Markov Games with Time-variant Types as a Framework for Human-robot Coordination

Shih-Yun Lo[1], Shani Alkoby[2], Benito Fernandez[1], and Peter Stone[2]

Abstract— **Coordination between humans and robots commonly happens when they co-exist in a shared workspace. Such scenario includes human-robot teaming on collaborative tasks, and human-robot conflict-resolving on limited shared resources. Mis-coordination reduces efficiency of both parties, and reduces the human's trust and patience with the robot. In this work, we propose a game-theoretic framework to analyze convergence in human-robot coordination. We also propose human behavior hypotheses on their decision-making mechanisms under this framework, to capture how the human 1) perception of the robot's capabilities, 2) personal preferences, 3) level of self-interest, and 4) social trust affect their policies and adaptability in dynamic environments. We provide human-robot path crossing as an instantiation of our framework, and use the hypothesized human behaviors to simulate real-world observed interaction patterns. Lastly, we simulate humans acting adaptively to their observed robot policies, as an initiative to incorporate effects on humans when designing robot algorithms using human-human interaction data.**

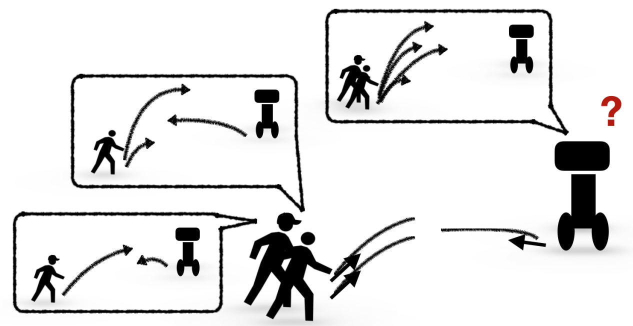

Fig. 1. Human-robot coordination in a crossing interaction. Since people have different prior assumptions on robot behaviors, their choice of actions differ.

I. Introduction

Human-robot interaction has received increased attention in recent years due to the emerging interest in deploying robots in human environments. Such environments may involve human-robot collaboration on given tasks, humans and robots working in a shared workspace, or service robots deployed in human environments. In such environments, robots may need to coordinate with humans with partially shared information and partially shared objectives; agents may need to reach agreement on one solution among multiple feasible choices, which makes the coordination non-trivial to settle. For a motivating example, see Fig. 1, which shows three agents coordinating at an intersection, where each agent has an individual goal to reach.

For a robot to engage coordination without confusing the human, it first requires the basic capabilities to understand human intent, and to respond in a legible manner [7]. Beyond those, to negotiate and agree on the coordinating solutions, the robot needs knowledge of human agents' behaviors to predict the outcome of its own actions. It also needs to known the reaction time humans need to update their policies [27], and to consider potential impacts of its own actions on humans' future decisions [10], [9]. Then the robot can plan and coordinate with humans in an intent-consistent fashion.

Past research has sought to improve human-robot coordination in a variety of ways, including: intent-expressive robot motion generation [7], [19], human preference-aware behavior modeling and its use for coordination [11], [5], human-robot mutual adaptation [23], [22], human expectations on robot capability [2], [17], as well as trust and comfort for long-term deployment[32].

While these topics share a focus on factors which affect human behaviors and their decision-making mechanisms, they lack a unifying framework to keep track of how those factors correlate with each other in different situations, how they affect human decisions and motions, and how they evolve over time as a function of robot interaction policies. Those questions are important for designing intent-consistent robots to interact with humans in daily situations. And they are important for designing robots that are aware of human adaptation to their long-term deployments. To fulfill such purposes, the main contribution of this work is a proposed game-theoretic model for human-robot coordination where agents have hidden preferences and time-variant adaptive behaviors based on online observations of others' behaviors.

To discuss human behaviors while interacting with robots, we point out some important factors that have been proposed in the human-robot interaction community, such as perceived robot capability, personal preferences, and trust on robot ability. We illustrate how they affect human behaviors through our proposed human decision-making model under the game framework, using a receding-horizon approach. We also propose two factors that we observed to be important in the social navigation domain: level of self-interest and social trust. We incorporate them along with others to analyze how people interact with the robot, with different personal preferences and assumptions about robot behaviors. We collect real-world human-robot interaction data to illustrate our model effectiveness.

We also use our decision-making model to discuss human adaptability to robots, and simulate human adaptation to robots after perceiving robot socially trust-worthy behaviors.

II. Related Work

Despite the growing interest in deploying robots in human populated workspaces, motion planning algorithms for smooth human-robot coordination have remained a challenge. While traditional planning algorithms have shown success in static environments, deploying such approaches alongside humans has shown insufficient adaptability to highly dynamic environments, and produce awkward motions making interpretability difficult [19], [7], [15]. Approaches considering time-variant factors, such as temporal constraints for multi-agent collision avoidance [31], have drawn attention for such applications; approaches considering human factors, such as human collision-avoidance behavior anticipation [12], have also gained attention for robot planning in human workspaces [28].

On the other hand, another community solves robot planning in human environments as a joint multi-agent planning problem, to incorporate the collective crowd behaviors with explicit modeling of the effect of agent actions on its surround agents [29], [16], [21]. Such joint modeling methodology has shown to be effective at outputting smooth human-mimicking trajectories at the same time acting responsively to its surrounding agents.

One major drawback applying these approaches interacting with humans is that the multi-agent joint dynamics models are typically learnt from data collected by human demonstrations, whereas humans do not act the same way around a robot compared to in pure-human environments. This problem has shown inefficiency of such approaches when humans use unexpected behaviors around robots – behaviors that humans will not present in front of another human [26].

