Learning-Aware Safety for Interactive Autonomy

Anonymous Author(s)
Affiliation
Address
email

Abstract: One of the outstanding challenges for the widespread deployment of robotic systems like autonomous vehicles is ensuring safe interaction with human users and stakeholders without sacrificing efficiency. Existing safety analysis methods often neglect the robot's ability to learn and adapt at runtime, leading to overly conservative behavior. This paper proposes a new closed-loop paradigm for synthesizing safe control policies that explicitly account for the system's evolving uncertainty under possible future scenarios. The formulation reasons jointly about the physical dynamics and the robot's learning algorithm, which updates its internal belief over time. We leverage adversarial deep reinforcement learning (RL) for scaling to high dimensions, enabling tractable safety analysis even for implicit learning dynamics induced by state-of-the-art prediction models. We demonstrate our framework's ability to work with both Bayesian belief propagation and the implicit learning induced by a large pre-trained neural trajectory predictor.

Keywords: Learning-Aware Safety Analysis, Behavior Prediction, Active Information Gathering, Adversarial Reinforcement Learning.

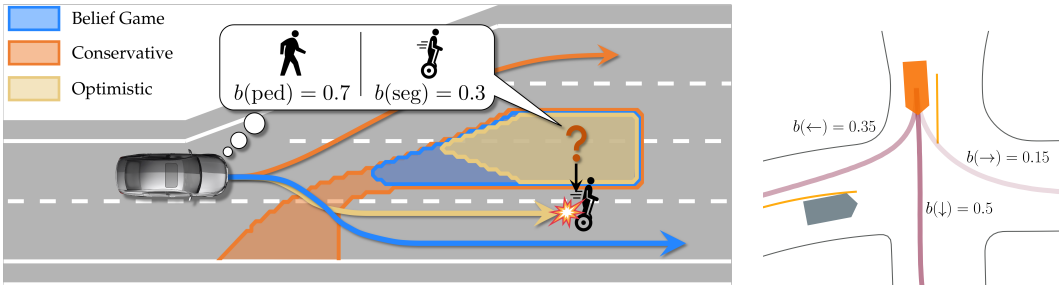
1 Introduction

The deployment of robotic systems in human-populated environments has increased in recent years as the capabilities of these technologies have improved. These systems must always maintain safety when working close to humans. However, overly conservative decision-making can lead to undesirable "deadlock" situations, such as the freezing robot problem [1], negatively impacting robots' ability to complete intended tasks. Therefore, it is of the utmost importance to balance safety and efficiency for intelligent robotic systems such as autonomous cars when interacting with other agents.

A major technical challenge arises due to *interaction uncertainty*—uncertainty induced by others' intents (e.g., does a human driver want to merge, cut behind, or stay in the lane?), responses (e.g., if the robot accelerates, how will the human react?), and semantic class (e.g., is this person a pedestrian or cyclist?). These aspects are all unknown *a priori*, so the robot must learn about them at runtime. As the robot and the human take action over time, the robot gathers more information about the human's internal state and can therefore update its uncertainty to make more effective decisions. Unfortunately, existing safety methods predominantly ignore this online learning component and instead make static modeling assumptions, unnecessarily constraining the robot's behavior. Specifically, their safety analysis only accounts for the evolution of the physical system state while assuming that the information available to the robot remains constant throughout the planning/prediction horizon. This simplification can often lead to overly conservative robot decisions and—in extreme cases— catastrophic safety failures. In this work, we argue that a key step towards realistic safety analysis methods is for safe control policies to be *explicitly aware* of how interaction uncertainty evolves *during the robot's operation*.

Related work. A popular safety analysis approach in multi-agent robot decision-making is to compute *robust predictions* of other agents' behavior. Forward reachability methods such as [3, 4, 5] use

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.



(a) Road crossing scenario (running example). **(b)** Deep trajectory prediction.

Figure 1: Our proposed Belief Game framework improves safety and efficiency of uncertain, multi-modal interactions with other agents by closing the safety decision loop around the robot's runtime learning algorithm. *Left:* An autonomous vehicle is uncertain of the type of a human crossing the road. Our method computes a safe, yet non-conservative control policy by closing the safety loop around the robot's runtime learning algorithm. *Right:* Our framework works with modern data-driven trajectory predictors such as the Motion Transformer [2].

