tower, which required more effort to assemble. Users were awarded 20 points for building the stable tower (which never fell) and 105 for building the unstable tower (which collapsed 80% of the time). Because the expected utility of building the unstable tower was higher, our *Noisy Rational* baseline anticipated that participants would make the unstable tower.

Independent Variables. We varied the agent's model of its human partner with two levels: *Noisy Rational* and *Risk-Aware*. The *Risk-Aware* agent uses our formalism from Section 3 to anticipate how humans make decisions under uncertainty and risk.

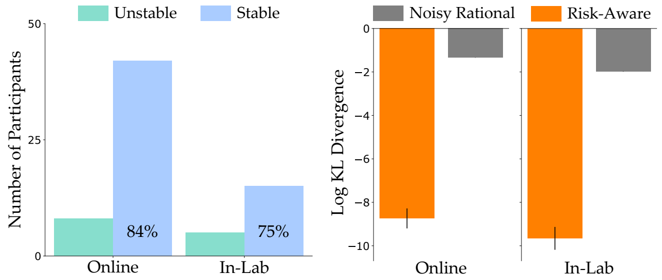


Figure 5: Results from online and in-person user studies during the collaborative cup stacking task.

5.1 Anticipating Collaborative Human Actions

Our **online** user study extended the results from the autonomous driving domain to this collaborative cup stacking task. We focused on how *accurately* the agent anticipated the participants' choices.

Hypothesis 2. Risk-Aware agents will better anticipate which tower the collaborative human user is attempting to build.

Results. Our results from the **online** user study are summarized in Fig. 5. During this scenario—where the human is collaborating with the agent—we observed a bias towards risk-averse behavior. Participants overwhelmingly preferred to build the stable tower (and take the guaranteed reward), even though this choice was suboptimal. Only the *Risk-Aware* agent was able to capture and predict this behavior: inspecting the right side of Fig. 5, we found a statistically significant improvement in model *accuracy* across the board ($t(59) = -21.1$, $p < .001$). Focusing only on the **online** users, the *log KL divergence* for *Risk-Aware* reached -8.7 ± 3.0, while *Noisy Rational* remained at -1.3 ± 0.01 ($t(29) = -13.1$, $p < .001$). Overall, these results match our findings from the autonomous driving domain, and support hypothesis H2.

5.2 Planning with Risk-Aware Human Models

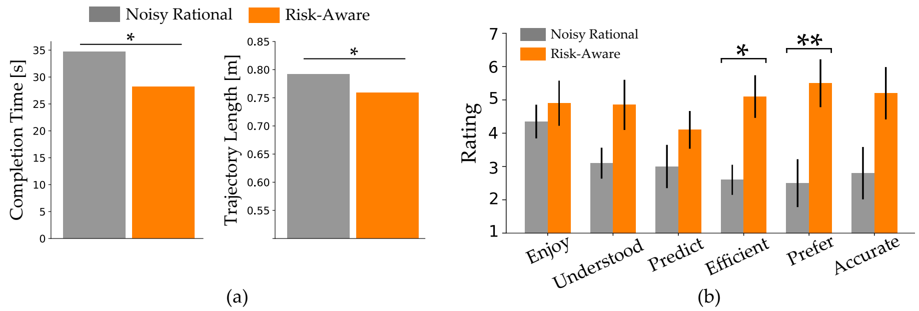
Having established that the *Risk-Aware* agent more accurately models the human's actions, we next explored whether this difference is *meaningful* in practice. We performed an **in-lab** user study comparing *Noisy Rational* and *Risk-Aware* collaborative agents. We focused on how agents can leverage the *Risk-Aware* human model to improve *safety* and *efficiency* during collaboration.

Dependent Measures. To test *efficiency*, we measured the time taken to build the tower (Completion Time). We also recorded the Cartesian distance that the agent's end-effector moved during the task (Trajectory Length). Because the agent had to replan longer trajectories when it interfered with the human, Trajectory Length was an indicator of *safety*.