To model joint behaviors among agents – how one's action affects the other – another way is to incorporate individual's action values with correlations with other agents' actions. Such joint behavior formulation is widely studied in Game theory: with different player assumptions, games evolve with different outcomes. In the artificial intelligence community, such game formulation has been incorporated with Markov Decision Process for multi-agent reinforcement learning [20]. Such incorporation enables strategy design for multi-agent robotic systems with mutual learning.

Yet, to simulate human-robot interaction, it requires agents modeled after humans, who have distinctive learning mechanisms and decision-making processes from reinforcement learning agents; moreover, agents have different behavioral types, as humans have personal preferences; agents online adapt to other agents, as humans observe the others and plan accordingly. Those features are widely studied in the human-robot interaction community, but not yet well formulated by one unifying framework. We therefore propose an extensive framework on Markov games with time-variant types to incorporate human-robot interaction with multi-agent learning problems to eliminate the gap.

III. Preliminaries

A. Game Definition

Consider a game G with k players, where each player $i \in \{1...k\}$ has a finite action set A^i. The set of action profiles is denoted as $\Sigma = A^1 \times A^2 \times ... \times A^k$. The utility of an agent i is a function, denoted as $f^i : \sigma \to \mathbb{R}$, evaluated at $\sigma \in \Sigma$.

Games that repeat for more than one action per player are defined as *repeated games* G^T, in which players receive cumulative utilities over a time horizon T, defined as:

$$V^i = \Sigma_{t=0}^{T} f_t^i(\sigma_t). \tag{1}$$

The action profiles σ_t over time collectively represent the strategy profile $s = \sigma_0 \times \cdots \sigma_T \in S$

We model human-robot interaction as a game, because each agent's utility not only depends on his or her own action, but also on others' actions. Let $a^H \in A^H$ and $a^R \in A^R$ be the action space of humans and robots, respectively. Consider a two-player game with agent R and H, a strategy profile $s^* = (a_{0:T}^{R*}, a_{0:T}^{H*})$ is a *Nash Equilibrium*, if no agent benefits from unilateral deviation from his or her current actions:

$$\forall i \in \{H, R\}, t, a_t^{i*} \in A^i, f^i(a_t^{i*}, a_t^{-i*}) \geq f^i(a_t^i, a_t^{-i*}). \tag{2}$$

Here a_t^{-i} refers to actions at time t from all agents but i, e.g. $a_t^{-H} = a_t^R$.

In repeated games, agents not only need to consider the current outcome of an action, but also its impact on the other agents' future actions, which affect their own expected cumulative rewards.

B. Markov Games

Markov games [20] are defined on top of Markov Decision Process, with finite state space $s \in S$, finite action space $a \in A$, and other agents' action space $a' \in A'$. Agent reward function is a function of the state, action, as well as other agents' actions: $r(s, a, a')$. The framework is commonly used for multi-agent reinforcement learning, where agent action value $Q(x, a)$ is defined to take in the other agents' actions:

$$Q(s, a, a') = r(s, a, a') + \gamma \Sigma_{s'} \mathcal{T}(s'|s, a, a')V(s'), \tag{3}$$

where γ is the discount rate, s' is the state transition from (s, a, a'), $\mathcal{T}$ is the transition probability, and V is the value function.

We propose a model, extended from Markov Games, to incorporate types as in Bayesian games, which assign game outcomes based on both agent actions and types.The model can also take in continuous-space input space U to apply to real-world robotics domains.

IV. Markov Games with Time-variant Types

We model agents $i = 1, \cdots, k$ to work in a joint state space X, where each individual has its own state representation $x_t^i \in X^i$ at time t, and control input from a bounded input space $u_t^i \in U^i$. $x_t = (x_t^1, \ldots, x_t^k) \in X$.

While actions $a_t^i \in A^i$ define the *high-level, finite* actions an agent can take to affect the game outcome, u_t^i defines the low-level continuous-space realization of such options, by:

$$u_t^i = g^i(x_t, a_t^i, \theta_t^i), \tag{4}$$

where $\theta_t^i \in \Theta^i$ is some parametrization of the agent's (potentially) time-variant behavior. $g^i : X \times A^i \times \Theta^i \to U$ can take in any model formulation, possibly stochastic, to sample inputs from $p(u_t^i|x_t, a_t^i, \theta_t^i)$. One example for high-level actions is table-turning directions; the low-level inputs to realize such actions are manipulator torque inputs. Note that, u_t^i is a function of x_t, not solely of x_t^i, since agents adjust their motions based on other agents' status in the joint state space.

The state transition function of agent i, $\mathcal{T}^i : X^i \times U^i \rightarrow X^i$, can then also be represented through $p(x^i_{t+1}|x_t, a^i_t, \theta^i_t)$, by marginalizing over u^i_t:

$$p(x^i_{t+1}|x_t, a^i_t, \theta^i_t) = \int_{u^i \in U^i} p(x^i_{t+1})|x^i_t, u^i_t)p(u^i_t|x^t, a^i_t, \theta^i_t)du^i. \tag{5}$$

$\mathcal{T}^i : X \times A^i \times \Theta^i \rightarrow X$, as agent i is part of the joint state; whereas the joint state transition function $\mathcal{T}$ takes in the form: $\mathcal{T} : X \times \Sigma \times \Theta \rightarrow X$, as a collective behavior among all agents.

