Autonomous Robots
https://doi.org/10.1007/s10514-021-10023-8

Bayes–Nash: Bayesian inference for Nash equilibrium selection in human-robot parallel play

Shray Bansal[1] · **Jin Xu**[1] · **Ayanna Howard**[2] · **Charles Isbell**[1]

Received: 1 February 2021 / Accepted: 27 September 2021

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.

Keywords Human–Robot interaction · Parallel play · Multi-agent systems · Game theory · Cooperative AI

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 (Parten 1932; Park and Howard 2010). 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 (Sadigh et al. 2016a; Bansal et al. 2018; Koppula and Saxena 2015). When working with others, people are often considerate of their intents and beliefs due to Theory-of-Mind (Premack and Woodruff 1978; Engel et al. 2014), 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.

Game Theory provides us tools to model this interdependence 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 (Leyton-Brown and Shoham 2008). A Nash equilibrium implicitly captures the interdependence between

✉ Shray Bansal
sbansal34@gatech.edu

Jin Xu
jxu81@gatech.edu

Ayanna Howard
howard.1727@osu.edu

Charles Isbell
isbell@cc.gatech.edu

[1] Georgia Institute of Technology, Atlanta, GA, USA

[2] Ohio State University, Columbus, OH, USA

Published online: 05 November 2021

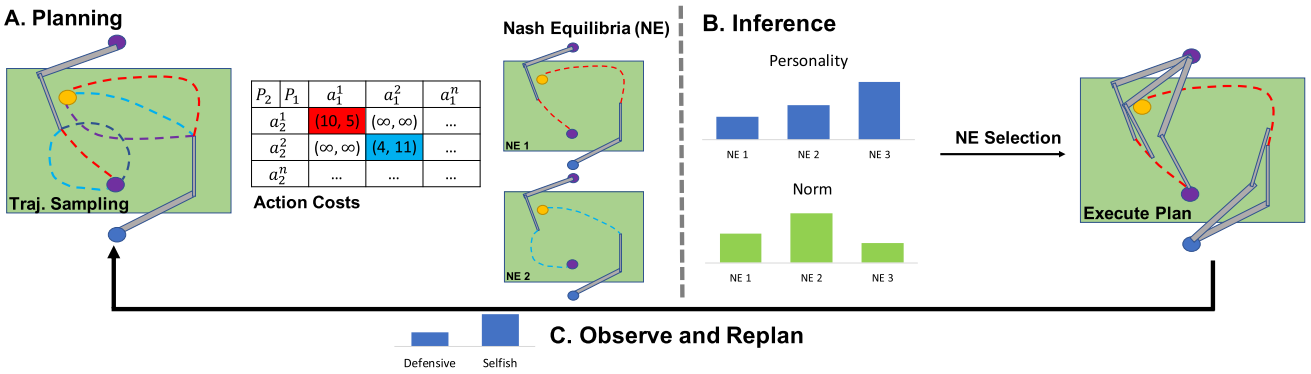

Fig. 1 Overview. Our approach has three main stages. **a** In the planning stage, we sample goal-reaching trajectories for both agents and compute their costs and use it to find the Nash equilibria (NE). **b** In the inference stage, we compute the probability for each equilibrium based on its adherence 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

agents, and our approach plans by finding equilibria strategies to enable better coordination.

Although humans have been shown to play to Nash equilibrium (Mailath 1998), 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 (Ho et al. 2016). Here, a norm refers to a set of abstract instructions that agents follow, and expect others to follow. In Fig. 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 (Gabler et al. 2017; Mainprice et al. 2016) 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 Cost formulation for interaction

In this section we describe a formulation of the cost function for multi-agent scenarios and define different interaction types based on the parameterization of this cost. The goal is to help illustrate the difference between them based on agent objectives and interaction dynamics. Table 1 briefly lists the interaction types.

We define an interaction as a game between multiple agents. Each agent i minimizes its cost c_i,

$$c^i(s_i, s_{-i}) = \alpha c^i_{native}(s_i) + (1-\alpha)c^i_{interactive}(s_i, s_{-i}). \quad (1)$$

Table 1 Interaction paradigms

Interaction	Description	Examples
Independent play	Agent is not influenced by other agents	Classic Atari games and board games (Brockman et al. 2016), classic control problems (*e.g.*inverted pendulum Barto et al. 1983)
Competitive play	Agent goals are in direct conflict leading to competition	Backgammon (Tesauro 1995), Go (Silver et al. 2016), or drone racing (Spica et al. 2020)
Collaborative play	Agents share the same goals and costs	Simulated driving (Bansal et al. 2018), cooperative games (Carroll et al. 2019), table-top manipulation (Nikolaidis et al. 2017)
Parallel Play	Agents have separate goals but shared space leads to conflict in achieving them.	Manipulation (Gabler et al. 2017), navigation (Sadigh et al. 2016a; Turnwald and Wollherr 2019)

Here, s_i is the state of agent i, s_{-i} refers to the states of the other agents, and c_{native} and $c_{interactive}$ refer to the cost derived by the agent's own state and by its relation to other agents, respectively. The cost for an agent is a linear combination of an individual component, which focuses on success in the individual task, as well as a social component, which considers the mutual effects of actions of other agents in the environment. Now, we consider a few different types of interaction.