After participants completed the task with each type of agent (*Noisy Rational* and *Risk-Aware*) we administered a 7-point Likert scale survey. Questions on the survey focused on four scales: how enjoyable the interaction was (Enjoy), how well the agent understood human behavior (Understood), how accurately the agent predicted which cups they would stack (Predict), and how efficient users perceived the agent to be (Efficient). We also asked participants which type of agent they would rather work with (Prefer) and which agent better anticipated their behavior (Accurate).

Hypothesis 3. Users interacting with the Risk-Aware agent will complete the task more safely and efficiently.
Hypothesis 4. Users will subjectively perceive the Risk-Aware agent as a better partner who accurately predicts their decisions and avoids grabbing their intended cup.

Results - Objective. We show example human and agent behavior during the **in-lab** collaborative cup stacking task in Fig. 9 in the Appendix. When modeling the human as *Noisy Rational*, the agent initially moved to grab the optimal cup and build the unstable tower. But in 75% of trials participants built the suboptimal but *stable* tower! Hence, the *Noisy Rational* agent often interfered with the human's actions. By contrast, the *Risk-Aware* agent was collaborative: it correctly predicted that the human would choose the stable tower, and reached for the cup that best helped build this tower. This led to improved safety and efficiency during interaction, as shown in Fig. 6(a). Users interacting



(a) (b)

Figure 6: (a) Objective results from our in-lab user study.(b) Subjective results from our in-person user study.

with the risk-aware agent completed the task in less time ($t(9) = 2.89$, $p < .05$), and the agent also traveled a shorter distance with less human interference ($t(9) = 2.24$, $p < .05$). These objective results support hypothesis H3.

Results - Subjective. We plot the user's responses to our 7-point surveys in Fig. 6(b). We first confirmed that each of our scales (*Enjoy*, *Understood*, etc.) was consistent, with a Cronbach's alpha > 0.9. We found that participants marginally preferred interacting with the *Risk-Aware* agent over the *Noisy Rational* one ($t(9) = 2.09$, $p < .07$). Participants also indicated that they felt that they completed the task more efficiently with the *Risk-Aware* agent ($t(9) = 3.01$, $p < .05$). The other scales favored *Risk-Aware*, but were not statistically significantly. Within their comments, participants noticed that the *Noisy Rational* agent clashed with their intention: for instance, *"it tried to pick up the cup I wanted to grab"*, and *"the agent picked the same action as me, which increased time"*. Overall, these subjective results partially support hypothesis H4.

6 Discussion and Conclusion

Many of today's agents model human partners as Noisly Rational. In real-life scenarios, however, humans must make choices subject to uncertainty and risk—and within these realistic settings, humans display a cognitive bias towards *suboptimal* behavior. We adopted Cumulative Prospect Theory from behavioral economics and formalized a human decision-making model so that agents can now anticipate suboptimal human behavior. Across autonomous driving and collaborative cup stacking environments, we found that our formalism better predicted user decisions under uncertainty. We also leveraged this prediction within the agent's planning framework to improve *safety* and *efficiency* during collaboration: our Risk-Aware agent interfered with the participants less and received higher subjective scores than the Noisy Rational baseline. We want to emphasize that this approach is *different from making agents robust* to human mistakes by *always* acting in a risk-averse way. Instead, when humans prefer to take safer but suboptimal actions, agents leveraging our formalism *understand* these conservative humans and increase overall team performance.

Limitations and Future Work. A strength and limitation of our approach is that the Risk-Aware model introduces additional parameters to the state-of-the-art Noisy Rational human model. With these additional parameters, agents are able to predict and plan around suboptimal human behavior; but if not enough data is available when the agent learns its human model, the agent could overfit. We point out that for all of the user studies we presented, the agents learned Noisy Rational and Risk-Aware models from the *same amount* of user data.

When learning and leveraging these models, the agent must also have access to real-world information. Specifically, the agent must know the rewards and probabilities associated with each outcome. We believe that agents can often obtain this information from experience: for example, in our collaborative cup stacking task, the agent can determine the likelihood of the unstable tower falling based on previous trials. Future work must consider situations where this information is not readily available, so that the agent can identify collaborative actions that are *robust* to errors or uncertainty in the human model.