Each agent receives an immediate reward r^i_t after taking input u^i_t at state x_t:

$$r^i_t = r(x_t, u^i_t, u^{-i}_t), \tag{6}$$

where u^{-i}_t is the control input of other agents at time t. The reward function r considers all agents' states x_t and inputs $u_t \in U = U^1 \times \cdots \times U^k$ for evaluation. As a result, $r : X \times U \rightarrow \mathbb{R}$; or, $r : X \times \Sigma \times \Theta \rightarrow \mathbb{R}$, considering the deterministic controller in Eq.4:

$$r^i_t = r(x_t, \sigma_t, \theta_t). \tag{7}$$

With the reward function r defined in Eq.7, the optimal policy is then to fine the strategy s^i that maximizes the cumulative rewards,

$$\begin{aligned} a^{i*}_{0:T} = \underset{a^i_{0:T}}{\operatorname{argmax}}\, \Sigma^T_{t=0} \mathbb{E}_{\theta_t, \sigma_t, x_t | \mathcal{T}, \sigma_{0:t-1}} \left[r^i(x_t, \sigma_t, \theta_t) \right] + \\ V^i_{T+1}(x_T, \sigma_T, \theta_T), \end{aligned} \tag{8}$$

where V^i_{T+1} some cost-to-go function for terminating at σ_T at $t = T$. Despite the general formalism for this framework to consider planning directly with the low-level control inputs u^i_t, discrete actions are computationally cheaper for plan evaluation and oftentimes easily obtained for domain-specific applications. Discrete action planning also well approximate human planning with hierarchical reasoning.

We present the framework, as Markov games with unknown time-variant types, as a tool to analyze the convergence criteria of human-robot coordinating process. We first introduce general solutions for the framework in both precomputation and online setting in Sec. V; we then introduce our approximation of human decision-making mechanism to incorporate research topics on human interaction behaviors with robots in Sec. VI; lastly, we use the proposed model to simulate human-robot navigation with path crossing.

V. Human-Robot Game in Real-time

To formally discuss the coordinating process between humans and robots, we define the real-time game setting through game start and termination criteria. When players sense the game has started, they plan by taking other agents' behaviors into consideration, until they sense game termination.

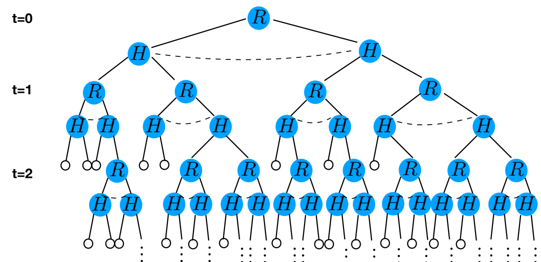

Fig. 2. Two-player human-robot interaction, modeled as a simultaneous-move game.

1) Game start criteria: In a scenario where agents' actions affect the outcome of each other, we define as it meets the criterion of *potential interaction*, $\mathcal{C}_{PI}$; when agents become aware of each other and start acting according to the perceived strategy of the other agents, it meets the criterion of *mutual awareness*, $\mathcal{C}_{MA}$. Once the two criteria hold, the game begins.

2) Game horizon and termination criteria: When interactions start at time $t = 0$, the game repeats *finitely* until a termination criteria $\beta : X \times \Sigma \rightarrow \{True, False\}$ is satisfied. It may be a predefined criteria, $t > T$, where T is a prespecified time frame of mandatory co-working. Termination criteria may also take in a dynamic format, such as to end the game whenever $\mathcal{C}_{PI}$ no longer holds. An example would be crowd interactions: when pedestrians are past the route intersections with one another, the game terminates; as no further intervention is expected.

With the above definition of game start and termination, the overall game, in a simultaneous-move fashion, is shown in Fig. 2 as an extensive-form game.

A. Simultaneous-move game tree

The tree contains the following:

1) Decision nodes: the solid circles where players make choices, $a^i_t \in A^i$, based on current state x_t
2) Terminal nodes: the hollow circles at the bottom where game outcomes $V(x_T, \sigma_T, \theta_T)$ are assigned
3) History set: the observed history plays before current time I^i_t. decision nodes connected with a dashed line share the same history set; players cannot distinguish nodes with the identical history sets. Therefore, in games with simultaneous moves, players share the same history set in one period.

The policy of how an agent i makes the high-level decision can be abbreviated as a function of his or her type θ^i_t,

$$a^i_t \sim \pi^i(x_t, \theta^i_t|\sigma_{0:t-1}), \tag{9}$$

without loss of generality.

More specifically, at each decision node, the player chooses an action based on current state x_t, and the policy, as to optimize Eq. 8, is to compute the following at each decision node:

1) $p(\sigma_t|I^i_t, x_t)$: game action profile probability given history set
2) r_t or V_T: reward estimate, or value estimate at termination nodes

Note that, to compute the game outcome, here we use either cumulative reward, as commonly seen in Markov Decision

Process, or terminal value, as commonly seen in extensive-form games. The form of game outcome has no effect on the solution algorithm.

Due to the continuous-space state space formulation, backward induction, a common solution for extensive-form games, is no longer applicable. Instead, forward-search approaches such as Monte Carlo Tree search can be applied: at each iteration, randomly sample an action a^i_t and then expand the search tree by sampling the action profile $p(\sigma^t|I^R_t, x_t)$. To compute the reward of a stage game r_t, or the value at termination nodes V_T, Eq. 6 can be applied by sampling u_t from Eq. 4, given prior on other agents' types $p(\theta^{-i}_t)$.

B. Online game strategies

When planning for long-horizon purposes, agents ideally want to optimize their outcome considering full-horizon accumulation, as introduced in Eq. 8. However, due to the high uncertainty in dynamic environments, pre-computed solutions may not fit well to newly received observation data, which affects agent's inference of the other agent: their future action profiles, their action realizations, or state transitions. Therefore, when playing in real-time, agents need constant online re-planning to adapt to unmodeled dynamic situations.