(possibly learned) predictive agent models alongside principled Hamilton-Jacobi (HJ) reachability analysis to produce robust predictions at each runtime planning cycle. Another major category of safety methods focuses on synthesizing *robust control policies* for the robot, which encode how the robot should react to possible behaviors of the other agents. Backward reachability analysis naturally captures the interaction as a zero-sum game [6] whose conservativeness can be modulated by adapting the set of admissible human controls based on observations at runtime [7].

The above methods implicitly assume a *static* understanding of the human's internal state during the entire safety analysis horizon. In reality, however, the robot will gain information about the other agent's intent, semantic class, etc., throughout the interaction. The robot may reduce conservativeness by reasoning the *future information* it might receive. For example, in the context of navigation in occluded environments, safety can be maintained less conservatively by leveraging the future observability of hypothetical yet-undetected objects [8, 9]. Additionally, applying forward reachability analysis in the robot's belief space allows predicting the worst-case time it may take to reveal a human agent's intention, allowing for safer contingency planning [10].

High-dimensional reachability analysis has traditionally been intractable for systems with more than five continuous state variables due to the need to numerically solve a partial differential equation. Recent developments have shown promise in scaling safety analysis to high-dimensional systems using neural approximations: deep self-supervised learning takes the HJ PDE as a learning target [11], whereas deep reinforcement learning (RL) approaches use the reachability payoff as a reward signal for iterative control policy improvement [12, 13].

Contributions. We propose a novel *Belief Game* safety analysis framework that closes the loop between the robot's prediction-planning-control pipeline and its runtime learning process, by *jointly* reasoning over the physical state and the robot's belief state (Figure 1). We scale our approach to high-dimensional problems by leveraging a recently-introduced approximate solution [13] to zero-sum dynamic games via model-free adversarial RL—enabling tractable safety analysis even for implicit learning dynamics induced by neural prediction algorithms. We demonstrate our framework's effectiveness in conjunction with a standard Bayesian filter and the black-box learning dynamics of the Motion Transformer [2], a highly proficient state-of-the-art multi-modal trajectory predictor.

2 Preliminaries and Problem Formulation

Multi-agent dynamical system. We consider a class of discrete-time nonlinear dynamics encoding the interaction between an "ego" autonomous robot (e) and one or multiple "opponent" agents (o):

$$x_{t+1} = f(x_t, u_t^e, u_t^o, w_t), \tag{1}$$

where $x_t = (x_t^e, x_t^o) \in \mathcal{X} \subseteq \mathbb{R}^n$ is the joint state vector, $u_t^e \in \mathcal{U}^e \subset \mathbb{R}^{m_e}$ and $u_t^o \in \mathcal{U}^o \subset \mathbb{R}^{m_o}$ are the control vectors of the robot and the opponent, respectively, $w_t \in \mathcal{W} \subset \mathbb{R}^{m_w}$ is a bounded

process noise (external disturbance), and $f : \mathcal{X} \times \mathcal{U}^e \times \mathcal{U}^o \times \mathcal{W} \to \mathcal{X}$ is a continuous nonlinear function that describes the physical system dynamics.

Background: Hamilton-Jacobi reachability. Hamilton-Jacobi (HJ) reachability analysis [6, 14] uses robust optimal control theory to compute provable policies to reach and/or avoid regions of the state space. Define the target and failure sets as $\mathcal{T} := \{x \mid \ell(x) \leq 0\} \subseteq \mathbb{R}^n$ and $\mathcal{F} := \{x \mid g(x) > 0\} \subseteq \mathbb{R}^n$, encoded by Lipschitz-continuous *margin functions* $\ell(\cdot)$ and $g(\cdot)$. Hamilton-Jacobi-Isaacs (HJI) analysis captures the opponent's worst-case behavior by formulating an infinite-horizon zero-sum dynamic game. In discrete time, the game's solution can be obtained via the fixed-point Isaacs equation [15] (the game-theoretic counterpart to the Bellman equation):

$$V(x) = \max\left\{g(x), \min\left\{\ell(x), \min_{u^e \in \mathcal{U}^e} \max_{u^o \in \mathcal{U}^o, w \in \mathcal{W}} V\big(f(x, u^e, u^o, w)\big)\right\}\right\}. \tag{2}$$

The value function $V(\cdot)$ encodes the reach-avoid set $\mathcal{RA}(\mathcal{T}, \mathcal{F}) := \{x \mid V(x) \leq 0\}$, from which the ego agent is guaranteed a control strategy to safely reach the target set without entering the failure set. The optimal ego and opponent policies are obtained as byproducts of solving (2).