2.1 Independent play

We refer to scenarios where an agent's success depends solely upon their own actions, and not those of other agents, as independent play. Such agents are unaffected by the agents. Therefore, they only minimize the native cost from Eq. 1, *i.e.*,

$$\alpha = 1, \quad c^i(s_i, s_{-i}) = c^i_{native}(s_i). \tag{2}$$

Any task involving only a single agent is an example of independent play. This also includes scenarios where agent influence is limited to itself, *e.g.*, driving on a road with only one car occupying a lane and barriers between lanes.

2.2 Competitive play

In competitive play scenarios, the success of an agent depends on the failure of the other agents. Examples of competitive play includes games like chess, Go or a race. Since the cost for an agent depends upon the state all agents are in, so, it will only include the interactive cost. Thus,

$$\alpha = 0, \quad c^i(s_i, s_{-i}) = c^i_{interactive}(s_i, s_{-i}). \tag{3}$$

For two-agent zero-sum games, $c_{interactive}$ will be the inverse for the two agents,

$$c^0_{interactive}(s_0, s_1) = -c^1_{interactive}(s_1, s_0) \tag{4}$$

2.3 Collaborative play

In a collaboration the payoff usually depends upon the states of both agents as well, *i.e.*, $c_{interactive}$ and thus, $\alpha = 0$ again. However, unlike the competitive scenario, this cost will be the same for all agents.

$$\alpha = 0, \quad c^i_{interactive}(s_i, s_{-i}) = c^j_{interactive}(s_i, s_{-i}) \; \forall j. \tag{5}$$

Examples of collaborative, shared-reward tasks include robots assisting humans in parts assembly, or robot teleoperation, *etc.* .

2.4 Parallel play

We define human-robot interaction scenarios that involve agents with separate goals but shared space as parallel play. In such scenarios, both native and interactive rewards play a role since agents have separate individual goals (c_{native}) but also aim to avoid interference, usually in the form of collision-avoidance, with the other agents ($c_{interactive}$). Hence, α in Eq. 1 is not a fixed value and depends on the scenario,

$$\alpha \in (0, 1). \tag{6}$$

Next, we identify some scenarios from previous work that fit our definition of parallel play.

2.4.1 Example: Close-proximity manipulation

Gabler et al. (2017) perform table-top manipulation in a close-proximity Human-Robot team, with the aim of performing this task efficiently by avoiding mutual interference. Although they introduce this as a collaboration, their model has separate costs for each agent. This is not the shared cost that characterizes collaborative play according to our definition. The cost of each action is determined by the following,

$$c^i = c^i_{native} + c^i_{interactive}. \tag{7}$$

Their cost function fits our formulation from Eq. 1 with $\alpha = 0.5$. They compute the Nash equilibria to plan the robot's actions.

2.4.2 Example: Navigation

Sadigh et al. (2016a) introduce an approach to plan actions for the simulated autonomous cars based on the effect it will have on the human drivers. They used Inverse Reinforcement Learning to model the human's reward and defined the robot's reward separately. Their reward functions include a linear combination of native costs, to incentivize lane driving, and interactive costs, to discourage collisions. These functions also fit our parallel play formulation with $\alpha \in (0, 1)$.

3 Related work

There has been extensive work in HRI for planning a robot to work on tasks around humans in domains like parts assembly (Hawkins et al. 2014; Gabler et al. 2017), motion planning, and autonomous driving (Sadigh et al. 2016a; Bansal et al. 2018). 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 emphasized accurately modeling the human's rational goal-driven behavior. This includes learning human preferences to predict low-level trajectories through a reward function obtained by inverse reinforcement learning (Ziebart et al. 2009; Sadigh et al. 2016a). It also includes methods focused on predicting action timing using apriori task-structure knowledge, *e.g.* parts delivery (Hawkins et al. 2014; Gombolay et al. 2015). Some of these models were also designed to be adaptable to the preferences of the particular human with whom the robot was interacting (Nikolaidis et al. 2015).

Human adaptive planning A common approach to planning is by first predicting the human's behavior and then finding a best-response to the predicted behavior. This approach works very well in scenarios where the robot shares the human's utility (collaborative play) and assumes an assistive role. For instance, in parts delivery (Hawkins et al. 2014; Unhelkar et al. 2014), where the robot intends to avoid interactions by keeping out-of-the-way while ensuring that the human doesn't wait. In close-proximity manipulation tasks, it has been used to plan robot trajectories that did not intersect with predicted human plans (Mainprice et al. 2016; Li and Shah 2019). An inherent assumption here is that, although the human's plan depends on the situation, the prediction is independent of the robot's plan. So, in situations where the agents have separate utilities, like in parallel play, the robot will choose overly conservative behaviors which can lead it to freeze when trying to navigate crowds (Trautman and Krause 2010) or fail to merge in traffic (Sadigh et al. 2016a).

Mutually adaptive planning Recent work has addressed this by considering the human's influence-ability as well (Turnwald and Wollherr 2019; Sadigh et al. 2016a; Fisac et al. 2019). Their model includes both the influence of the human and their goals on the robot, as well as the influence of the robot on the human. Similar to us, they also utilize game-theoretic tools to model this cyclical influence. Turnwald and Wollherr (2019) 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. Sadigh et al. (2016a) modeled driving as a Stackelberg game where the robot planned first and the human planned in response; they showed that this model can successfully influence human behavior in simulated driving tasks. Fisac et al. (2019) extended this to longer time horizons by computing a Nash equilibrium for high-level actions and optimizing low-level trajectories for executing them. Gabler et al. (2017) 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 this strategy (Gabler et al. 2017) or only find one equilibrium due to their problem structure (Sadigh et al. 2016a; Fisac et al. 2019).