With that said, instead of running search algorithms to solve for the total horizon at $t = 0$, agents run *belief updates* whenever new observations arrive, and *replan for certain horizon* from current time t, for as much lookahead horizon as computational resource allows. We assume agents have knowledge of the termination timing even in the dynamic setting β_t.

1) belief updates: During real-world interactions, observations can be from direct measures, such as the relative positions and velocities of all agents. Observations can also be implicit as to infer private messages, through ways like eye contacts or body languages, to express messages such as intent [14].

In either form of observations $o_{0:t}$, for long-horizon planning on Markov games with time-variant types, agents update their beliefs of the following:

(a) type estimate of other agents $b^i_t(\theta^{-i}_t)$: $p(\theta^{-i}_t|o_{0:t})$, to sample state transitions $\mathcal{T}$ and compute reward function r_t

(b) future strategy profiles of other agents $b^i_t(\sigma_{t:T})$: $p(\sigma_h|I^i_h)$, where $t \leq h \leq T$ is the future time step of interest

Belief updates are usually computationally cheap, so agents may run it at a high rate (potentially higher than replan frequency) to deal with noisy observations.

2) Receding-horizon planning: Due to the potential demand of fast re-planning in dynamic environments, agents may only be computationally available to plan for finite-step lookahead $H < T$. Therefore, at each time step $t = h, h < T$, agents instead try to optimize for $t = h : H'$, where $H' = min(h + H, T)$, based on updated model of future strategy anticipation of other agents' $p(\sigma_{h:H'}|I^i_h)$. This online replanning for finite horizon strategy, known as receding-horizon planning, can be formulated as the following:

$$a^{i*}_{h:H'} = \underset{a^i_{h:H'}}{\operatorname{argmax}} \Sigma^{H'}_{t=h} \mathbb{E}_{\theta_t,\sigma_t,x_t|\mathcal{T},I^i_t}\left[r^i(x_t, \sigma_t, \theta_t)\right] + \hat{V}^i_{H'+1}(x_H, \sigma_H, \theta_H), \tag{10}$$

where $\hat{V}_{H+1}(x_H, \sigma_H, \theta_H)$ is the cost-to-go estimate for following σ_H from $t = H + 1$ to T.

For coordination games, it is usually true that the earlier the termination, the better the final outcome is: similar to games with benefit discounts. Take the table turning task for example, the faster the two agents reach agreement on the direction to go, the faster progress they make, despite the potentially longer routes. Therefore, biased search to strategies with early termination has its empirically advantage; agents can even trade off computation depth with breadth to better explore its action profile.

VI. Human Behavior and Decision-Making Model

As one of the main purposes of our proposed framework is to model *human* interaction with AI agents, here we introduce our hypotheses on human behaviors and their decision-making mechanism; by implementing these hypotheses into the framework, we have can better analyze how different paradigms in human-robot coordination affect the overall convergence.

A. Adaptability to other agents

Humans observe other agents and adapt their strategies accordingly [22], [32]. In the proposed framework, an agent's type θ^H_t is parameterized by a static parameter z^H, which is associated with his or her personal preferences that do not change in short period of time, and by a *time-variant* parameter $b^H_t(\theta^{-H}_t)$, to capture the agent's dynamic *beliefs* of other agents' types. More specifically,

$$\theta^H_t = \begin{bmatrix} z^H \\ b^H_t(\theta^{-H}_t). \end{bmatrix} \tag{11}$$

z^H, the static type parameter, features personal preferences such as travel efficiency versus travel energy in navigation domains. $b^H_t(\theta^{-H}_t)$, the time-variant parameter, features human's perception on other agents' types. As for the detailed form in the belief, we hypothesize that humans possess certain *information budget* to reason about other agents' behaviors.

Below, we propose our hypothesis on the human's information budget, which corresponds to their computational capability to infer about other agents' decision-making model. The decision-making model is the policy that outputs the agent's high-level action(s):

$$a^i_{t:t+H} \sim \pi^i(x_t, \theta^i_t) \tag{12}$$

1) $b^i_t(\theta^{-i}_t) = \emptyset$: Agents keep no information of the other agents' behavior; they parametrize their policies solely based on personal preferences, but make no use of other agents' behaviors to characterize the game outcome. We assume human behaviors to not belong to this category while interacting with robots: $b^H_t(\theta^{-R}_t)$.

2) $b^i_t(\theta^{-i}_t) = \hat{z}^{-i}$: Agents assume the other agents maintain no information of themselves, but only act according to their static preferences $\theta^{-i}_t = z^{-i}$. Therefore, agents adapt to the others as if their own actions have no impact on other agents' decision-making model: $p(a^{-i}_{t+H}|x_t, z^{-i})$. This is referred to as the *one-layer* inference.

3) $b_t^i(\theta_t^{-i}) = [\hat{z}^{-i}, \hat{\hat{z}}^i]$: Agents assume the other agents are also adaptive to themselves, $\theta_t^{-i} = [z^{-i}, \hat{z}^i]$, with one-layer inference. Therefore, when planning for more than one period, agents act adaptively, at the same time evaluating their actions' potential impacts on the other agents' future strategies. When planning for one period, Agents have the budget to compute two-layer inference: to plan according to what they predict the others' predictions about themselves $a_t^i \sim \pi^i(x_t, b_t^i(\pi^{-i}(x_t, b_t^{-i}(\pi^i(x_t, \hat{\theta}^i)))))$. This is the maximum budget we assume humans can afford for real-time inference, and is not applied for planning due to its intrinsic complexity.

Therefore, with the information budget assumption $b_t^H(\theta_t^H) = \hat{z}^{-H}$, we consider human policies being adaptive to their perceived, presumably static, behavior of other agents. The higher adaptation rate is, the more flexible they appear in the joint policy.