Opponent uncertainty. We model opponent agents as taking actions from an *uncertain control set* affected by their unknown individual characteristics $\theta \in \Theta$, which we denote their *type* following the game theory literature. This type θ may encode aspects about the agent, including its *semantic class* (e.g., a pedestrian or a cyclist) and its *internal state* (e.g., intended destination, attention state). We assume that the hypothesis space Θ is finite, and that each hypothesized type θ is associated with a known (possibly state-dependent) control set $\mathcal{U}^o_\theta(x_t)$ of *admissible* actions for an opponent agent of this type.

Reducing uncertainty via learning. The robot represents its uncertainty over the agent type $\theta \in \Theta$ through a probabilistic *belief* $b(\theta) \in \Delta$. Since Θ is a finite set, b is a *categorical distribution* and the belief space Δ is a $(|\Theta| - 1)$-simplex. Our approach is agnostic to the specific runtime learning algorithm used for propagating the belief state, which includes, among others, Bayesian inference and transformer neural networks [2]. We, therefore, consider general *learning dynamics* of the form

$$b_{t+1} = f_L(b_t, y_t), \qquad y_t = h(x_t, u^o_t, v_t), \tag{3}$$

where $y_t \in \mathcal{Y} \subseteq \mathbb{R}^{n_y}$ is the robot's indirect observation of the opponent at time t from state x_t, $v_t \in \mathcal{V} \subset \mathbb{R}^{m_v}$ is a bounded measurement noise, and $h : \mathcal{X} \times \mathcal{U}^o \times \mathcal{V} \to \mathcal{Y}$ is the observation model. Defining the joint state vector $z := (x, b)$, we can combine (1), (3) into a joint dynamical system:

$$z_{t+1} = F(z_t, u^e_t, u^o_t, w_t, v_t) := \begin{bmatrix} f(x_t, u^e_t, u^o_t, w_t) \\ f_L\big(b_t, h(x_t, u^o_t, v_t)\big) \end{bmatrix}, \tag{4}$$

which compounds the physical evolution of the system with the robot's belief update. We note that the learning dynamics need not impose (or be aware of) the type-specific control bounds $\mathcal{U}^o_\theta(x_t)$.

Running example: In Figure 1, the autonomous vehicle (ego) is uncertain about whether the crossing human (opponent) is a pedestrian or a segway rider. The ego and opponent are modeled as a 5D kinematic bicycle model and a 2D particle with velocity control. We model the opponent's unknown type as $\theta \in \Theta := \{\text{ped}, \text{seg}\}$, with associated control sets $\mathcal{U}^o_{\text{ped}} = [-0.75, 0.75] \times [-2, 2]$ m/s and $\mathcal{U}^o_{\text{seg}} = [-0.75, 0.75] \times [-8, 0]$ m/s. The joint state space is defined as $z := (x^e, x^o, b(\text{ped})) \in \mathbb{R}^8$. The robot uses Bayesian inference learning dynamics f_L.

3 Closing the Safety Loop Around Runtime Learning: Belief-Space Games

Predictive control bound. Our key effort in reducing the conservativeness of safety analysis is to model the opponent's control bound as *dynamically coupled* to the robot's internal runtime learning. In particular, we define the *predictive bound* on the opponent's control action as:

$$\widehat{\mathcal{U}}^o(z_t) := \bigcup_{\theta \in \Theta} \widehat{\mathcal{U}}^o_\theta(z_t), \quad \widehat{\mathcal{U}}^o_\theta(z_t) := \begin{cases} \mathcal{U}^o_\theta(x_t), & \text{if } b_t(\theta) \geq \epsilon_\theta \\ \emptyset, & \text{if } b_t(\theta) < \epsilon_\theta \end{cases} \tag{5}$$