References

[1] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[2] Andrew Gray, Yiqi Gao, J Karl Hedrick, and Francesco Borrelli. Robust predictive control for semi-autonomous vehicles with an uncertain driver model. In *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 208–213. IEEE, 2013.

[3] Vasumathi Raman, Alexandre Donzé, Dorsa Sadigh, Richard M Murray, and Sanjit A Seshia. Reactive synthesis from signal temporal logic specifications. In *Proceedings of the 18th international conference on hybrid systems: Computation and control*, pages 239–248. ACM, 2015.

[4] Michael P Vitus and Claire J Tomlin. A probabilistic approach to planning and control in autonomous urban driving. In *52nd IEEE Conference on Decision and Control*, pages 2459–2464. IEEE, 2013.

[5] Ram Vasudevan, Victor Shia, Yiqi Gao, Ricardo Cervera-Navarro, Ruzena Bajcsy, and Francesco Borrelli. Safe semi-autonomous control with enhanced driver modeling. In *2012 American Control Conference (ACC)*, pages 2896–2903. IEEE, 2012.

[6] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2000.

[7] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*. ACM, 2004.

[8] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 301–308. IEEE Press, 2013.

[9] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems*, 2017.

[10] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proc. AAAI*, pages 1433–1438, 2008.

[11] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*, pages 49–58, 2016.

[12] Malayandi Palan, Nicholas C Landolfi, Gleb Shevchuk, and Dorsa Sadigh. Learning reward functions by integrating human demonstrations and preferences. *arXiv preprint arXiv:1906.08928*, 2019.

[13] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

[14] Dylan P Losey, Krishnan Srinivasan, Ajay Mandlekar, Animesh Garg, and Dorsa Sadigh. Controlling assistive robots with learned latent actions. *arXiv preprint arXiv:1909.09674*, 2019.

[15] Muhammad Awais and Dominik Henrich. Human-robot collaboration by intention recognition using probabilistic state machines. In *19th International Workshop on Robotics in Alpe-Adria-Danube Region (RAAD 2010)*, pages 75–80. IEEE, 2010.

[16] Anca D Dragan and Siddhartha S Srinivasa. *Formalizing assistive teleoperation*. MIT Press, July, 2012.

[17] Dylan P Losey and Marcia K O'Malley. Enabling robots to infer how end-users teach and learn through human-robot interaction. *IEEE Robotics and Automation Letters*, 4(2):1956–1963, 2019.

[18] Brian D Ziebart, Nathan Ratliff, Garratt Gallagher, Christoph Mertz, Kevin Peterson, J Andrew Bagnell, Martial Hebert, Anind K Dey, and Siddhartha Srinivasa. Planning-based prediction for pedestrians. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3931–3936. IEEE, 2009.

[19] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, 36(5-7):618–634, 2017.

[20] Truong-Huy Dinh Nguyen, David Hsu, Wee-Sun Lee, Tze-Yun Leong, Leslie Pack Kaelbling, Tomas Lozano-Perez, and Andrew Haydn Grant. Capir: Collaborative action planning with intention recognition. In *Seventh Artificial Intelligence and Interactive Digital Entertainment Conference*, 2011.

[21] Haoyu Bai, Shaojun Cai, Nan Ye, David Hsu, and Wee Sun Lee. Intention-aware online pomdp planning for autonomous driving in a crowd. In *2015 ieee international conference on robotics and automation (icra)*, pages 454–460. IEEE, 2015.

[22] Dorsa Sadigh, S. Shankar Sastry, Sanjit A. Seshia, and Anca D. Dragan. Planning for autonomous cars that leverage effects on human actions. In *Proceedings of Robotics: Science and Systems (RSS)*, June 2016.

[23] Dorsa Sadigh, S Shankar Sastry, Sanjit A Seshia, and Anca Dragan. Information gathering actions over human internal state. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 66–73. IEEE, 2016.