B. Bounded memory belief updates

As introduced in Sec. V-B.1, we also assume humans maintain their beliefs of other agents' types as well as their future strategy profile through out online planning. Here, we further assume humans to either run Bayesian updates, or possess *bounded* memory on past observations and interaction history for belief updates[22]: $b_t^H(\sigma_t|I_{t-(t-n)}^i$, and $b_t^H(\theta_t^{-i}|o_{t-n:t})$.

C. Finite-step lookahead

As introduced in Sec. V-B.2, we also assume humans to replan online with finite-step lookahead.

1) 0-step lookahead, H=0: Agents act as if the current game is the termination game.

2) multi-step lookahead, H>0: When agents plan as the game has more than one period, adaptive behaviors due to belief updates are expected [22].

D. Anticipation of other agents' policy

As pointed out in Sec. VI-A, humans adapt their policies based on their beliefs of other agents' behavior. When planning at time t, agents predict about other agents' action profile $\sigma_{t:t+H}$ based on past interactions:

$$a_t^H \sim \pi^H(x_t, b_t^H(a_t^{-H}|\sigma_{0:t-1})). \tag{13}$$

To anticipate the other agents' actions $a_t^{-H}|s_{0:t}$, one may assume their policies are non-adaptive, $a_t^{-H} \sim \pi^{-H(x_t, z^{-H})}$, associated with the one-layer inference. which is of two-layer inference for planning. In this paper, we discuss human decisions based on only one-layer inference.

VII. Explaining influential factors in human-robot interaction

As pointed out in Sec. VI, human policies for high-level decisions are parametrized by θ_t^i, which include that person's static type z^H and dynamically perceived types $b_t^H(\theta_t^{-H})$ of other agents. With different perceptions on the other, humans act differently. For example, pedestrians take over others' roads when they are in a hurry; however, they yield when encountering the elderly. Similar situation applies to human-robot interaction: people have distinctive behaviors based on their prior assumptions of robots, which affect their policies and their adaptabilities when new observations come in.

Here we discuss how to use the proposed model to explain phenomena in human-robot interactions.

A. Static preferences

1) Personal preferences: As people have different perception and long-time experience in interactions with the environment, they preserve distinctive characteristics in their behaviors that do not change in short period of time. Therefore, when planning considering joint actions, robots should be aware of such types to plan accordingly for agent comfort and overall efficiency [11]. In our proposed framework, personal preferences contribute to agents' policy realizations, transition functions, and affect the joint performances. In the reward function, they can be characterized as feature weighting y^i [5]:

$$r^i = -y^{iT} C, \tag{14}$$

where C is some vector of cost function.

2) Level of self-interest: When agents are deployed in public environments, the notion of public welfare plays in to assess policy fairness [8]. While the public welfare is the self-interest in collaborative tasks, in non-cooperative games, agents have incentives to deviate from cooperative behaviors for personal benefits [10]. When individuals plan in a shared workspace with personal objectives to achieve, resource conflicts may occur. While cooperative policies are the most efficient for social welfare, agents may gain more resource allocation when playing selfishly. This notion of fairness can be characterized as weighting on all parties' interest α^i:

$$r^{i'} = \alpha^i r^i + \frac{1-\alpha^i}{k-1} \sum_{-i} r^{-i}. \tag{15}$$

Level of self-interest α^i and personal preferences y^i jointly contribute to the static preferences z^i.

B. Perceived robot capability

The gap between true robot capability and human perceived robot capability has shown to deteriorate both joint and individual work efficiency in different task domains [6].

Here, we characterize human perceived robot capability for human-robot interaction into two categories: 1) functional capability and 2) social inference capability:

1) Functional capability: This includes the belief of whether the robot is able to *identify interaction*. Before engaging interactions, agents need to ensure that $\mathcal{C}_{PI}$ and $\mathcal{C}_{MA}$ are met by all parties, or confusions may arise. Due to lack of social signaling capability, such as gazes, humans may be uncertain whether the robot is aware of the potential interactions. This may result in distantly-avoiding behaviors through out the interaction, since people are not sure whether to engage [6]. This also includes the knowledge of robot action set A^R, and the confidence of whether the robot is able to *succeed in its target actions* a_t^R, especially when complex domains are considered [3].

2) Social inference capability: This involves whether they perceive the robot to be capable of engaging the game with, where implicit communication may involve to interact [14]. This includes the ability to *identify human's intended high-level actions* based on the inference from past observations,

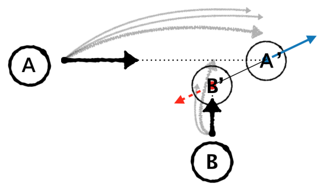

Fig. 3. Collision avoidance actions based on closest position estimate. Red dashed arrow (social force direction): yielding agent with later arrival timing at the intersection; blue solid arrow: passing-first agent.

and to *express its own intended action* in a clear, context-aware, or, legible manner [7]. Failing to present such capabilities out of human expectations deteriorates human patience and overall efficiency on given tasks [2].

The above two perceived capabilities are prerequisites for humans to engage in the coordination process with robots. In complex domains, those criteria have shown to be challenging [14], therefore prior experiences on cross-training, teaching, and learning [34] may be required to prepare for natural engagement.

C. Social trust

Perceived capabilities are preliminary for human trust to interact with robots [32], since the knowledge of the action set of the robot enables human prediction of robot future actions. When there are resource conflicts in shared workspaces, self-interested agents may not be socially compliant to the other agents. In this case, trust on social collaborativity in conflicted situations, captured by human's belief of robot static type $b_t^H(z^R)$, affects human policies through their anticipation of robot policy. While perceiving robots as socially trust-worthy agents, humans predict robots to have non-hostile behaviors and may cooperate at ease.