Online model inference Although different people can perform the same task in multiple ways, past work generally modeled the behavior of users with only a single mode. In recent years, techniques to infer aspects of the behavior of the person or people that they are interacting with have been developed, we discuss some instances next. Nikolaidis et al. (2015) cluster human behavior into different *types* and predict their actions based on the inferred *type*. Nikolaidis et al. (2016) groups people by their adaptability to the robot's actions to decide the plan for the robot. Chen et al. (2018)

explicitly model human *trust* on the robot's ability during decision-making and infer this parameter during the interaction. Sadigh et al. (2016b) use information-gathering actions to infer whether a human driver is attentive or not, and use this to guide the robot's decision-making. We define a latent variable that represents human personality and infers this online, for each participant, in order to coordinate better with them. Recently, Schwarting et al. (2019) proposed a method for simulated autonomous driving around human drivers using Nash equilibrium. They also use a single parameter to represent human-personality and perform inference to find it. However, in their model, this parameter is part of the human cost function while we use it to choose between Nash equilibria. Also, they define a continuous action-space and use a local approximation for computing Nash equilibrium, which finds a single equilibrium. Their results indicate that inferring human preference in combination with finding Nash equilibrium improves prediction and helps achieve coordination with humans.

Coordination The importance of coordinating with a human in non-competitive games was highlighted by Carroll et al. (2019) where they learn human models and used them to train reinforcement learning agents that achieve performance superior to self-play. Also, Ho et al. (2016) proposed using social norms for improving coordination in multi-agent environments. Peters et al. (2020) used particle filtering to align autonomous cars to a single Nash equilibrium solution in an intersection navigation task which admits multiple equilibria.

Although, no standard metrics exist in collaborative HRI tasks to measure individual and team performance. Hoffman (2019) provides a guide for common metrics used to evaluate team fluency. This includes task completion time for every agent, total task time, the ratio of the time the human spent idle. While parallel play is not purely collaborative, we believe these metrics still capture important aspects of the interaction that we would like to measure and we will use these for evaluations. However, it should be noted that the utility of a metric depends on the task and stakeholder goals.

4 Interactive planning as a game

We model the multi-agent interactive planning task as a non-cooperative game represented as tuple G, $G = (P, A, c)$ (Leyton-Brown and Shoham 2008). Here, $P = \{P_1, \ldots, P_N\}$ is a finite set of N players, $A = A_1 \times \cdots \times A_N$ 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, $a = (a_1, \ldots, a_N)$. We define a cost representing the unfavorability of an action profile for agent i as $c_i : A \mapsto \mathbb{R}$ and $c = (c_1, \ldots, c_N)$ includes the mapping for all agents.

In our scenario, each agent p is a robot arm, each action set A_p 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.

$$a_i^* \in \arg\min_{a_i} c_i(a_1^*, ., a_i, ., a_N^*) \quad \forall i \in N. \tag{8}$$

Although generally, only one (mixed) equilibrium is guaranteed to exist for a game (Leyton-Brown and Shoham 2008), 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 $c_i(a^{*1}) < c_i(a^{*2}) \forall i \in N$. Next, we present an approach that can help the agent select an action by choosing between Pareto-optimal equilibria.

5 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,

$$p(a) = p_n(a) p_\alpha(a). \tag{9}$$

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.

5.1 Norm

Similar to Ho et al. (2016), 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,

$$p_n(a) \propto e^{-\lambda_n min_i(c_i(a))}, \tag{10}$$

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

5.2 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,

$$p^t_\alpha(a) = p(a|H_t). \tag{11}$$

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

$$p^{t=0}_\alpha(a) = \text{uniform}(a) \quad \forall a. \tag{12}$$

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

$$p(a^0|H) \propto e^{-\lambda_\alpha f_{dist}(a^0,H)}. \tag{13}$$

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. 13, 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 all of the information required for predicting the agent's next action, which makes a^t conditionally independent of H^t given θ. So,

$$p(a^t|H^t) = \sum_\theta p(\theta|H^t)p(a^t|\theta). \tag{14}$$

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

$$p(a|\theta) = \frac{\mathbb{1}_\theta(a)}{\sum_{a'} \mathbb{1}_\theta(a')}. \tag{15}$$

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,

$$p(\theta|H^t) = \sum_{a^0} p(a^0|H^t)p(\theta|a^0). \tag{16}$$

Since we assume a uniform prior on θ, $p(\theta)$ is a constant. Combining with Eq. 15, we get,

$$p(\theta|a^0) = \frac{p(\theta)p(a^0|\theta)}{p(\theta)\sum_{\theta'} p(a^0|\theta')} = \frac{p(a^0|\theta)}{\sum_{\theta'} p(a^0|\theta')} \tag{17}$$

We use $p(\theta|a^0)$ and $p(a^0|H^t)$ (Eq. 13) to get $p(\theta|H^t)$ in Eq. 16. This allows us to find $p(a^t|H^t)$ from Eq. 16 by using $p(a^t|\theta)$ from Eq. 15, which gives us $p_\alpha(a^t)$ in Eq. 11

6 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 Fig. 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.