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

[25] Andrea Thomaz, Guy Hoffman, Maya Cakmak, et al. Computational human-robot interaction. *Foundations and Trends® in Robotics*, 4(2-3):105–223, 2016.

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

[27] Erdem Biyik, Malayandi Palan, Nicholas C. Landolfi, Dylan P. Losey, and Dorsa Sadigh. Asking easy questions: A user-friendly approach to active reward learning. In *Proceedings of the 3rd Conference on Robot Learning (CoRL)*, October 2019.

[28] Erdem Biyik, Daniel A. Lazar, Dorsa Sadigh, and Ramtin Pedarsani. The green choice: Learning and influencing human decisions on shared roads. In *Proceedings of the 58th IEEE Conference on Decision and Control (CDC)*, December 2019.

[29] Chandrayee Basu, Erdem Biyik, Zhixun He, Mukesh Singhal, and Dorsa Sadigh. Active learning of reward dynamics from hierarchical queries. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, November 2019.

[30] Daniel S Brown, Wonjoon Goo, and Scott Niekum. Ranking-based reward extrapolation without rankings. *arXiv preprint arXiv:1907.03976*, 2019.

[31] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pages 2586–2591, 2007.

[32] Michael Bloem and Nicholas Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *53rd IEEE Conference on Decision and Control*, pages 4911–4916. IEEE, 2014.

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

[34] Stefania Pellegrinelli, Henny Admoni, Shervin Javdani, and Siddhartha Srinivasa. Human-robot shared workspace collaboration via hindsight optimization. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 831–838. IEEE, 2016.

[35] Lisa Ordonez and Lehman Benson III. Decisions under time pressure: How time constraint affects risky decision making. *Organizational Behavior and Human Decision Processes*, 71(2):121–140, 1997.

[36] Adele Diederich. Dynamic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology*, 41(3):260–274, 1997.

[37] Diana L Young, Adam S Goodie, Daniel B Hall, and Eric Wu. Decision making under time pressure, modeled in a prospect theory framework. *Organizational behavior and human decision processes*, 118(2):179–188, 2012.

[38] Espen Moen Eilertsen. Cumulative prospect theory and decision making under time pressure. Master's thesis, 2014.

[39] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

Appendix

Formalism

Risk-Aware Model. Without loss of generality, we assume that each of the K rewards are ordered in decreasing order, i.e. $R_H^{(i+1)}(a) \leq R_H^{(i)}(a)$ for all $i \in \{1, 2, \ldots, K-1\}$ and $a \in \mathcal{A}_H$. Then, the probability transformation is as follows:

$$\begin{aligned}
\pi\left(C(a_H)\right) &= \left(\pi^{+}(C(a_H)), \pi^{-}(C(a_H))\right)\\
\pi^{+}(C(a_H)) &= \left(w^{+}\left(p^{(1)}\right), w^{+}\left(p^{(1)}+p^{(2)}\right)-w^{+}\left(p^{(1)}\right), \ldots\right)\\
\pi^{-}(C(a_H)) &= \left(\ldots, w^{-}\left(p^{(K)}+p^{(K-1)}\right)-w^{-}\left(p^{(K)}\right), w^{-}\left(p^{(K)}\right)\right)
\end{aligned}$$

Finally, we normalize probabilities so that $\pi(C(a_H))$ sums to 1:

$$\overline{\pi}_j(C(a_H)) = \frac{\pi_j(C(a_H))}{\sum_{i=1}^{K}\pi_i(C(a_H)))}, \forall j \in \{1, 2, \ldots, K\}$$

When $\gamma, \delta \in (0, 1)$, the probability transformations capture biases humans are reported to have [13] by overweighting smaller probabilities and underweighting larger probabilities.