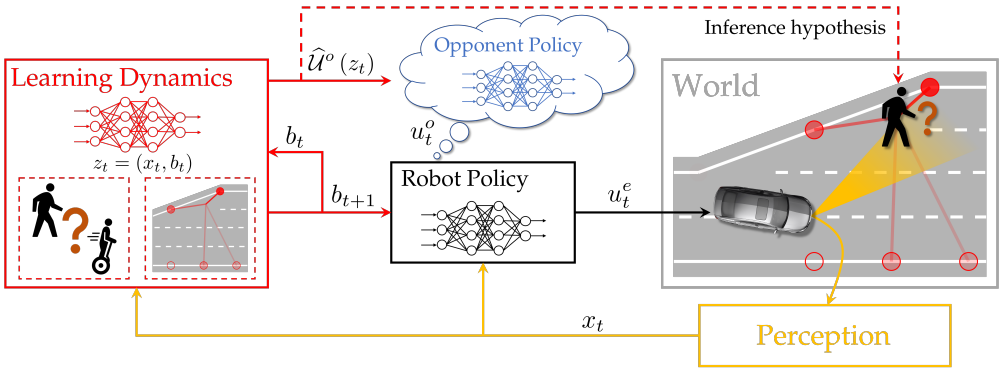


Figure 2: Deployment of the Belief Game policy trained with adversarial deep RL in closed-loop with a deep neural network learning dynamics model. The robot policy takes as inputs both the physical state x_t and belief state b_t, and operates under the *inference hypothesis* (dashed red line) that the opponent behaves according to the predictive control bound $\widehat{\mathcal{U}}^o(z_t)$ imposed by the learning dynamics.

where $\epsilon_\theta \geq 0$ is a small probability threshold. Model (5) assumes that, when $b_t(\theta)$ falls below ϵ_θ, indicating that θ is *sufficiently unlikely*, the corresponding control set $\mathcal{U}_\theta^o(x_t)$ can be discarded, resulting in a smaller predictive bound $\widehat{\mathcal{U}}^o(z_t)$ and thereby reducing the opponent action uncertainty.

The inference hypothesis. The belief-based restriction of the opponent's behavior is not *causal*, but *inferential*: it is based on the premise that an agent cannot take actions only explained by type θ without the robot's belief placing nonnegligible probability on this type. There is also an implied limit to abruptness: by the time of such an action, the robot must *already* have a belief of at least ϵ_θ in θ. Note that (5) does *not* preclude the scenario of an unlikely hypothesis becoming likely again, i.e., a previously small $b(\theta)$ may regain probability and surpass threshold ϵ_θ, at which point the corresponding control set $\mathcal{U}_\theta^o(x_t)$ is automatically added back to the overall predictive control bound $\widehat{\mathcal{U}}^o(z)$.

Belief-space reachability. The coupling between the robot's learning dynamics and the opponent's action uncertainty induces a new safety problem: a *belief-space reach-avoid game* in which the ego agent attempts to robustly ensure safety against the worst-case behavior of an opponent whose actions must be plausible in light of past behavior, yet may be strategically deceptive. To solve this game, we extend the HJI reachability formulation (Section 2) to the joint state space. While the target set $\mathcal{T} := \{z \mid \ell(x) \leq 0\} \subseteq \mathcal{X} \times \Delta$ and failure set $\mathcal{F} := \{z \mid g(x) > 0\} \subseteq \mathcal{X} \times \Delta$ are still defined in terms of only the *physical state*, the joint dynamics (4) lead to a reach-avoid solution that generally depends on the joint physical-belief state z, given by the belief-space Isaacs equation

$$V(z) = \max\left\{g(z), \min\left\{\ell(z), \min_{u^e \in \mathcal{U}^e} \max_{u^o \in \widehat{\mathcal{U}}^o(z), w \in \mathcal{W}, v \in \mathcal{V}} V\left(F(z, u^e, u^o, w, v)\right)\right\}\right\}. \tag{6}$$

Note that the ego and opponent controls given by solving (6) are optimal in the sense of a zero-sum partially observable stochastic game [16] (or dual control [17, 18]): they not only aim to influence the physical state x, but also the ego's uncertainty about the opponent, encoded by the belief state b, which then determines the opponent's predictive control bound $\widehat{\mathcal{U}}^o(z)$ at future time steps. In particular, the ego's control u^e strategically forces the opponent to reveal information about its type θ whenever reducing the uncertainty is critical to safely reaching the target. The opponent's control u^o is also dual: to sabotage the ego's reach-avoid task completion, it may behave deceptively to maintain strategic ambiguity that it can exploit later, for example, by forcing a collision (safety failure). This is formalized in the result below, which is a direct consequence of HJI reachability theory [19].

Proposition 1. *If the opponent satisfies the predictive control bound* (5)*, i.e* $u_t^o \in \widehat{\mathcal{U}}^o(z_t), \forall t \geq 0$*, then the sublevel set* $\{z \mid V(z) \leq 0\}$ *is a robust reach-avoid set from which the robot has a guaranteed control strategy—the minimizer of* (6)*—to drive the joint system* (4) *into* $\mathcal{T}$ *while avoiding* $\mathcal{F}$*.*

Running example: The ego vehicle's target set is defined as $\mathcal{T} = \{z \mid p_x^e \geq 85 \text{ m}\}$, where p_x^e denotes the ego vehicle's longitudinal position. Failure set $\mathcal{F}$ captures the ego colliding with the opponent or

driving out of the road boundary. The belief-space reach-avoid set $\mathcal{RA}(\mathcal{T}, \mathcal{F})$ computed numerically (with simplified motion models) on a grid is visualized in blue in Figure 1a.