VIII. Problem Instantiation

We instantiate the framework on 2-player human-robot navigation with path crossing, shown in Fig. 1.

A. Intention-aware social navigation

Great progress has been made in the past two decades in pedestrian simulations [13], [33], robot navigation with human predictive models [29], [16], and socially-friendly robot planning [21], [4], to smoothly deploy robots in human workspaces. In human-robot interaction community, robot motion legibility has gained attention for human understanding [7], and the concept of trajectory interpretability has been applied on robot navigation in crowded environments to identify goals of pedestrians [1], [30].

Here we make use of the social force model with explicit collision prediction [33], which was inspired by the social-force model but equipped with explicit avoidance behaviors. Other choices of human behavioral model for action realizations $g(x_t, a_t^H, \theta_t^H)$ can also fit in the framework for simulation; we choose this model for its underlying motion *legibility* concept in distinguishing choices of avoidance behaviors, shown in Fig. 3.

When two agents have intersecting paths and have similar timing to arrive at the intersection, explicit avoidance behaviors are involved. Particularly, an agent has to decide whether to avoid in front or behind the other. This leaves the coordination to agree on two action combinations: (passing_first,yielding), or (yielding,passing_first).

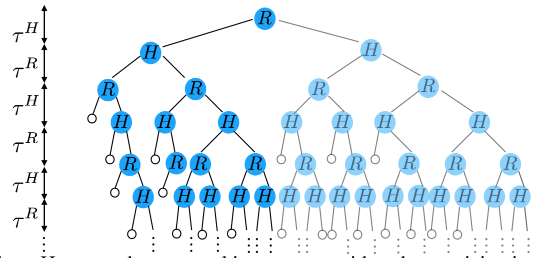

Fig. 4. Human-robot turn-taking games with robot anticipating action value based on finite-horizon predictions. τ^R and τ^H are approximate time intervals for robot and human response times. The game may terminate at either player's decision nodes.

We define the state space as the following: $x_t^R = \begin{bmatrix} p_t^R \\ v_t^R \end{bmatrix}$, and $x_t^H = \begin{bmatrix} p_t^H \\ v_t^H \end{bmatrix}$ where p_t^R, p_t^H are the positions of the robot and human, and v_t^R, v_t^H are the velocities of the robot and human, respectively. All variables are described in 2D, therefore, $p_t^R, p_t^H, v_t^R, v_t^H \in \mathbb{R}^2$.

B. Game start and termination criteria

As pointed out in Sec. V, to initiate interactive behaviors between agents, the potential-interaction criteria $\mathcal{C}_{PI}$ and mutual-awareness criteria $\mathcal{C}_{MA}$ are prerequisites to start the game. In path crossing, $\mathcal{C}_{PI}$ is met if:

1) two agents have path intersections in near future: $\exists t^H < t^{rH}, t^R < t^{rR}$, such that $v_t^R \times t^R + p_t^R = v_t^H \times t^H + p_t^H$
2) the arrival timing difference is within certain threshold: $|t^H - t^R| < \delta$

Here, the reaction time t^{rH} is the timing that agents decide to engage in the scenario. For pedestrian avoidance, experimental results suggest that t^{rH} is on average 4 secs before reaching the intersection [25]. The arrival timing difference depends on agent velocities, the safety margin to keep from other pedestrians, and the noise in estimation. Here, we use an estimate value of 1.5 sec, which is the upper bound on time difference for which agents respond to the intersecting scenario.

The game terminates whenever two agents have passed each other, or the minimum relative distance has passed.

C. Intentions in path crossing and time delay

The action set is defined as the following: $A^R = \{a^{pf}, a^y\}$, $A^H = \{a^{pf}, a^y\}$, where a^{pf} is corresponding to the class of trajectory realizations to pass first, in front of the other agent; whereas a^y is corresponding to the class to yield to the other.

Empirical study on human-robot crossing has suggested distinctive velocity and trajectory profiles among two classes of avoidance actions [24]. Agents who intend to pass first often accelerate and bend their trajectories away from the other; agents who intend to yield often slow down. Such changes in motion profiles are clear signals for agent passing intents, and one can observe responses within short period of time, around 0.7 sec.

The minimum time for human agents to react to their action changes, τ^H, is here assumed to be between 0.7s to 0.9s. For more complex domains, higher values should be considered.

D. Response-time-adapted turn-taking structure

While the interaction process is modeled as a simultaneous-move game in Sec. V, with the time delays in action realization, it is often played as a *turn-taking* game. With proper waiting time after the initial action, the decision timing of both agents can be clearly separated to prevent oscillations. Despite the nature in the simultaneous-move setting for interactions, turn-taking games simplify the simulation setting and therefore save computation. The overall turn-taking game used in this instantiation is illustrated in Fig. 4.

E. Experiments

We hypothesize that our proposed framework can explain different types of avoidance behaviors when their paths intersect with robots. We use our framework to simulate a particular type, cautious behavior, in comparison with collected trajectory data, to show that the framework can capture distinctive interaction patterns among people. We deploy a mobile robot into a building atrium, and record human crossing trajectories using tracking packages on laser scans [18]. Eight participants were requested to head towards a set of goals, while the robot followed a given route that intersected with human eight times. The robot was implemented with a local planner with emergent slow-down when sensing an object within 1 meter.

With the same prior on robot strategy profile, with 0-step lookahead, virtual pedestrians act compliantly after observing the other agent's action, due to the assumption of immediate termination. With 1-step lookahead, agents assume the game may end with the robot playing minimax strategy, therefore they play based on conservative value estimate. With 2-step lookahead, virtual pedestrians are capable of planning for the optimistic (2nd step) in the worst case robot strategy (1st step). Therefore the behavior is less conservative compared to the previous two options.