6.1 Action planning

To plan for this task, we first sample $k-1$ plans for each agent in configuration space using a Rapidly-exploring Random

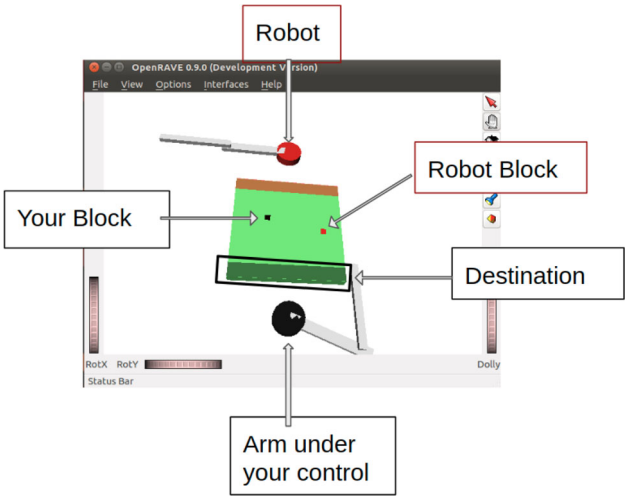

Fig. 2 Task setup

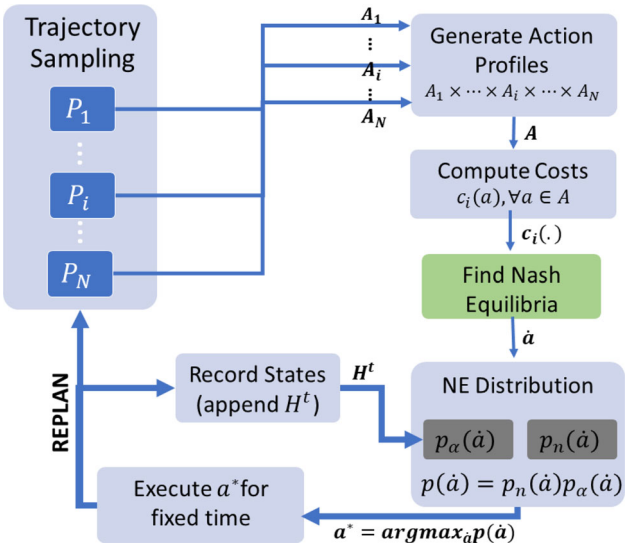

Fig. 3 Framework. We first plan trajectories, then use the cost to compute all Pareto-optimal Nash equilibria, then combine the norm and inferred-personality distributions to select the most likely equilibrium. This action is partially executed before repeating the whole process until the task completes

Tree (RRT) (Lavalle 1998) 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, $A = A_1 \times \cdots \times A_N$. We compute a cost for each action profile by simulating it and use Eq. 8 to find the Nash equilibria. We choose the NE profile a that maximizes the distribution $p(a)$ from Eq. 9, 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 Fig. 3.

6.2 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.

6.3 Time complexity

Our approach described in Fig. 3 includes five main components. The first one is trajectory sampling which has a time complexity of $O(Nkt_{RRT})$, where k is the number of trajectories sampled per agent, N is the number of agents, and t_{RRT} is the time that it takes to plan a single trajectory. The second component includes generating action profiles and computing the cost associated with each of them, this takes $O(k^N T)$ time, where T is the number of time-steps in each trajectory. The next component finds the Nash equilibria and takes $O(k^N)$ time. The last component computes the distribution over Nash equilibria and also takes $O(k^N)$ time. So, the second component dominates and the total time complexity of our approach is $O(k^N T)$.[1]

6.4 Baselines

We define three baselines to compare with our approach.

1. 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

[1] This analysis assumes a parameterization of the RRT algorithm such that it completes in a reasonable amount of time.

complete their task.

$$a_R = \arg\min_{a_R \in A_R} \; max_{a_H \in A_H} c_R(a_R, a_H). \tag{18}$$

2. 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.

$$a_R \in \mathrm{a}^*, \mathrm{a}^* = \arg\min_{\mathrm{a}^*} c_R(\mathrm{a}^*). \tag{19}$$

3. Norm-Nash. Chooses an equilibrium profile that maximizes the norm distribution p_n from Eq. 10. 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.

$$a_R \in \mathrm{a}^*, \mathrm{a}^* = \arg\max_{\mathrm{a}^*} p_n(\mathrm{a}^*). \tag{20}$$

6.5 Implementation details

A 3D simulation environment was created using the open-source Open Robotics Automation Virtual Environment (OpenRAVE) (Diankov 2010) 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) (Sucan et al. 2012). 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. 10) 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 $\theta = 1$ selects equilibrium profiles a that favor agent 2, $c_2(a) < c_1(a)$. We set $\lambda_\alpha = 10$ in the personality distribution (Eq. 13).

7 Simulated human study

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

7.1 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.

7.2 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.

7.3 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 Fig. 5. 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.

7.4 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. Next, we perform an experiment to validate this trade-off in human interaction.