Embracing data-driven predictors with model-free deep RL. Since the Isaacs equation (6) is generally intractable, and an explicit Markovian model of the belief dynamics may not always be available, we introduce a model-free deep adversarial RL approach to approximately solve (6) based on the recently developed Iterative Soft Adversarial Actor-Critic for Safety (ISAACS) framework [13].

Our adversarial reinforcement learning pipeline follows ISAACS to approximate $\mathcal{RA}(\mathcal{T}, \mathcal{F})$ offline by considering a time-discounted variant of the Isaacs equation (6). The learning scheme employs three neural networks: a critic $Q(z)$, an ego policy $\pi^e(z)$, and an opponent policy $\pi^o(z)$. These networks are trained through simulated gameplay, with more frequent updates to the critic and opponent policy than to the ego policy, thereby granting an informational advantage to the opponent [20]. During training, the same runtime learning algorithm that will be used at deployment is queried at each time step with the current observation (y or, depending on the algorithm, a *history* of recent observations) to produce an updated belief state b^+. The only requirement on the runtime learning algorithm is that they can be executed at training time; in particular, its learning dynamics need not have an explicit Markovian model f_L as in (3). The physics simulator enforces the *inference hypothesis* during training by projecting the opponent's action $\pi^o(z)$ onto the predictive control bound $\widehat{\mathcal{U}}^o(z)$ defined in (5). At runtime, the ego agent receives the current state estimate x and an observation y of the opponent, queries the learning dynamics for a belief state b, and executes control action $u^e = \pi^e(z)$. Figure 2 presents a schematic of our proposed safe control framework.

Running example: Consider the same road-crossing scenario, but now the robot is also uncertain about the human's goal position (visualized as red circles in Figure 4). In addition, we use a Transformer-based trajectory prediction model [2] (details can be found in Appendix B) as $f_L(\cdot)$. We use the above adversarial RL approach to integrate the predictor and scale up the safety analysis.

4 Case Studies

In this section, we illustrate the strength of our approach with three simulated driving examples that differ in problem scale, computation approach, and runtime learning mechanism. *Our main hypothesis is that the proposed Belief Game policy yields a planning performance close to an optimistic baseline while achieving a much lower failure rate.*

Warm-up: Dynamic programming solution. We first demonstrate our proposed approach on a simplified variant of the road-crossing scenario in the running example, in which the ego vehicle's motion is described by a 3D point mass model, the human is restricted to travel along the $p_y^o = 90$ m line, with possible types $\theta \in \{\text{ped}, \text{seg}\}$. With the above simplification, the joint system (4) is now 5-dimensional, which lends itself to numerical grid-based dynamic programming [21] for solving Isaacs equation (6) to high accuracy. We compare our proposed Belief Game policy to three baseline methods (all synthesized with the same grid resolution as the proposed method): The first is a maximum a posteriori (MAP) aligned policy [22], which *optimistically* uses a standard (belief-less) reach-avoid policy based on the MAP estimator $\hat{\theta} \in \arg\max_{\theta \in \Theta} b(\theta)$. The second baseline is a contingency policy [10, 23], which discards *currently* unlikely hypotheses for which $b(\theta) < \epsilon_\theta$ (also called "branching" in [10]) while safeguarding against all remaining hypotheses. The last baseline is a robust policy that *conservatively* safeguards against all hypotheses $\theta \in \Theta$ at all times.

We consider two metrics: the *failure rate*, defined as $N_{\text{fail}}/N_{\text{trial}} \times 100\%$, where N_{trial} is the number of trials and N_{fail} is the number of failed trials in which $z_t \in \mathcal{F}$ for some t, and the *completion time* averaged across all trials, which measures a policy's ability to make progress. A trial is considered complete if $p_x^e \geq 100$ m. For each policy, we simulated the scenario for $N_{\text{trial}} = 1000$ times with different initial states within the safe set and the (hidden) opponent type under the same random seed. Table 1 displays results obtained from all 1000 trials. Our proposed Belief Game policy maintained a zero failure rate, which empirically validated Proposition 1, and achieved an average completion time close to that of the optimistic but unsafe MAP baseline.

