Bayesian Inference for Human-Robot Coordination in Parallel Play

Shray Bansal[1]*, **Jin Xu**[1], **Ayanna Howard**[2], **Charles Isbell**[1]
[1] Georgia Institute of Technology
[2] Ohio State University

Abstract

We consider shared workspace scenarios with humans and robots acting to achieve independent goals, termed as parallel play. We model these as general-sum games and construct a framework that utilizes the Nash equilibrium solution concept to consider the interactive effect of both agents while planning. We find multiple Pareto-optimal equilibria in these tasks. We hypothesize that people act by choosing an equilibrium based on social norms and their personalities. To enable coordination, we infer the equilibrium online using a probabilistic model that includes these two factors and use it to select the robot's action. We apply our approach to a close-proximity pick-and-place task involving a robot and a simulated human with three potential behaviors - defensive, selfish, and norm-following. We showed that using a Bayesian approach to infer the equilibrium enables the robot to complete the task with less than half the number of collisions while also reducing the task execution time as compared to the best baseline. We also performed a study with human participants interacting either with other humans or with different robot agents and observed that our proposed approach performs similar to human-human parallel play interactions.

1 Introduction

People often perform activities in shared spaces with other people achieving their own individual goals. This includes driving to work while sharing the road with other cars, navigating around other shoppers when pushing a cart in a grocery store, and sharing counter-space and utensils in a kitchen. Although these situations are neither purely collaborative nor competitive, the actions of other participants have bearing on each person's own success or failure. We refer to these activities as *parallel play*, related to its psychology namesake that refers to activities in early social development, where children play *besides* instead of *with*, other children [17, 16]. In the Human-Robot Interaction (HRI) context, we define *parallel play* to refer to those activities where people and robots have separate individual goals but interact due to shared space. We aim to derive a framework that helps a robot plan effectively for parallel play with human participants, and apply it to a close-proximity pick-and-place scenario between a robot and a human.

Planning a robot's action in HRI usually involves considering the robot's goals as well as predictions of future human actions [20, 1, 9]. When working with others, people are often considerate of their intents and beliefs due to Theory-of-Mind [18, 4], and so, the human's action is influenced by their predicted plans of the other participant's, including the robot. Modeling this cyclical-dependence, of the human's predicted plan on the robot's and vice-versa, is important for accurately representing the interaction dynamics in HRI.

*sbansal34@gatech.edu

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

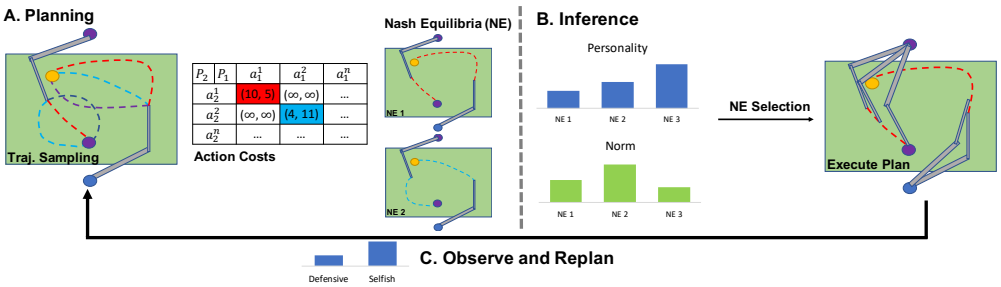


Figure 1: Overview. Our approach has three main stages. (A) In the planning stage, we sample goal-reaching trajectories for both agents and compute their costs and use it to find the Nash equilibria (NE). (B) In the inference stage, we compute the probability for each equilibrium based on its adherence to the social norm, as well as the inferred personality of the other agent. We select the most-likely equilibrium under this distribution and partially execute the action. In (C), we observe the states of both agents, infer the personality and start replanning. This process repeats until the agent has reached its goal.

Game Theory provides us tools to model this inter-dependence of rational interacting agents. A (pure) Nash equilibrium (NE) is a set of actions, one for each agent in the game, which is optimal, assuming the actions of others remain fixed [11]. A Nash equilibrium implicitly captures the inter-dependence between agents, and our approach plans by finding equilibria strategies to enable better coordination.

Although humans have been shown to play to Nash equilibrium [12], we find that multiple equilibria can exist in a game and it is not clear how the robot should choose between them. Figure 1 shows an example of two equilibria in a pick-and-place scenario, where each favors a different agent by allowing them to reach their goal first. A collision is likely if both agents choose an equilibrium that favors them, highlighting the importance of coordination. Humans can coordinate social behavior in non-competitive games by learning and following social norms [8]. Here, a norm refers to a set of abstract instructions that agents follow, and expect others to follow. In Figure 1, a norm might favor solutions that allow the agent closer to their goal to reach for it first and, if followed by both agents, would lead to coordination. However, sometimes agents can ignore the norm in favor of their personal preferences. For example, a selfish agent might ignore the norm and expect the other to always yield to them when reaching for an object. Coordinating with such agents requires the ability to infer this preference from experience.