7.5 Qualitative results

Figure 4 shows end-effector trajectories that lead to Nash equilibrium in this task. It shows all the pareto-optimal equilibria found in three different goal object location configurations. We find between one and three equilibria in each configuration with different costs for the two agents. These differences in equilibria are caused due to the relative location of the objects as well as the stochasticity of the planner. For instance, the higher relative difference between c_1 and c_2 in configuration 3 as compared to 2 can be attributed to the objects being closer to the respective robots in configuration 2. Optimal plans for one agent in configuration 2 thus do not disadvantage the other agent as much as in goal configuration

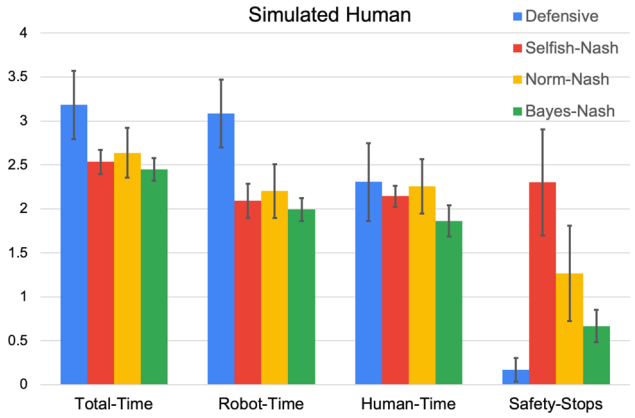

Fig. 5 Results from the experiment involving a simulated human. Note that our proposed approach, Bayes–Nash, is safer than all the non-defensive baselines and similar in task completion time to the Selfish-Nash baseline. Error bars represent standard error of the mean (SEM)

3. Also note that these equilibria were computed when the robots were in their initial states and the equilibria found will change as the task progresses and agents move to different states.

8 Human–human study

To investigate the natural interaction between two people, we recruited 4 male students aged 28–32 from our university campus 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.

8.1 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.

8.2 Results

Figure 6a shows that the total task completion times for the Selfish and Defensive humans are similar when interacting

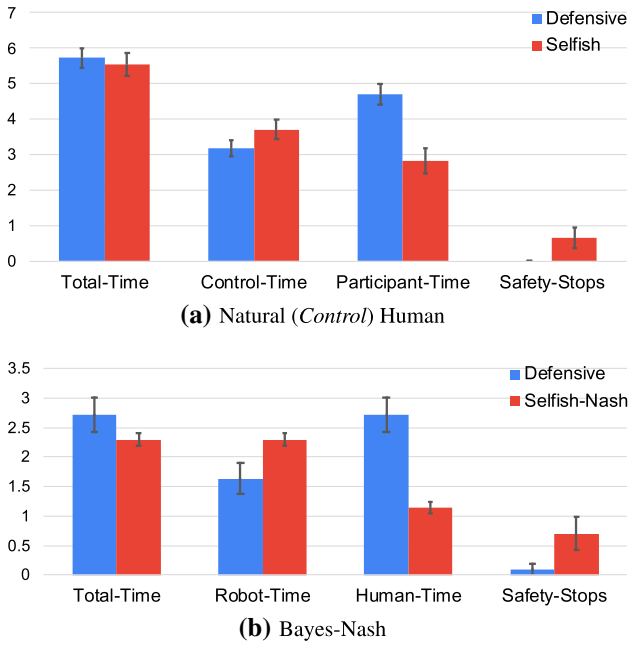

Fig. 6 **a** Plots the interaction task metrics for the naturally acting human in the presence of either a selfish or defensive participant in the human-human study; **b** The same metrics but for the interaction of Bayes–Nash with the simulated human. The similarity in the relative trends across **a** and **b** highlight the similarity of Bayes–Nash to a real human agent. Error bars represent SEM

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 Fig. 6b 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.

8.3 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 Fig. 6a. Similar trends for Bayes–Nash, Fig. 6b, 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