Metric/Method	MAP (optimistic)	Contingency	Robust (w/o learning)	Belief Game (ours)
Failure rate	11 %	3.1 %	0 %	0 %
Completion time (s)	4.49 ± 0.60	5.20 ± 1.12	6.27 ± 0.86	4.73 ± 0.97

Table 1: Failure rate and mean completion time in 1000 randomized trials with policies synthesized with numerical dynamic programming [21]. The proposed Belief Game policy ensures safety in all trials while closely matching the performance of an optimistic, but unsafe MAP policy.

Metric/Method	Opponent policy	MAP (optimistic)	Contingency	Robust (w/o learning)	Belief Game(ours)
Failure rate	Belief Adv. (Modeled)	40.7 %	31.2 %	22.3 %	5.7 %
	Belief Adv. (Unmodeled)	62.5 %	51.3 %	8.1 %	36.4 %
	Non-belief Adv.	5.2 %	1.5 %	7.4 %	1 %
	Noisily-rational	10.3 %	6 %	5.2 %	0.5 %

Table 2: Failure rate in 1000 randomized trials with policies trained with simulator-based deep adversarial RL. The Belief Game policy achieves a consistently lower failure rate than the MAP and contingency baselines.

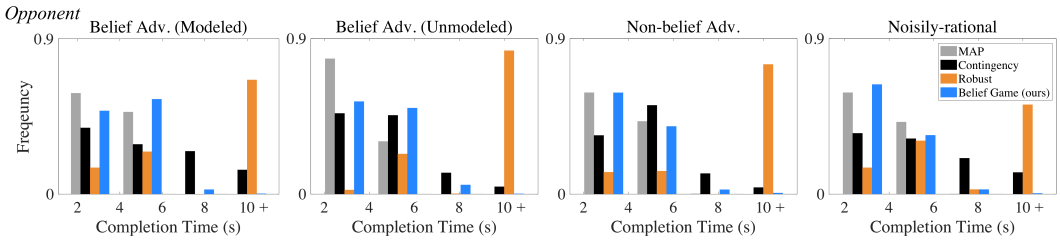


Figure 3: Completion time in 1000 randomized trials with policies trained with simulator-based deep adversarial RL. The performance of Belief Game closely matches that of the optimistic MAP baseline. The robust baseline is overly conservative and requires over 10 s to complete most trials.

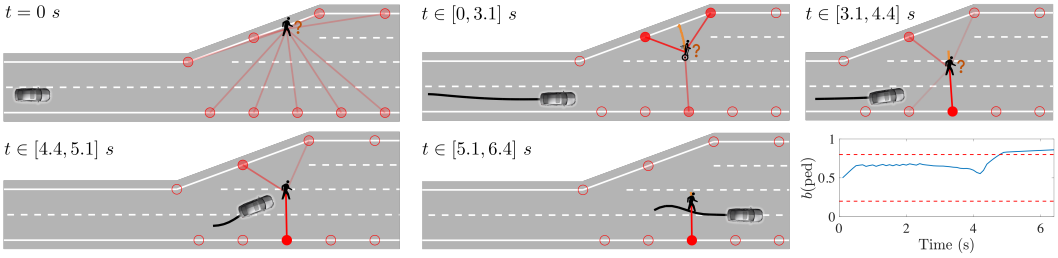


Figure 4: Simulation snapshots and belief (type hypothesis) trajectory of the running example. Both the robot and the human use the Belief Game policy synthesized with adversarial RL. Red circles denote the human's goal hypotheses and their color intensity represents the probability. A line drawn between the human and a goal hypothesis g_i means that $b(g_i) \geq \epsilon_\theta$. The pedestrian and Segway icons indicate the ground-truth human type. The orange question mark implies that neither hypothesis has a probability below the threshold. Despite the human's adversarial and deceptive motion, the ego vehicle managed to safely complete the task.