We design a framework that finds Nash equilibria for *parallel play* tasks; it models the strategy for choosing equilibria as a distribution composed of two aspects - (1) a domain-specific social norm, designed by an expert apriori and (2) an agent-specific individual preference, inferred online during the interaction. We hypothesize that this framework would lead to better coordination with humans in performing in such tasks, due to its modeling of the decision-making coupling between agents, as well as, its combination of expert knowledge with online adaptation. To validate, we apply this to a close-proximity pick-and-place task, designed to be similar to HRI tasks used to study team coordination and fluency [6, 13] with a simulated human. Our results show that this framework is able to shorten task execution time while also reducing the number of human collisions by half as compared to the best baseline when interacting with a simulated human with 3 potential personalities. We make the following contributions:

1. Introduce a novel framework that models norm-following social behavior and personality-based likelihood inference to interactive-planning with humans in parallel play activities. We do this by first computing the Nash equilibria and then using this framework to find a distribution over them.

2. Design task and metrics to benchmark the performance of interactive-planning algorithms for *parallel play*. This includes 3 baselines for simulating distinct human personalities that have similarity to human performance of the task, which enables testing adaptability of the algorithms to different human behavior.

2 Related Work

There has been extensive work in HRI for planning a robot to work on tasks around humans in domains like parts assembly, motion planning, and autonomous driving. Planning around people generally involves two aspects, predicting the human's behavior, and finding robot actions that achieve its goal in the presence of the human.

Human modeling. Prior work has placed emphasis on accurately modeling the human's rational goal-driven behavior. This includes learning the human's preferences to predict low-level trajectories [25] or modeling their high-level decision-making using a probabilistic task model [7]. A common approach to planning in this setting is by finding a best-response to the predicted human behavior. This approach works very well in scenarios where the robot assumes an assistive role like parts delivery [7]. An inherent assumption of this approach is that, although the human's plan depends on the situation but not on the robot's behavior. This leads to the robot choosing overly conservative behaviors which can lead it to freeze when trying to navigate crowds [23] or fail to merge in traffic.

Mutually adaptive planning. Recent work has addressed this by considering the human's influence-ability as well [24, 20, 5]. Similar to us, they also utilize game-theoretic tools to model this cyclical influence. For instance, [24] modeled robot navigation as a dynamic general-sum game and computed a Nash equilibrium to effectively plan the robot's trajectory among a crowd of pedestrians. [6] utilized the Nash equilibrium to find an order for object pick-up in a close-proximity pick-and-place task similar to ours; they found that considering the mutual adaptation allowed their framework to improve safety as well as human subjective preference. Our approach also uses the Nash equilibrium to plan goal-driven actions for the robot that consider the mutual adaptability between the two agents. However, our approach includes a strategy for selecting an equilibrium in case multiple are present, while others either have not mentioned their technique for doing so [6] or only find one equilibrium due to their problem structure [20, 5].

Online model inference. Although, human behavior is heterogeneous, prior work tends to use a single model to describe all human users with some exceptions. For *e.g.*, [15] groups people by their adaptability to the robot and this is used to guide the adaptable users to better collaborative solutions in a collaborative task, and [19] uses information-gathering actions to infer whether a driver is paying attention to the task and use this knowledge to make better plans for the robot. We use a latent variable to cluster human personality and infer this online for the interacting human participant. Similar to us, [21] developed an approach for autonomous driving which relies on computing the Nash equilibrium with the interacting human driver. This approach includes online inference of a parameter in the human cost function, that represents their degree of selfishness or altruism. While they use a local approximation for computing the Nash equilibrium for a continuous action space, and so only find one equilibrium, we find multiple in a discrete higher-level action space. Their results showed that inferring the human's preference online reduced error in human prediction and helped the robot achieve better coordination.

Coordination. The importance of coordinating with the human in non-competitive games was also highlighted in [2] where they learned human models and used it to trained reinforcement learning agents that achieved superior performance to those trained with self-play. Better coordination for collaborative tasks was also was achieved in [8] by using learning task-based norms for social behavior.

3 Interactive Planning as a Game

We model the multi-agent interactive planning task as a non-cooperative game represented as tuple G, $G = (P, A, c)$ [11]. Here, $P = \{P_1, ..., P_N\}$ is a finite set of N players, $A = A_1 \times ... \times A_N$ where A_i is the set of actions available to player i. We refer to the set of concurrent actions, one for each agent, as an action profile, a, $a = (a_1, ..., a_N)$. We define a cost representing the unfavorability of an action profile for agent i as $c_i : A \mapsto \mathbb{R}$ and $c = (c_1, ..., c_N)$ includes the mapping for all agents.