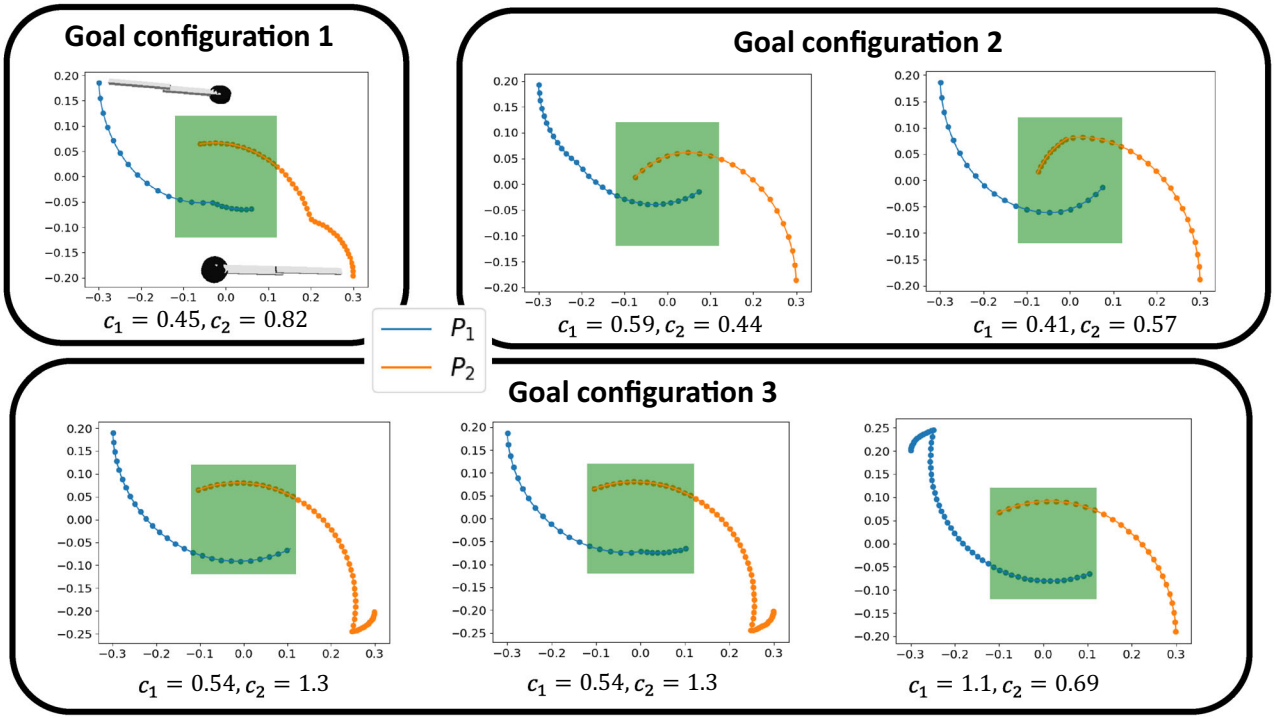

Fig. 4 Nash Equilibrium trajectories for different goal configurations. Goal configuration 1 has a single equilibrium. In the plot, the green rectangle represents the table and the end-effector trajectories for the two robots (P_1, P_2) are shown with the dots being sampled at equal time-intervals to show the speed of the arm during the trajectory. The costs for the trajectories are shown under each plot, where c_1, c_2 refer to those for P_1, P_2 respectively. In goal configuration 2 we find two equilibria, one favoring each agent, while goal configuration 3 has three equilibria where 2 of them are more favorable for P_1

well to real human participants. This leads to the following two hypotheses for a human-robot interaction study:

H1: A robot using Bayes–Nash will have significantly fewer collisions than a robot using a Selfish-Nash planner.

H2: A robot using Bayes–Nash will have faster task completion time than a robot using a Defensive planner.

9 Human–robot study

We test these hypotheses in a pilot experiment by pairing human participants with a robot controlled by our algorithm and the baselines.

9.1 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, as shown in Fig. 7. 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.

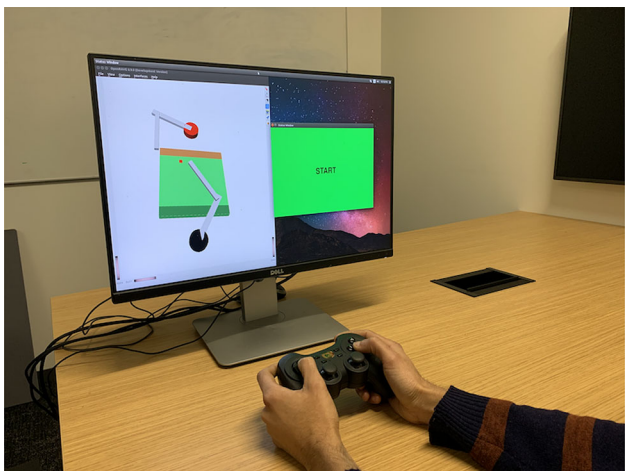

Fig. 7 A user controlling the robot arm during the Human–Robot study

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.

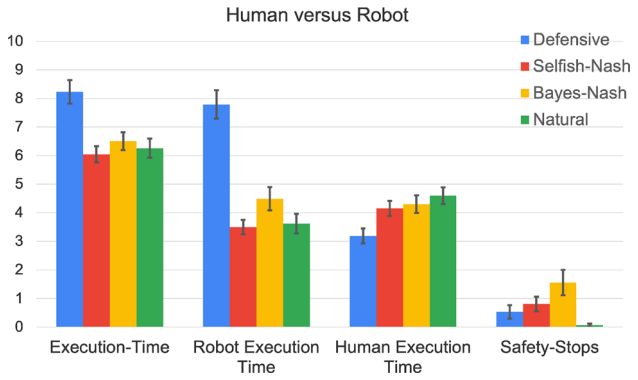

Fig. 8 Human–Robot Study. Defensive and Selfish are baselines, Bayes–Nash is our approach and the Natural agent refers to the human–human study results where the participants acted naturally. Error bars represent SEM

9.2 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.

9.3 Results

A total of 6 students aged 25–29 ($M = 27.8$, $SD = 1.3$, 5 male) were recruited from our university and were randomly assigned to one possible order of the experimental conditions. Figure 8 compares the performance of the three agents. We also compare them to a naturally acting human by including the results of two naturally acting humans from the human-human study. The total task completion time was the longest for the Defensive robot while the other three agents were similar but significantly faster supporting hypothesis **H2**. 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 and Natural the most, the selfish and defensive conditions had similarly small safety stops, contrary to hypothesis **H1**.