Based on these two transformations, we now extend the human decision making model with the Risk-Aware model:

$$\begin{aligned}
R_H^{\text{CPT}}(a_H) &= \overline{\pi}_1(C(a_H)) \cdot v\left(R_H^{(1)}(a_H)\right) + \ldots + \overline{\pi}_K(C(a_H)) \cdot v\left(R_H^{(K)}(a_H)\right)\\
P(a_H) &= \frac{\exp\left(\theta \cdot R_H^{\text{CPT}}(a_H)\right)}{\sum_{a \in \mathcal{A}_H}\exp\left(\theta \cdot R_H^{\text{CPT}}(a)\right)}\ .
\end{aligned} \tag{2}$$

In contrast to the Noisy Rational model, the Risk-Aware model's expressiveness allows it to model *both* optimal and suboptimal human decisions by assigning larger likelihoods to those actions.

Formal Model of Interaction. We describe how we can integrate them into a partially observable Markov decision process (POMDP) formulation of human-agent interaction. We model the world where both the human and the agent take actions as a POMDP, which we denote with a tuple $\langle \mathcal{S}, \mathcal{O}, O, \mathcal{A}_H, \mathcal{A}_R, T, r_H, r_R \rangle$. $\mathcal{S}$ is the finite set of states; $\mathcal{O}$ is the set of observations; $O : \mathcal{S} \to \mathcal{O}$ defines the shared observation mapping; $\mathcal{A}_H$ and $\mathcal{A}_R$ are the finite action sets for the human and the agent, respectively; $T : \mathcal{S} \times \mathcal{A}_H \times \mathcal{A}_R \times \mathcal{S} \to [0, 1]$ is the transition distribution. r_H and r_R are the reward functions that depend on the state, the actions and the next state. In this POMDP, we assume the agents act simultaneously. Having a first-order ToM, the human tries to optimize her own cumulative reward given an action distribution for the agent, $P(a_R|s)$. The human value function $V_H(s)$ can then be defined using the following Bellman update:

$$V_H(s) = \max_{a_H \in \mathcal{A}_H} \mathbb{E}_{a_R|s}\mathbb{E}_{s'|s,a_H,a_R}\left[R_H(s, a_H, a_R, s') + \gamma V_H(s')\right]\ .$$

Table 1: Autonomous Driving. Users were given different amounts of *information* about the likelihood that the light would turn red. Under *risk*, we list two tested probabilities of the light turning red.

Information	***Time*** [b]
None: *With some probability*	**Timed**: 8 s [t]
the light will turn red.	**Not Timed**: no limit
Explicit: *There is a 5% chance*	[t]
the light will turn red.	***Risk***
Implicit: *Of the previous 380 cars*	**High**: 95% [t]
that decided to accelerate,	**Low**: 5%
the light turned red for 19 cars.	[b]

We then use the fact that

$$P(s, s', a_R \mid o, a_H) = P(s \mid o) \cdot P(a_R \mid s) \cdot P(s' \mid s, a_H, a_R)$$

to construct a set $C(a_H)$ for the current observation o that consists of the pairs $(P(s, s', a_R \mid o, a_H), V_H(s'))$ for varying s, s', and a_R. When modeling the human as zeroth-order ToM, $P(a_R \mid s)$ will simply be a uniform distribution.

Having constructed $C(a_H)$, we can define the human's utility function for different values of s, and s'. The utility functions for both Noisy Rational and Risk-Aware models are defined as follows:

Noisy Rational:

$$R_H(o, a_H) = \sum_{s,s' \in \mathcal{S}} \sum_{a_R \in \mathcal{A}_R} P(s, s', a_R \mid o, a_H) \cdot V_H(s') \; .$$

Risk-Aware:

$$R_H^{CPT}(o, a_H) = \sum_{s,s' \in \mathcal{S}} \sum_{a_R \in \mathcal{A}_R} \bar{\pi}_i(P(s, s', a_R \mid o, a_H)) \cdot v\left(V_H(s')\right) \;\; ,$$

where the index i corresponds to the event that leads to s' from s with a_R, a_H. An optimal human would always pick the action a_H that maximizes $\mathbb{E}_{s|o}\mathbb{E}_{a_R|s}\mathbb{E}_{s'|s,a_H,a_R}\left[V_H(s')\right]$. The agent can obtain $P(a_H|o)$ using Eq. 1, Eq. 2 and use it to maximize its own cumulative reward.