In our scenario, each agent P_i is a robot arm, each action set A_i is a set of goal-driven trajectories, each trajectory is a sequence of joint-space positions and velocities sampled using a planner, and the cost $c_i(a)$ encourages each robot to minimize task completion time while avoiding collisions with other agents. The goal for an agent i is to take an action $a_i \in A_i$ in profile a, which minimizes its cost. However, its cost depends upon the actions chosen by the other agents in the profile a. We

assume that all agents are rational and have Theory-of-Mind, *i.e.*, they choose actions to minimize their own cost and are aware of the states and goals of the other agents. These assumptions allow us to utilize the Nash Equilibrium (NE) solution concept for this game. An action profile is a NE, for a single-stage game, if no agent has an incentive to choose a different action for themselves given that all the other actions are fixed.

$$\mathrm{a}_i^* \in \arg\min_{a_i} c_i(\mathrm{a}_1^*, ., a_i, ., \mathrm{a}_N^*) \quad \forall i \in N. \tag{1}$$

Although generally, only one (mixed) equilibrium is guaranteed to exist for a game [11], in our problem, one pure equilibrium is always present and we find that multiple equilibria are frequently present. For the planning agent, some of these equilibria can be eliminated for being Pareto-sub-optimal, *i.e.*, worse for all agents. For example, Nash profile, a^{*1} Pareto-dominates a profile a^{*2}, if $c_i(\mathrm{a}^{*1}) < c_i(\mathrm{a}^{*2}) \forall i \in N$. Next, we present an approach that can help the agent select an action by choosing between Pareto-optimal equilibria.

4 Equilibrium Selection Strategy

Our strategy chooses between equilibria using two aspects of human social behavior, norm-following and personality-adaptation. We model the distribution over equilibria as a product of its probability under the norm, p_n and its probability given the predisposed personality, p_α,

$$p(a) = p_n(a) p_\alpha(a). \tag{2}$$

Here, and in the rest of this section, a refers to a NE action profile. Next, we explain the norm for this problem and how we use observations to update the personality distribution.

Social Norm. Similar to [8], we define a norm to be a set of, situation-dependent, abstract, instructions that agents follow with the expectation that others will follow them as well. They help agents coordinate in the absence of prior knowledge of the agents they are interacting with. For example, a first-come-first-leave norm can help decide how cars navigate a four-way stop. Here, we model it as a probability distribution over NE. Different games will have different norms and the choice of a norm should be based on expert knowledge or learned from data. For our problem, we use a simple min-norm, that prioritizes the equilibrium which achieves minimum cost for any of the agents,

$$p_n(a) \propto e^{-\lambda_n min_i(c_i(a))}, \tag{3}$$

where λ_n is a parameter of the exponential distribution that we set. This norm it encourages the agent with the shortest unobstructed path to its goal to reach it first.

Online Preference Estimation

Although norms can help in coordination, people sometimes have strong preferences that guide them towards certain equilibria regardless of the norms. For example, an aggressive driver may decide to cross an intersection first, despite the norm, expecting the other drivers to adapt their strategy. We model this as a distribution over equilibria, inferred at time t using the history H_t of the past interaction,

$$p_\alpha^t(a) = p(a|H_t). \tag{4}$$

Here, H_t refers to the history of the interaction, *i.e.*, $H_t = \{(\{s_{i\in N}^0\}), ..., (\{s_{i\in N}^{t-1}\})\}$, and s_i^t is the state of agent i at time t. We set it to the uniform distribution at the start,

$$p_\alpha^{t=0}(a) = \text{uniform}(a) \quad \forall a. \tag{5}$$

We define an exponential distribution on the distance between a past trajectory, H, to an action profile, a,

$$p(a|H) \propto e^{-\lambda_\alpha f_{dist}(a,H)}, \tag{6}$$

Here, $f_{dist}(a, H)$ is defined as the Euclidean distance between the sequence of states in H to those in a and λ_α is a parameter of the exponential distribution. From Eq. 6, we know $p(a^0|H^t)$, however,

we would like to find $p(a^t|H^t)$. For this, we first define a latent variable θ. θ refers to the intrinsic personality of an agent and so is assumed to remain constant for every agent during the interaction. Now, we derive $p(a^t|H^t)$ by using the personality, θ, as follows,

$$p(a^t|H^t) = \sum_{\theta} p(a^t, \theta|H^t),$$
$$p(a^t|H^t) = \sum_{\theta} p(\theta|H^t)p(a^t|\theta, H^t).$$

We assume that the personality, θ, encodes the information required for predicting the agent's next action, which makes a^t conditionally independent of H^t given θ. So,

$$p(a^t|H^t) = \sum_{\theta} p(\theta|H^t)p(a^t|\theta). \tag{7}$$

We define θ such that each action profile a that belongs to a personality is equally likely to be chosen,

$$p(a|\theta) = \frac{\mathbb{1}_\theta(a)}{\sum_{a'} \mathbb{1}_\theta(a')}. \tag{8}$$