9.4 Analysis

It took the Defensive agent 36.2% more time to complete the entire task than the Selfish agent. These results confirm that the Defensive agent acts in an overly cautious manner. When comparing with the naturally-acting human we also noticed that the task completion time for the Bayes–Nash approach performs almost equally well. However, the safety stops results are surprising in two respects: (1) the higher number of safety stops for Bayes–Nash as compared to Natural and other conditions, (2) the much fewer safety stops for all conditions when compared to the simulated human study. We also noticed the latter in the human-human study, indicating that people are better at avoiding collisions as compared to the robots. We explore potential causes for these findings in the next section.

10 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. Figure 9 shows that the number of safety stops decrease 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 Sadigh et al. (2016a) 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. Also, our framework could use knowledge from previous interactions with the same agent to improve the interaction, e.g., by using a prior on the α inferred from the previous trial.

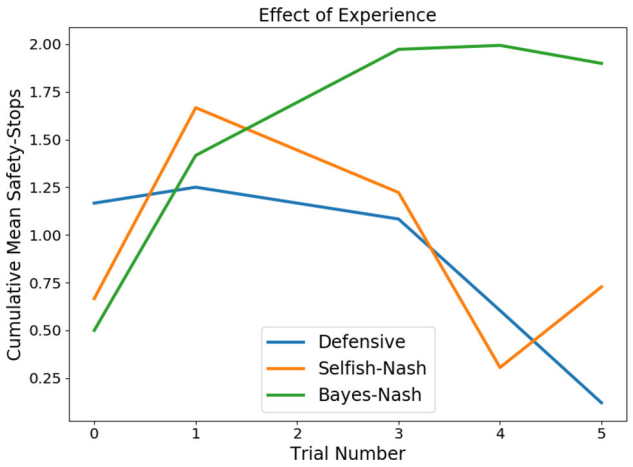

Fig. 9 Effect of experience in the Human-Robot study. It shows the cumulative safety stops for the robot averaged over the trials. For the Defensive and Selfish-Nash conditions, the number of stops decreases with more trials, indicating that the human learns to avoid collision. However, for Bayes–Nash, the safety stops first increase and then remain constant, perhaps due to the difficulty in adapting to an agent whose behavior is not fixed

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 (Nikolaidis et al. 2015; Gabler et al. 2017). 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 (Schwarting et al. 2019). In the future we would like to test it in the presence of more agents.

The action planning in our approach utilizes an RRT planner to compute high-level trajectories for the pick-and-place task. This allows us to efficiently compute all of the pure Nash equilibrium strategies present. However, since the planner is ignorant of the presence of the other agents, the variations in the sampled trajectories are random and not adapted to the goals of the other agents which can make the interactions less efficient. This can be improved in the current framework by conditioning new plan samples on previously sampled plans of the other agent.

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

Bansal, S., Cosgun, A., Nakhaei, A., & Fujimura, K. (2018). Collaborative planning for mixed-autonomy lane merging. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics, 5,* 834–846.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). Openai gym. arXiv preprint arXiv:1606.01540.

Carroll, M., Shah, R., Ho, MK., Griffiths, T., Seshia, S., Abbeel, P., & Dragan, A. (2019). On the utility of learning about humans for human-ai coordination. In *Advances in Neural Information Processing Systems* (pp 5175–5186).

Chen, M., Nikolaidis, S., Soh, H., Hsu, D., & Srinivasa, S. (2018). Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction* (pp. 307–315).

Diankov, R. (2010). *Automated construction of robotic manipulation programs*. PhD thesis, Carnegie Mellon University, Robotics Institute.

Engel, D., Woolley, A. W., Jing, L. X., Chabris, C. F., & Malone, T. W. (2014). 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)

Fisac, J. F., Bronstein, E., Stefansson, E., Sadigh, D., Sastry, S. S., & Dragan, A. D. (2019). Hierarchical game-theoretic planning for autonomous vehicles. In *2019 International conference on robotics and automation (ICRA)* (pp 9590–9596). IEEE.

Gabler, V., Stahl, T., Huber, G., Oguz, O., & Wollherr, D. (2017). A game-theoretic approach for adaptive action selection in close proximity human-robot-collaboration. In *2017 IEEE international conference on robotics and automation (ICRA)*.

Gombolay, M. C., Gutierrez, R. A., Clarke, S. G., Sturla, G. F., & Shah, J. A. (2015). Decision-making authority, team efficiency and human worker satisfaction in mixed human-robot teams. *Autonomous Robots, 39*(3), 293–312.

Hawkins, K. P., Bansal, S., Vo, N. N., & Bobick, A. F. (2014). Anticipating human actions for collaboration in the presence of task and sensor uncertainty. In *2014 ieee international conference on Robotics and automation (ICRA)*.

Ho, M. K., MacGlashan, J., Greenwald, A., Littman, M. L., Hilliard, E., Trimbach, C., Brawner, S., Tenenbaum, J., Kleiman-Weiner, M., & Austerweil, J. L. (2016). Feature-based joint planning and norm learning in collaborative games. In *CogSci*

Hoffman, G. (2019). Evaluating fluency in human-robot collaboration. *IEEE Transactions on Human-Machine Systems, 49*(3), 209–218.

Koppula, H. S., & Saxena, A. (2015). Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(1), 14–29.

Lavalle, S. M. (1998). Rapidly-exploring random trees: a new tool for path planning. Tech. rep.