We simulate virtual pedestrians with type $\theta_t^H = [z^H, b_t^H(z^R)]$, 2-step lookahead, memory bound on interaction history = 2, in the turn-taking formulation. This is associated with the decision-making capacity to reason: "what would the robot do after seeing my action, and what I could do to in reaction to that action?". Other combinations can be considered. We have the robot choose avoidance actions purely based on arrival timing estimates. It is a simple model for demonstration clarify on human responses; more complex models can be used.

1) Perceived capabilities: Complete set of perceived capabilities are specified in Sec. VII-B as prerequisites for humans to engage in the coordination process with robots. For social navigation, when people see a robot coming, they may be unsure whether 1) the robot sees them (of $\mathcal{C}_{MA}$), 2) the robot is aware of the potential collision (of $\mathcal{C}_{PI}$), 3) when they choose an avoidance action, the robot can identify the underlying intention (of social inference capability on perception), and 4) when the robot avoid, humans can identify the underlying intention (of social inference capability on motion generation).

2) Types in path crossing: Agents perceived other agents' behaviors and update their own type θ_t^H. In social navigation, personal spacing is a common example, as people act repellently more to unfamiliar agents. People's urgency to travel to their own goals are common factors to describe navigation preferences. In simulation, we simulate agents to choose avoidance actions based on the following cost formulation:

$$r_t^H = -\eta\alpha C_t^H - (1-\alpha)C_t^{-H}, \tag{16}$$

where C_t^H is the estimated time delay, $\eta \in [0,1]$ is the urgency level, and $\alpha \in [0,1]$ is the level of self-interest.

3) Social trust: When navigation around a robot, the assumption of whether it is socially compliant affects human's avoidance decision. We commonly observe, at the beginning of the experiment, participants slow down to wait for the robot to pass first, even when the robot has a later arrival timing; people speed up and stay far from the robot when trying to pass in front, even when the robot is still far from the intersection. We refer to this type of behavior as cautious, simulated in Fig. 5 through varying the perceived robot level of self-interest: $b_t^H(z^R)$.

4) Human exploration and adaptation: We observed many conservative decisions on human crossing, where pedestrians constantly yield to the robot. There was one agent who tried to pass in front of the robot after a couple of times of crossing and then decided to pass first. We refer this process as the agent was gradually gaining social trust on the robot; she updated her belief of robot behavior and then adapted her own strategy. Similar behavior was simulated on virtual human with high belief update rate, shown in Fig 6.

IX. Conclusion and Future Work

We propose a Markov Game model with time-variant types to analyze human-robot coordination outcome, and propose a human decision-making model to describe phenomena in human-robot interactions. With the framework, we simulate virtual pedestrians with cautious type behaviors on human-robot crossing and compare the results with real-world recording. We also simulate adaptive human behaviors through belief updates on robot policy. In future work, we will further investigate human adaptability criteria based on different prior assumptions on robot behaviors and game convergence given false human assumptions.

References

[1] Haoyu Bai, Shaojun Cai, Nan Ye, David Hsu, and Wee Sun Lee. Intention-aware online pomdp planning for autonomous driving in a crowd. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 454–460. IEEE, 2015.

[2] Elizabeth Cha, Anca D Dragan, and Siddhartha S Srinivasa. Perceived robot capability. In *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*, pages 541–548. IEEE, 2015.

[3] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. Planning with trust for human-robot collaboration. *arXiv preprint arXiv:1801.04099*, 2018.

[4] Yu Fan Chen, Michael Everett, Miao Liu, and Jonathan P How. Socially aware motion planning with deep reinforcement learning. *arXiv preprint arXiv:1703.08862*, 2017.

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

[6] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 51–58. ACM, 2015.

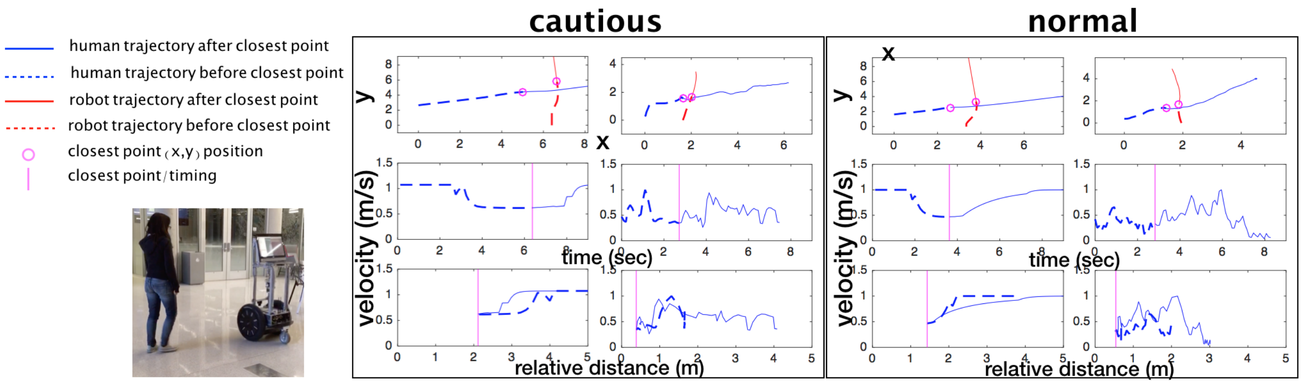

Fig. 5. SIMULATED pedestrian yielding, compared with REAL recording. We observed many types of behaviors during the experiments, and illustrate one common type in human-robot crossing: cautious, shown in the left box. Compared with people with gradual slow-down and speed-up (noted as "normal", shown in the right box), cautious agents wait for the robot until it passes the intersection (shown in the bottom-left photo). We simulate agent yielding and compare the x-y trajectory (top) and velocity profile over time (middle) and relative distance (bottom) with the true recordings, right-side of each box.