Next, we find $p(\theta|H^t)$ by taking its joint distribution with a^0 and marginalizing it out,

$$p(\theta|H^t) = \sum_{a^0} p(\theta, a^0|H^t),$$
$$p(\theta|H^t) = \sum_{a^0} p(a^0|H^t)p(\theta|a^0, H^t).$$

From the conditional independence between θ and H^t given a^0, we get,

$$p(\theta|H^t) = \sum_{a^0} p(a^0|H^t)p(\theta|a^0). \tag{9}$$

Since we assume a uniform prior on θ, $p(\theta)$ is a constant and so,

$$p(\theta|a^0) = \frac{p(\theta)p(a^0|\theta)}{p(\theta)\sum_{\theta'} p(a^0|\theta')} = \frac{p(a^0|\theta)}{\sum_{\theta'} p(a^0|\theta')} \tag{10}$$

We use $p(\theta|a^0)$ and $p(a^0|H^t)$ (Eq. 6) to get $p(\theta|H^t)$ in Eq. 9. This allows us to find $p(a^t|H^t)$ from Eq. 9 by using $p(a^t|\theta)$ from Eq. 8, which gives us $p_\alpha(a^t)$ in Eq. 4

5 Pick-Place Task

The pick-and-place task involves two 2-dof articulated arms moving on a 2D surface with the goal to pick up their designated object, by moving their end-effector close to it for grasping, and placing it, by bringing the grasped object to the destination area. The scenario is depicted in Figure 2, where the arm with a red base was controlled by our approach, and the other one was either simulated as a human or controlled by a human. Henceforth, the former will be referred to as the robot and the latter as the human.

Action Planning. To plan for this task, we first sample $k-1$ plans for each agent in configuration space using a Rapidly-exploring Random Tree (RRT) [10] and add a static plan where the agent does not move, $A_i = \{\tau_{j\in k}\}$. We use them to generate an action set by taking the outer product of the trajectories for each agent, $A = A_1 \times ... \times A_N$. We compute a cost for each action profile by simulating it and use Eq. 1 to find the Nash equilibria. We choose the NE profile a that maximizes the distribution $p(a)$ from Eq. 2, and select the agent's action from a. This action, along with the action taken by the human, is executed until a collision is detected or if the time before replanning is reached. After this, we update the history H^t and replan. This process continues until the robot completes the task and is depicted in Figure 3.

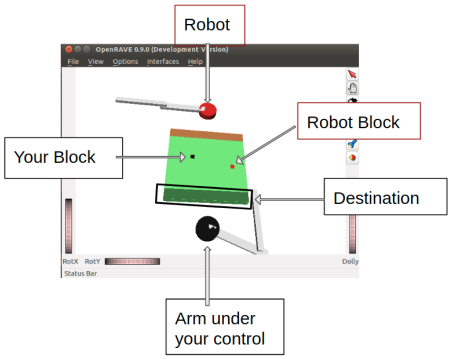


Figure 2: Task setup.

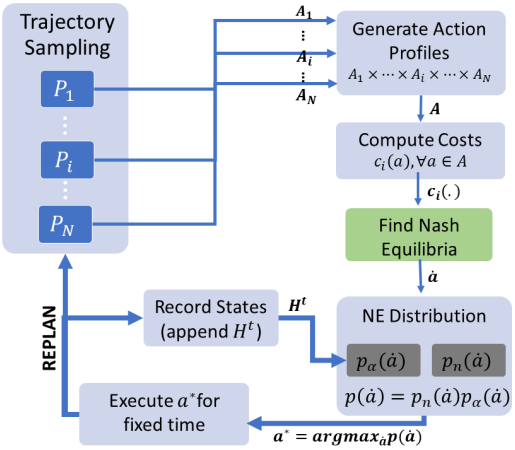


Figure 3: Framework.

Task Costs. We define a simple cost function that encourages the robot to complete the task quickly and avoid collisions. The cost of an action profile, $a = (a_R, a_H)$, where a_R, a_H, are the robot and human actions respectively, is the trajectory duration if it is successful in reaching the goal and infinity if it leads to a collision. We sample goal-reaching trajectories for each agent that are independent of the goal and state of other agents. We assume that the human is also goal-driven and collision-avoidant and so assume an analogous cost function where their cost depends on the human task completion time. Under these conditions, one pure Nash equilibrium is guaranteed as long as there exists a trajectory for the robot to reach the goal. For instance, say that we select an action profile with the robot's action being the shortest trajectory to its goal and the human's action as the shortest trajectory that does not collide with the robot's plan (including the static action). This profile will be a Nash equilibrium since neither agent has an incentive to modify their actions. For the robot, the action is optimal, and, for the human, this action is optimal assuming the robot's action as fixed due to the infinite cost of a collision.