Autonomous Driving Experiment Details

Participants and Procedure. We conducted a within-subjects study on Amazon Mechanical Turk and recruited 30 participants. All participants had at least a 95% approval rating and were from the United States. After providing informed consent, participants were first given a high-level description of the autonomous driving task and were shown the example from Fig. 1. In subsequent questions, participants were asked to indicate whether they would accelerate or stop. We presented the Timed questions first and the Not Timed questions second. For each set of Timed and Not Timed questions, we presented questions in the order of their informativeness from None to Explicit. The risk levels were presented in random order.

Online Collaborative Cup Stacking Experiment Details

Participants and Procedure. We recruited 14 Stanford affiliates and 36 Amazon Mechanical Turkers for a total of 50 users (32% Female, median age: 33). Participants from Amazon Mechanical Turk had at least a 95% approval rating and were from the United States. After providing informed consent, each of our users answered survey questions about whether they would collaborate with the agent to build the *efficient but unstable* tower, or the *inefficient but stable* tower. Before users made their choice, we explicitly provided the rewards associated with each tower, and implicitly gave the probability of the tower collapsing. To implicitly convey the probabilities, we showed videos of humans working with the agent to make stable and unstable towers: all five videos with the stable tower showed successful trials, while only one of the five videos with the unstable tower displayed success. After watching these videos and considering the rewards, participants chose their preferred tower type.

Dependent Measures. We aggregated the participants' decisions to find their *action distribution* over stable and unstable towers. We fit *Noisy Rational* and *Risk-Aware* models to this action distribution, and reported the *log KL divergence* between the actual and model tower choices.

Offline Collaborative Cup Stacking Experiment Details

Participants and Procedure. Ten members of the Stanford University community (2 female, ages $20-36$) provided informed consent and participated in this study. Six of these ten had prior experience interacting with the Fetch robot. We used the same experimental setup, rewards, and probabilities described at the beginning of the section. Participants were encouraged to build towers to maximize the total number of points that they earned.

Each participant had ten familiarization trials to practice building towers with the agent. During these trials, users learned about the probabilities of each type of tower collapsing from experience. In half of the familiarization trials, the agent modeled the human with the *Noisy Rational* model, and in the rest the agent used the *Risk-Aware* model; we randomly interspersed trials with each model. After the ten familiarization trials, users built the tower once with *Noisy Rational* and once with *Risk-Aware*: we recorded their choices and the agent's performance during these final trials. The presentation order for these final two trials was counterbalanced.

Complex POMDP settings

To investigate how well *Risk-Aware* and *Noisy Rational* model humans in more complex POMDP settings, we designed two different maze games. Each game consists of two 17-by-15 grids and these two grids have the exact same structure of walls, which are visible to the player. In each grid, there is one *start* and two *goal* squares. Players start from the same square, and reach either of the goals. Each square in the grids has an associated reward, which the player can also observe. The partial observability comes from the rule that the player does not exactly know which grid she is actually playing at. While she is in the first grid with 95% probability, there is a 5% chance that she might be playing in the second grid. We visualize the grids for both games in Fig. 7, and also attach the full mazes in the supplementary material. We restricted the number of moves in each game such that the player has to go to the goals with the minimum possible number of moves. Finally, we enforced a time limit of 2 minutes per game.