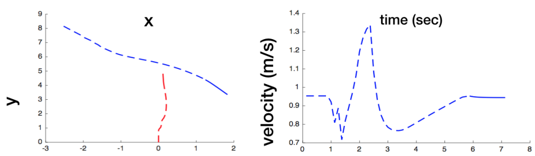

Fig. 6. SIMULATED cautious agent updating belief on robot strategies. While the simulated pedestrian yields at the beginning (with -0.7 sec arrival timing difference), after observing robot yielding behavior, she updates her belief and changes the action in the next time frame.

[7] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, pages 301–308. IEEE, 2013.

[8] Ernst Fehr and Urs Fischbacher. Social norms and human cooperation. *Trends in cognitive sciences*, 8(4):185–190, 2004.

[9] Jakob N Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. *arXiv preprint arXiv:1709.04326*, 2017.

[10] Takako Fujiwara-Greve. *Non-cooperative game theory*, volume 1. Springer, 2015.

[11] Matthew C Gombolay, Cindy Huang, and Julie A Shah. Coordination of human-robot teaming with human task preferences. In *AAAI Fall Symposium Series on AI-HRI*, volume 11, page 2015, 2015.

[12] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[13] Ioannis Karamouzas, Peter Heil, Pascal van Beek, and Mark H Overmars. A predictive collision avoidance model for pedestrian simulation. In *International Workshop on Motion in Games*, pages 41–52. Springer, 2009.

[14] Ross A Knepper, Christoforos I Mavrogiannis, Julia Proft, and Claire Liang. Implicit communication in a joint action. In *Proceedings of the 2017 acm/ieee international conference on human-robot interaction*, pages 283–292. ACM, 2017.

[15] Thibault Kruse, Patrizia Basili, Stefan Glasauer, and Alexandra Kirsch. Legible robot navigation in the proximity of moving humans. In *Advanced Robotics and its Social Impacts (ARSO), 2012 IEEE Workshop on*, pages 83–88. IEEE, 2012.

[16] Markus Kuderer, Henrik Kretzschmar, Christoph Sprunk, and Wolfram Burgard. Feature-based prediction of trajectories for socially compliant navigation. In *Robotics: science and systems*. Citeseer, 2012.

[17] Minae Kwon, Malte F Jung, and Ross A Knepper. Human expectations of social robots. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 463–464. IEEE Press, 2016.

[18] Angus Leigh, Joelle Pineau, Nicolas Olmedo, and Hong Zhang. Person tracking and following with 2d laser scanners. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 726–733. IEEE, 2015.

[19] Christina Lichtenthäler, Tamara Lorenzy, and Alexandra Kirsch. Influence of legibility on perceived safety in a virtual human-robot path crossing task. In *RO-MAN, 2012 IEEE*, pages 676–681. IEEE, 2012.

[20] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*, pages 157–163. Elsevier, 1994.

[21] Christoforos I Mavrogiannis and Ross A Knepper. Decentralized multi-agent navigation planning with braids. In *Proceedings of the Workshop on the Algorithmic Foundations of Robotics. San Francisco, USA*, 2016.

[22] Stefanos Nikolaidis, Anton Kuznetsov, David Hsu, and Siddharta Srinivasa. Formalizing human-robot mutual adaptation: A bounded memory model. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 75–82. IEEE Press, 2016.

[23] Stefanos Nikolaidis and Julie Shah. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 33–40. IEEE Press, 2013.

[24] Sébastien Paris, Julien Pettré, and Stéphane Donikian. Pedestrian reactive navigation for crowd simulation: a predictive approach. In *Computer Graphics Forum*, volume 26, pages 665–674. Wiley Online Library, 2007.

[25] Julien Pettré, Jan Ondřej, Anne-Hélène Olivier, Armel Cretual, and Stéphane Donikian. Experiment-based modeling, simulation and validation of interactions between virtual walkers. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 189–198. ACM, 2009.

[26] Mark Pfeiffer, Ulrich Schwesinger, Hannes Sommer, Enric Galceran, and Roland Siegwart. Predicting actions to act predictably: Cooperative partial motion planning with maximum entropy models. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 2096–2101. IEEE, 2016.

[27] Julie Shah, James Wiken, Brian Williams, and Cynthia Breazeal. Improved human-robot team performance using chaski, a human-inspired plan execution system. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 29–36. ACM, 2011.

[28] Masahiro Shiomi, Francesco Zanlungo, Kotaro Hayashi, and Takayuki Kanda. Towards a socially acceptable collision avoidance for a mobile robot navigating among pedestrians using a pedestrian model. *International Journal of Social Robotics*, 6(3):443–455, 2014.

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

[30] Vaibhav V Unhelkar, Claudia Pérez-D'Arpino, Leia Stirling, and Julie A Shah. Human-robot co-navigation using anticipatory indicators of human walking motion. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 6183–6190. IEEE, 2015.

[31] Jur Van Den Berg, Stephen J Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In *Robotics research*, pages 3–19. Springer, 2011.

[32] X Jessie Yang, Vaibhav V Unhelkar, Kevin Li, and Julie A Shah. Evaluating effects of user experience and system transparency on trust in automation. In *HRI*, pages 408–416, 2017.

[33] Francesco Zanlungo, Tetsushi Ikeda, and Takayuki Kanda. Social force model with explicit collision prediction. *EPL (Europhysics Letters)*, 93(6):68005, 2011.

[34] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. Plan explicability and predictability for robot task planning. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 1313–1320. IEEE, 2017.