Baselines. We define three baselines to compare with our approach. **Defensive:** The robot chooses an action assuming that the human wants to maximize the robot's cost while still achieving its goal leading to a maximin formulation. Thus, the agent will act defensively by preferring actions that do not lead to collision with the sampled human trajectories and will often lead it to wait for the human to complete their task.

$$a_R = \arg\min_{a_R \in A_R} \quad max_{a_H \in A_H} c_R(a_R, a_H). \tag{11}$$

Selfish: Chooses an equilibrium profile that minimizes the robot's cost. This strategy selects a trajectory that reaches the goal as quickly as possible assuming the other agent avoids collision.

$$a_R \in \mathrm{a}^*, \mathrm{a}^* = \arg\min_{\mathrm{a}^*} c_R(\mathrm{a}^*). \tag{12}$$

Norm-Nash: Chooses an equilibrium profile that maximizes the norm distribution p_n from Eq. 3. This leads to behavior that encourages the agent closer to their goal to reach them first. While the first two will lead to somewhat fixed behaviors, this strategy adapts to goal-achievability, which varies across tasks and also in state evolution within the same interaction.

$$a_R \in \mathrm{a}^*, \mathrm{a}^* = \arg\max_{\mathrm{a}^*} p_n(\mathrm{a}^*). \tag{13}$$

Implementation Details. A 3D simulation environment was created using the open-source Open Robotics Automation Virtual Environment (OpenRAVE) [3] with a time step of 0.1 seconds. The action-set, A_i was sampled using an RRT planner from the open-source Open Motion Planning Library (OMPL) [22]. We sampled $k = 8$ plans for each agent when planning and compute the cost as an $k \times k$ table by simulating the actions using OpenRAVE with a time-step of 0.8; we increased the time-step here to allow for fast computation of the nash solutions. Parameter λ_n of the norm distribution (Eq. 3) was set to 50. We set two personality types and use a binary latent variable $\theta = \{0, 1\}$. $\theta = 0$ selects equilibrium profiles a that favor agent 1, $c_1(a) < c_2(a)$, and $\theta = 1$

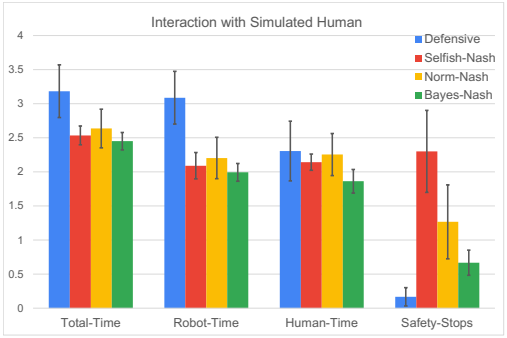


Figure 4: Results from the simulated human study.

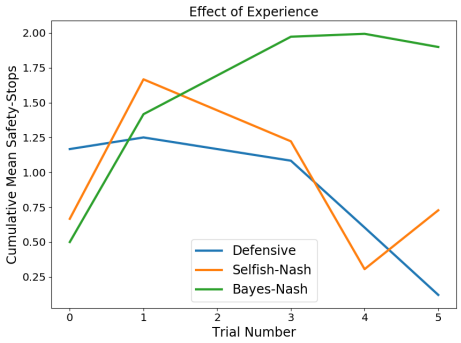


Figure 5: Learning effects.

selects equilibrium profiles a that favor agent 2, $c_2(a) < c_1(a)$. We set $\lambda_\alpha = 10$ in the personality distribution (Eq. 6).

6 Simulated Human Study

We simulate human behavior to create a controlled setting for our first experiment.

Simulated Human. We defined three human behaviors using the baselines: (1) **Defensive**, (2) **Selfish-Nash**, and (3) **Norm-Nash**. We chose the first two behaviors because of their clear intuitive distinctness and combine it with the third in accordance with our expectation that people also follow social norms.

Metrics. We measured the following task performance metrics: total task completion time and task time for each agent; we also counted safety stops, which are the number of times the simulation stopped the agents to avoid an impending collision. To keep these measures independent, we did not have any time penalty for a safety stop.

Results. To test how each algorithm fares with the different behaviors, we randomly pair the robot with one of these simulated human behaviors with random object locations for 30 trials. The averaged metrics for the three baselines and our proposed approach, Bayes-Nash, are presented in Figure 4. As expected, the Defensive robot was the safest, but its safe behavior also caused the highest robot and total task completion times. The Selfish-Nash was significantly faster than the Defensive robot but also led to the highest safety stops. Both Norm-Nash and Bayes-Nash performed comparably in time to Selfish-Nash but the Bayes-Nash was marginally faster. They were both significantly safer than Selfish-Nash and Bayes-Nash also had the fewest safety stops of the two.