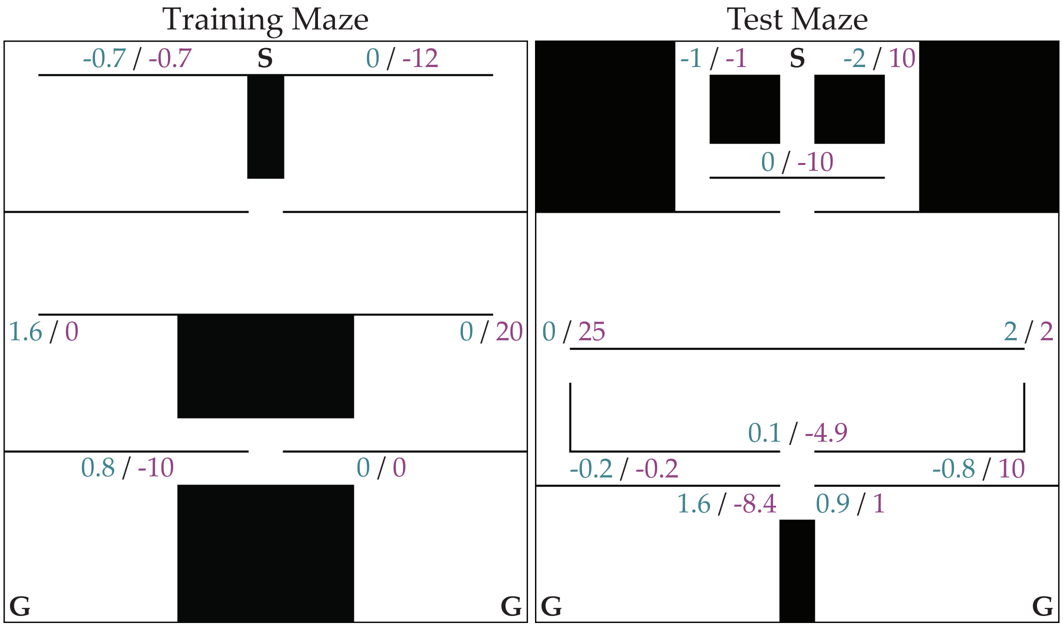


Figure 7: Summaries of two games. For each game, we have a maze. The values written on the mazes represent how much reward players can collect by entering those roads. The first numbers in each pair correspond to the 95% grid, and the second one to the 5% grid.

We investigate the effect of both risk and time constraints via this experiment. While it is technically possible for the players to compute the optimal trajectory that leads to the highest expected reward, time limitation makes it very challenging, and humans resort to rough calculations and heuristics.

Moreover, we designed the mazes such that humans can get high rewards or penalties if they are in the low-probability (5%) grid. This helps us investigate when humans become risk-seeking or risk-averse.

We recruited 17 users (4 female, 13 male, median age 23), who played both games. We used one game (two grids) to fit the model parameters independently for each user, and the other game (other two grids) to evaluate how well the models can explain the human behavior. As the human actions depend not only the immediate rewards, but also the future rewards, we ran value iteration over the grids and used the values to fit the models as we described in Sec. 3. We again employed Metropolis-Hastings to sample model parameters, and recorded the mean of the samples.