Leyton-Brown, K., & Shoham, Y. (2008). Essentials of game theory: A concise multidisciplinary introduction. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 2*(1), 1–88.

Li, S., Shah, J. A. (2019). Safe and efficient high dimensional motion planning in space-time with time parameterized prediction. In *2019 international conference on robotics and automation (ICRA)*.

Mailath, G. J. (1998). Do people play nash equilibrium? Lessons from evolutionary game theory. *Journal of Economic Literature, 36*(3), 1347–1374.

Mainprice, J., Hayne, R., & Berenson, D. (2016). Goal set inverse optimal control and iterative replanning for predicting human reaching motions in shared workspaces. *IEEE Transactions on Robotics, 32*(4), 897–908.

Nikolaidis, S., Kuznetsov, A., Hsu, D., & Srinivasa, S. (2016). Formalizing human-robot mutual adaptation: A bounded memory model. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 75–82). IEEE.

Nikolaidis, S., Nath, S., Procaccia, A. D., & Srinivasa, S. (2017). Game-theoretic modeling of human adaptation in human-robot collaboration. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction* (pp. 323–331).

Nikolaidis, S., Ramakrishnan, R., Gu, K., & Shah, J. (2015). Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *ACM/IEEE international conference on human-robot interaction*.

Park, H. W., & Howard, A. M. (2010). 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* (pp. 170–177).

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

Peters, L., Fridovich-Keil, D., Tomlin, C., & Sunberg, Z. (2020). Inference-based strategy alignment for general-sum differential games. In *AAMAS '20, international foundation for autonomous agents and multiagent systems*. https://github.com/lassepe/AAMAS2020-GameInference-Paper/blob/master/submission/ibsa-camera-ready-aamas2020.pdf.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(4), 515–526.

Sadigh, D., Sastry, S., Seshia, S. A., & Dragan, A. D. (2016a). Planning for autonomous cars that leverage effects on human actions. In *Robotics: Science and systems*.

Sadigh, D., Sastry, S. S., Seshia, S. A., & Dragan, A. (2016b). Information gathering actions over human internal state. In *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 66–73). IEEE.

Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S., & Rus, D. (2019). Social behavior for autonomous vehicles. *Proceedings of the National Academy of Sciences, 116*(50), 24972–24978.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature, 529*.(7587):484

Spica, R., Cristofalo, E., Wang, Z., Montijano, E., & Schwager, M. (2020). A real-time game theoretic planner for autonomous two-player drone racing. *IEEE Transactions on Robotics, 36*(5), 1389–1403. https://doi.org/10.1109/TRO.2020.2994881.

Sucan, I. A., Moll, M., & Kavraki, L. E. (2012). The open motion planning library. *IEEE Robotics & Automation Magazine*. https://doi.org/10.1109/MRA.2012.2205651.

Tesauro, G. (1995). Temporal difference learning and td-gammon. *Communications of the ACM, 38*(3), 58–68.

Trautman, P., & Krause, A. (2010). Unfreezing the robot: Navigation in dense, interacting crowds. In *2010 IEEE/RSJ international conference on intelligent robots and systems* (pp. 797–803). IEEE.

Turnwald, A., & Wollherr, D. (2019). Human-like motion planning based on game theoretic decision making. *International Journal of Social Robotics, 11*(1), 151–170.

Unhelkar, V. V., Siu, H. C., Shah, J. A. (2014). Comparative performance of human and mobile robotic assistants in collaborative fetch-and-deliver tasks. In *ACM/IEEE international conference on human-robot interaction (HRI)*.

Ziebart, B. D., Ratliff, N., Gallagher, G., Mertz, C., Peterson, K., Bagnell, J. A., Hebert, M., Dey, A. K., & Srinivasa, S. (2009). Planning-based prediction for pedestrians. In *2009 IEEE/RSJ international conference on intelligent robots and systems* (pp. 3931–3936). IEEE.

Shray Bansal is a Ph.D. student in CS at the Georgia Institute of Technology. He received his B.E. in Computer Engineering from the Delhi College of Engineering in 2010 and his M.S. in CS from the Georgia Institute of Technology in 2014. His research interests include artificial intelligence, human-robot interaction, and multi-agent learning.

Jin Xu is a Ph.D. student in Robotics at the Georgia Institute of Technology. He received his B.S. degree in Electrical Engineering in 2013, his M.S. degree in Electrical and Computer Engineering in 2020, and his M.S. degree in Computer Science in 2021, all from Georgia Institute of Technology. His research interests include human-robot interaction, machine learning, and rehabilitation robotics.

Dr. Ayanna Howard Ph.D. is the Dean of Engineering at The Ohio State University and Monte Ahuja Endowed Dean's Chair. She is also an IEEE and AAAI Fellow. Her research interests focus on artificial intelligence (AI), assistive technologies, and robotics.

Dr. Charles Isbell received his BS in CS from Georgia Tech and his PhD in CS from MIT. He is now the John P. Imlay, Jr. Dean of the College of Computing at Georgia Tech. The unifying theme of his research has been using machine learning to enable autonomous agents to engage in life-long learning in the presence of thousands of other intelligent agents, including humans. He is a Fellow of both the ACM and AAAI, and an Elected Member of the American Academy of the Arts and Sciences.