High-dimensional belief-space reachability with Bayesian learning dynamics. Now, we bring back the full setup of the running example with the 7D physical system, and *additionally* models the human's intent with hidden goal position $\theta_{\text{goal}} \in \Theta_{\text{goal}} := \{g_i\}_{i=1}^{N_{\text{goal}}}$, respectively, where $g_i \in \mathbb{R}^2$ denote goal hypotheses scattered along the road boundary and $N_{\text{goal}} = 9$ is the number of goal hypotheses. To cope with the high dimensionality of the joint state space, we apply the deep RL framework in Sec. 3 to approximately solve Isaacs equation (6). We again simulated for each policy the scenario 1000 times under the same random seed. The simulation was repeated for four different human policies: (i). *Belief-space adversarial (modeled):* The human uses adversarial policy $\pi^o(z)$ trained with the belief-space deep RL that approximates (6) and was projected to the predictive control bound $\widehat{\mathcal{U}}^o(z)$ at simulation time, (ii). *Belief-space adversarial (unmodeled):* Same as (i) but without the projection step, (iii). *Non-belief adversarial:* The opponent uses adversarial policy $\pi^o(x)$ trained with the non-belief deep RL that approximates (2) while the ego safeguards against all hypotheses, and (iv). *Noisily-rational:* The human uses a perturbed LQR policy to reach a goal position (randomized across trials). The failure rate and completion time results are shown in Ta-

ble 2 and Figure 3, respectively. The proposed Belief Game policy outperforms both the MAP and contingency baselines in all scenarios in terms of the failure rate. Remarkably, the completion time statistics of the Belief Game policy closely resemble that of the optimistic MAP policy in all four scenarios. The robust policy, albeit safe, was, however, overly conservative and unable to complete the task within 10 seconds in most trials. These results empirically support our main hypothesis.

In Figure 4, we closely examine one representative simulation trial of the running example initialized with a uniform prior belief. As the ego vehicle approached the human during $t \in [0, 4.4]$ s, the human *deceptively* alternated between applying pedestrian and Segway actions in order to hide its type. The ego vehicle then deliberately turned left, causing the adversarial human to move upwards and be exposed as a pedestrian. Due to the pedestrian's limited ability to move downwards compared to the Segway type, the ego vehicle could safely turn right and pass in front of the human.

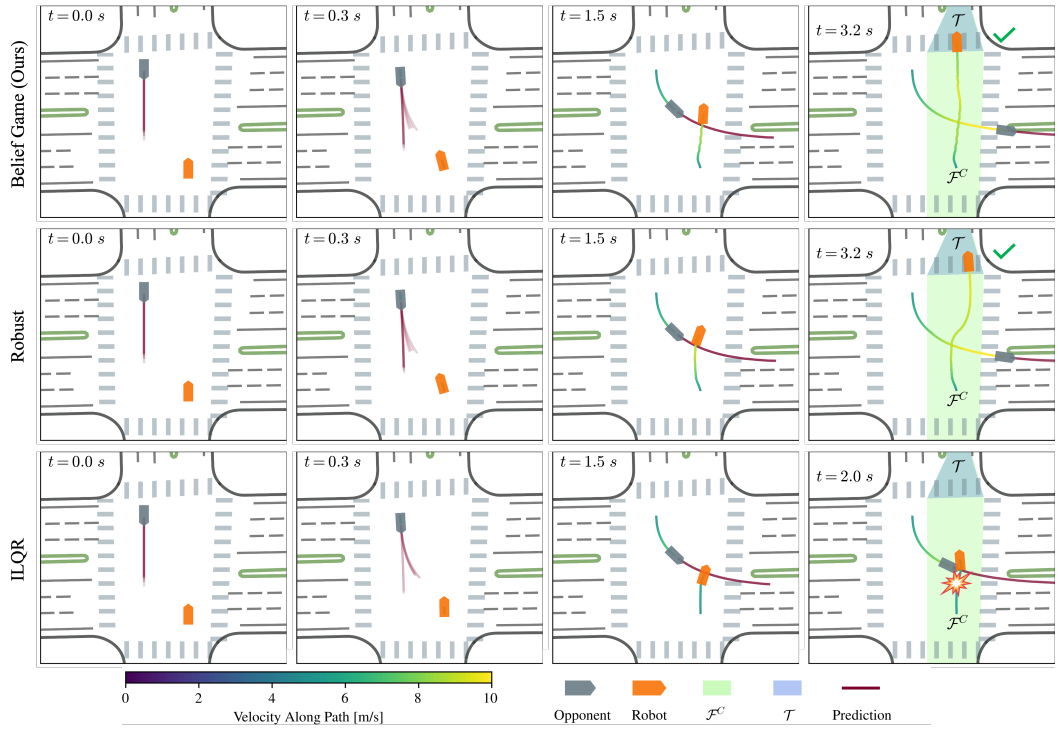


Figure 5: Belief Game with an MTR prediction model. The opponent suddenly turned left in the intersection. When using the Belief Game (top) and Robust policy (middle), the robot safely reached the target $\mathcal{T}$, while the ILQR policy (bottom) led to a collision. Compared to the robust policy, our proposed Belief Game policy generated a much smoother trajectory without radical evasive maneuvers.