Analysis. These results illustrate the trade-off between safety and efficiency present in the task, where the Defensive and Selfish-Nash agents sit at opposite extremes. Norm-Nash and Bayes-Nash are able to better trade-off these metrics due to their capability to adapt to the situation and the (simulated) human, respectively.

7 Human-Robot Study

Experimental Design. In order to validate the proposed approach, we design a within-subject human study. We examined the effects of three planning algorithms on their interaction with a human user and counter-balanced their order. Participants were asked to control a robot arm in simulation using a gamepad controller. The gamepad controller allows users to move the robot arm in eight different directions at 10 Hz. We used the same manipulation task as the previous two experiments, including keeping the object locations the same. Participants were informed that they might interact with different robots but not what these types were. There were three rounds of the task and each round involved six trials. In each round, the robot used one of the following conditions: (1) **Defensive**, (2) **Selfish-Nash**, and (3) **Bayes-Nash**. We used the same metrics as before.

Procedure. After giving informed consent, participants went through an overview of the experimental procedures. The study started with a pre-survey to collect demographic information. Then participants entered a practice session to familiarize themselves with the user interface and the gamepad controller. The purpose of this session was to erase potential novelty effects caused by the robot and the user interface. The practice round ended when the participants indicated they felt comfortable with the control and overall task. Then the participants went through three rounds of 'pick-and-place tasks' with the robot. Each round consisted of 6 trials and participants were asked to fill out a short survey after the last trial. After these three rounds, the participants were given a post-survey with open-ended questions about their experience.

Results. A total of 6 participants were recruited from a university campus and were randomly assigned to one possible combination of the experimental conditions. We compared the performance of the three agents (see Fig. 7 in the appendix for details). The total task completion time was the longest for the Defensive robot while the other three agents were similar but significantly faster. The Defensive robot also had the longest robot task execution time but also led to the shortest time for the human. Bayes-Nash was the least safe, the selfish and defensive conditions had similarly small safety stops.

8 Discussion, Limitations and Future Work

Figure 6 shows that the Bayes-Nash approach and the naturally-acting human are similar in their ability to adapt effectively to different personalities. However, in the human-robot study, Bayes-Nash had the most safety stops which contradicts **H1**. We believe there are two potential explanations.

First, due to human learning effects. Fig. 5 shows that the number of safety stops decreases with more trials for both the Defensive and Selfish-Nash conditions. This indicates that the user might have learned a collision-avoidant response to those agents over time. As for the Bayes-Nash condition, the safety stops first increase over trials and then remain constant. This might be because those two strategies had a somewhat fixed behavior that the human could easily adapt to. For example, if the robot moves to the goal without consideration of the human's presence every time, the human will learn that her optimal response is to wait for the robot initially. This is similar to the observation from [20] where the robot directly influenced human behavior. Although this led to better performance, in our scenario, it is questionable whether this fixed behavior will be desirable from a robot collaborator in the real-world. Second, due to the mismatch in human-robot speed. Although the maximum arm velocities were the same for both agents, the robot was able to act faster than the human. In the Selfish condition, after a couple of trials, people might have realized that it was easier to complete the task by waiting for the robot. We need further experiments that control for these variables to confirm these and plan for it in future work.

Although our approach can generalize to more agents, it was specifically designed for addressing scenarios involving human-robot cohabitation where usually only two agents are present [14, 6]. The time complexity of our algorithm is exponential in the number of agents, so can become intractable for large numbers of them. However, this may not be a limitation in real-world scenarios, since, even in cases where many agents are present (*e.g.*, driving on a highway), we only need to consider the interactive influence of a few close-by cars to generate human-like behavior [21]. In the future we would like to test it in the presence of more agents.

The primary contribution of our work is in developing a novel game-theoretic approach for HRI tasks. We also instituted a pilot study to validate this methodology and present descriptive statistics of the results. However, the small sample size did not allow us to run significance tests for the human experiments. Our plans for future work include a larger HRI study to provide evidence for generalization to a broad population.

References

[1] Shray Bansal, Akansel Cosgun, Alireza Nakhaei, and Kikuo Fujimura. Collaborative planning for mixed-autonomy lane merging. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[2] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. In *Advances in*

Neural Information Processing Systems, pages 5175–5186, 2019.

[3] Rosen Diankov. *Automated Construction of Robotic Manipulation Programs*. PhD thesis, Carnegie Mellon University, Robotics Institute, August 2010.

[4] David Engel, Anita Williams Woolley, Lisa X Jing, Christopher F Chabris, and Thomas W Malone. Reading the mind in the eyes or reading between the lines? theory of mind predicts collective intelligence equally well online and face-to-face. *PloS one*, 9(12), 2014.

[5] Jaime F Fisac, Eli Bronstein, Elis Stefansson, Dorsa Sadigh, S Shankar Sastry, and Anca D Dragan. Hierarchical game-theoretic planning for autonomous vehicles. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9590–9596. IEEE, 2019.