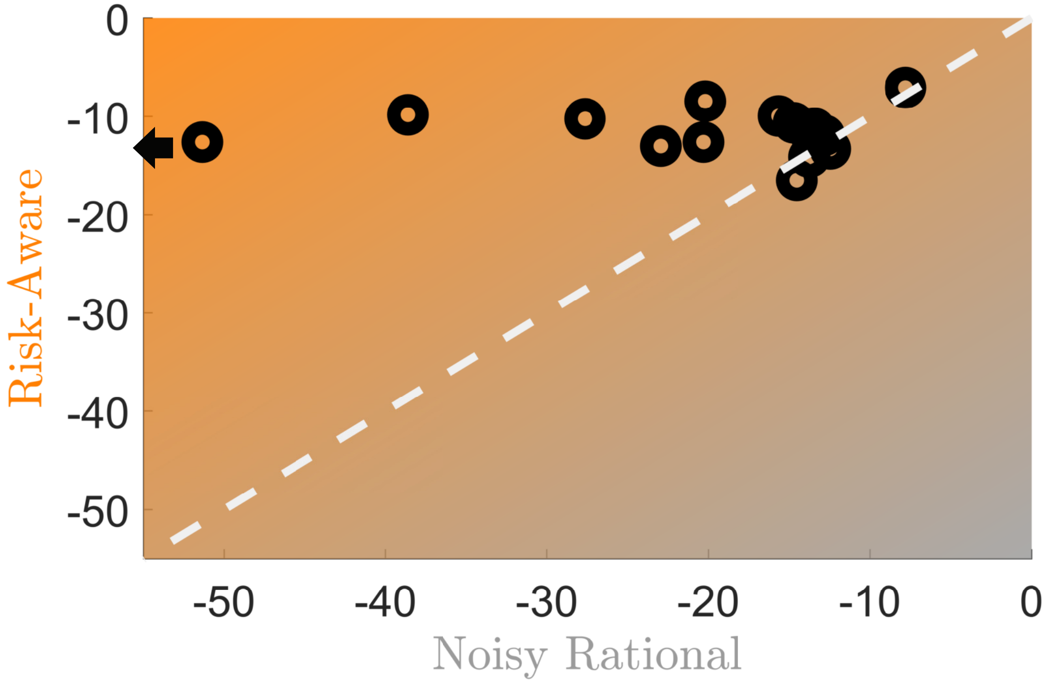


Figure 8: Log-Likelihood values by Risk-Aware and Noisy Rational models. One outlier point is excluded from the plot and is shown with an arrow.

Figure 8 shows the log-likelihoods for each individual user for Risk-Aware and Noisy Rational models. Overall, Risk-Aware explains the test trajectories better. The difference is statistically significant (paired t-test, $p < 0.05$). In many cases, we have seen risk-averse and risk-seeking behavior from people. For example, 12 out of of 17 users chose the risk-seeking action in the test maze by trying to get 25 reward with probability 5% instead of getting 2 with 100% probability. Similarly, 15 out of 17 users choose to guarantee 0.9 reward and gain 0.1 more with 5% probability instead of guaranteeing 1.6 reward and losing 10 with 5% probability. This is an example of suboptimal risk-averse action.

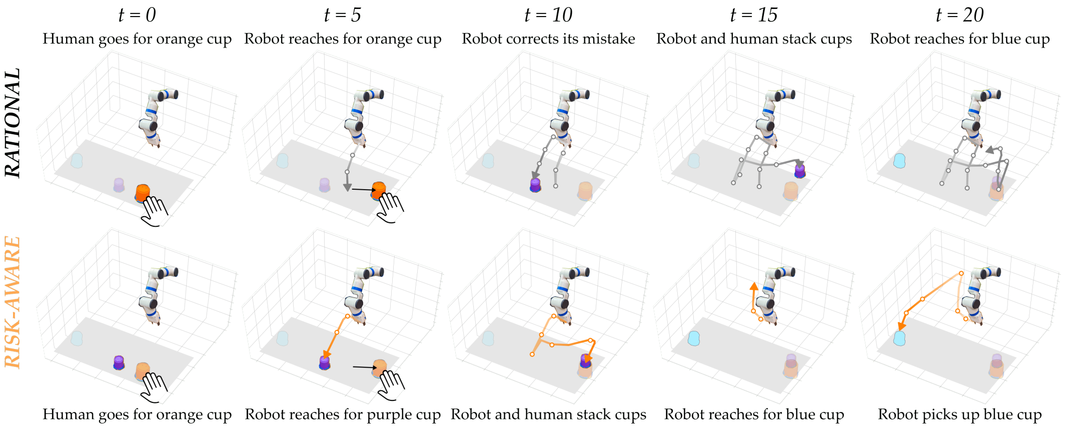

t = 0
Human goes for orange cup
t = 5
Robot reaches for orange cup
t = 10
Robot corrects its mistake
t = 15
Robot and human stack cups
t = 20
Robot reaches for blue cup
RATIONAL
RISK-AWARE
Human goes for orange cup
Robot reaches for purple cup
Robot and human stack cups
Robot reaches for blue cup
Robot picks up blue cup

Figure 9: Example agent and human behavior during the collaborative cup stacking user study. At the start of the task, the human reaches for the orange cup (the first step towards a stable tower). When the agent models the human as a *Noisy Rational* partner (top row), it incorrectly anticipates that the human will build the optimal but unstable tower; this leads to interference, replanning, and a delay. The agent leveraging our *Risk-Aware* formalism (bottom row) understands that real decisions are influenced by uncertainty and risk, and correctly predicts that the human wants to build a stable tower. This results in safer and more efficient interaction, leading to faster tower construction.