[6] Volker Gabler, Tim Stahl, Gerold Huber, Ozgur Oguz, and Dirk Wollherr. A game-theoretic approach for adaptive action selection in close proximity human-robot-collaboration. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[7] Kelsey P Hawkins, Shray Bansal, Nam N Vo, and Aaron F Bobick. Anticipating human actions for collaboration in the presence of task and sensor uncertainty. In *2014 ieee international conference on Robotics and automation (ICRA)*, 2014.

[8] Mark K Ho, James MacGlashan, Amy Greenwald, Michael L Littman, Elizabeth Hilliard, Carl Trimbach, Stephen Brawner, Josh Tenenbaum, Max Kleiman-Weiner, and Joseph L Austerweil. Feature-based joint planning and norm learning in collaborative games. In *CogSci*, 2016.

[9] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015.

[10] Steven M LaValle. Rapidly-exploring random trees: A new tool for path planning. 1998.

[11] Kevin Leyton-Brown and Yoav Shoham. Essentials of game theory: A concise multidisciplinary introduction. *Synthesis lectures on artificial intelligence and machine learning*, 2(1):1–88, 2008.

[12] George J Mailath. Do people play nash equilibrium? lessons from evolutionary game theory. *Journal of Economic Literature*, 36(3):1347–1374, 1998.

[13] Jim Mainprice, Rafi Hayne, and Dmitry Berenson. Goal set inverse optimal control and iterative replanning for predicting human reaching motions in shared workspaces. *IEEE Transactions on Robotics*, 32(4):897–908, 2016.

[14] Stefanos Nikolaidis, Ramya Ramakrishnan, Keren Gu, and Julie Shah. Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *ACM/IEEE international conference on human-robot interaction*, 2015.

[15] Stefanos Nikolaidis, Anton Kuznetsov, David Hsu, and Siddhartha Srinivasa. Formalizing human-robot mutual adaptation: A bounded memory model. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 75–82. IEEE, 2016.

[16] H. W. Park and A. M. Howard. Understanding a child's play for robot interaction by sequencing play primitives using hidden markov models. In *2010 IEEE International Conference on Robotics and Automation*, pages 170–177, 2010.

[17] Mildred B Parten. Social participation among pre-school children. *The Journal of Abnormal and Social Psychology*, 27(3):243, 1932.

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

[19] 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.

[20] Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and Systems*, 2016.

[21] Wilko Schwarting, Alyssa Pierson, Javier Alonso-Mora, Sertac Karaman, and Daniela Rus. Social behavior for autonomous vehicles. *Proceedings of the National Academy of Sciences*, 116(50):24972–24978, 2019.

[22] Ioan A. Şucan, Mark Moll, and Lydia E. Kavraki. The Open Motion Planning Library. *IEEE Robotics & Automation Magazine*, 2012. doi: 10.1109/MRA.2012.2205651.

[23] Peter Trautman and Andreas Krause. Unfreezing the robot: Navigation in dense, interacting crowds. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 797–803. IEEE, 2010.

[24] Annemarie Turnwald and Dirk Wollherr. Human-like motion planning based on game theoretic decision making. *International Journal of Social Robotics*, 11(1):151–170, 2019.

[25] 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.

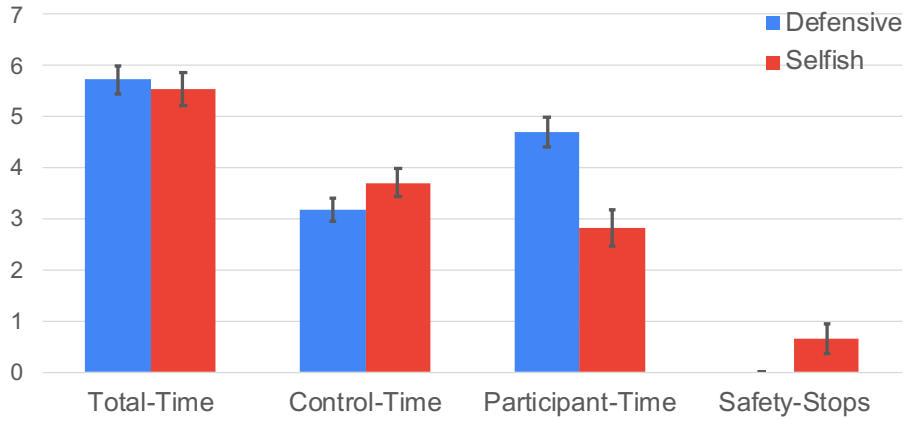
A Appendix: Human-Human Study

To investigate the natural interaction between two people, we recruited 4 participants to perform the same task, in a pilot experiment, where both interacting agents were human and controlled the simulated robot arm using a gamepad controller.

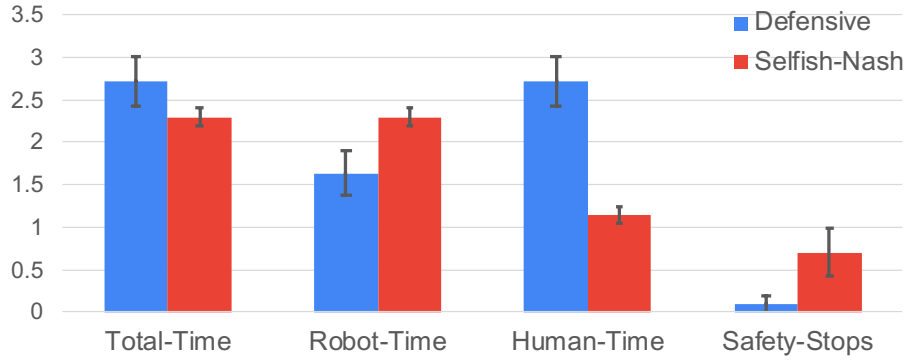
Experiment Design. We kept one of the human agents fixed throughout the experiment and will refer to them as *control*. The other agent (participant) evoked different behaviors in each experiment and performed 3 trials with the control. In the first trial, the participant was asked to behave **naturally**, by trying to increase efficiency while reducing task time. For the other two trials, the participant chose either (1) **Selfish** - completing this task efficiently or (2) **Defensive** - avoiding collisions with the other arm strategy. The *control* kept the same natural strategy throughout the interactions and was not made aware of the strategy that the participant was employing. Also, no verbal communication was allowed during the experiment. We measured the same metrics as in the simulated human experiment.

Results. Figure 6 (a) shows that the total task completion times for the Selfish and Defensive humans are similar when interacting with a naturally-acting human. However, in terms of their individual task completion times, the Defensive agent takes significantly longer as compared to the Selfish agent. The Selfish human also triggers more safety stops but the safety stops in this study were much less than in simulation. In Figure 6 (b) we show the simulated Defensive and Selfish-Nash behaviors when interacting with Bayes-Nash. We find similar comparative trends between behavior types for both task completion times and safety stops. However, the robot in (b) completes the task more quickly since the arm is allowed higher velocities in simulation.

Analysis. These results indicate that a naturally-acting human adapts well to both Selfish and Defensive behavior due to similar task metrics for both conditions, as shown in Figure 6(a). Similar trends for Bayes-Nash, Figure 6(b), indicate that it also adapts well to different strategies. The similarity between trends among personalities across experiments validate our interactive task design and metrics to benchmark performance for *parallel play*. Also, since the latent variable used to parameterize the equilibrium was designed to capture the agent's favorability in equilibria and not the specific behaviors of the baselines, we expect that Bayes-Nash to be able to adapt well to real human participants. This leads to the following two hypotheses for a human-robot interaction study:



(a) Natural (*Control*) Human



(b) Bayes-Nash

Figure 6: (a) Plots the interaction task metrics for the naturally acting human in the presence of either a selfish or defensive participant in the human-human study; (b) The same metrics but for the interaction of Bayes-Nash with the Selfish and Defensive baselines. The similarity in the relative trends across (a) and (b) highlight the similarity of Bayes-Nash to a real human agent. Error bars represent SEM.

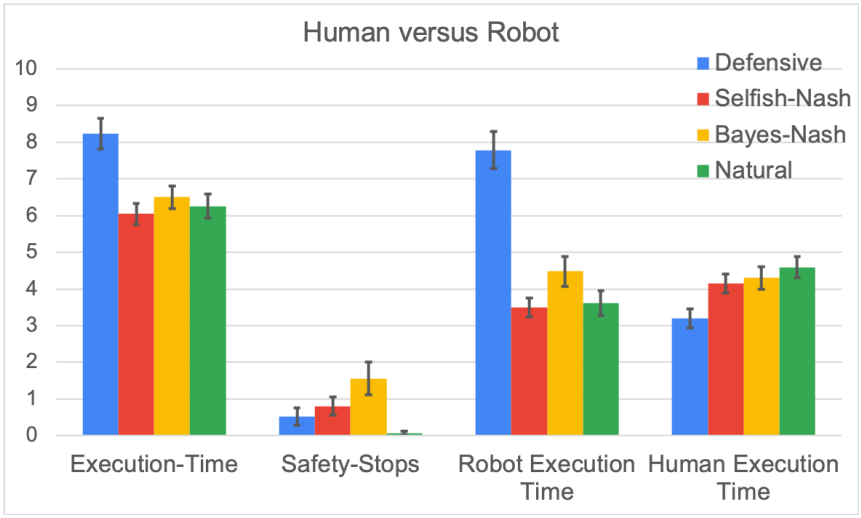


Figure 7: Human-Robot Study. Defensive and Selfish are baselines, Bayes-Nash is our approach and the Natural agent refers to the human-human study results where the participants acted naturally. Error bars represent SEM.