Black-box trajectory prediction model. Now, we demonstrate that our method applies further to a data-driven non-Markovian learning process. Consider the scenario with two identical vehicles, when a robot aims to traverse the intersection without collision or violating the road constraints. The robot infers the opponent's future actions by predicting a Gaussian Mixture Model of trajectories from the scene's history. We utilize a proportional controller to track the mean trajectory of each mixture component (mode θ), from which we can obtain an opponent's nominal action prediction. Following (5), a small $d_t^o(\theta)$ around each predicted action is set for constructing the predictive control bound $\widehat{\mathcal{U}}^o$. We simulated a traffic intersection scenario extracted from Waymo Open Motion Dataset [24] in closed-loop with a *pre-trained* trajectory predictor, Motion Transformer (MTR) [2], which was used as the *implicit* learning dynamics $f_L(\cdot)$ for producing the belief states over time. Both the robot and opponent's control policies were trained using the deep RL approach in section 3, taking as inputs the physical states, predicted nominal actions, and the corresponding probabilities. Two baselines, set up with the same opponent policy, are introduced for comparison: a robust policy that safeguards against the entire opponent action space $\mathcal{U}^o$, and an iterative linear quadratic regulator (ILQR) policy that re-plans using MTR predictions during each planning cycle.

By closing the safety and learning loop, the Belief Game policy produced proactive ego behaviors, which prepared the robot for potential shifts in belief. As an example, in Figure 5, at $t = 0.3$ s, the robot predicted the opponent's left turn and maintained a lower velocity for a larger safety margin, which made it possible to safely pass the opponent later at $t = 1.5$ s when the opponent slowed down, attempting to block the robot. The robust policy produced more radical evasive actions to protect against the left-turning opponent regardless of belief. However, due to road boundaries and maneuverability limitations, the robust policy was unable to find feasible actions, as shown in Figures 7 and 8 in Appendix B. The ILQR policy failed all tests as it cannot account for the opponent's ability to change the robot's prediction in the future.

A close-up in Figure 6 confirms that the opponent using the Belief Game policy indeed took adversarial actions to *actively* shift the trajectory predictions in order to intercept the robot. This enables the opponent to pursue the robot within the predictive action bound and indicates that our framework is able to capture the implicit learning dynamics represented by a deep neural network. Moreover, even with one of the state-of-the-art trajectory predictors, small deviations from the nominal action can dramatically shift the robot's belief of the opponent over a short period. Therefore, classical methods (including the baseline ILQR policy) that rely on "accurate" trajectory prediction in an open-loop fashion [25] and frequently replan are insufficient for safe interactions. The intelligent robot needs to close the loop of learning by accounting for its future ability to learn and adapt.

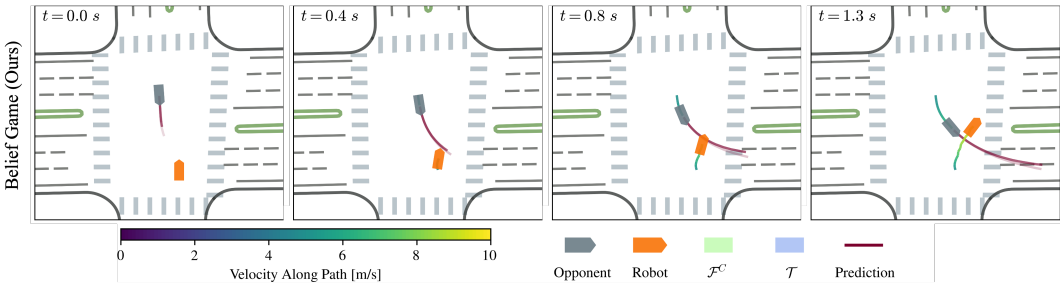


Figure 6: Our proposed framework enables the adversarial opponent to learn a policy to pursue the robot by quickly drifting the robot's beliefs from the black-box prediction model.

5 Limitations and Future Work

While empirical results of our proposed *Belief Game* framework in Section 4 have shown that it significantly improves the efficiency and maintains a low failure rate compared to baselines, we are not able to provide a safety guarantee for the robot when the solution of the game is approximated via reinforcement learning. In addition, our approach relies on a user-designed predictive control bound as part of our operational design domain for safety analysis. Selecting proper predictive control bound is a non-trivial task when deploying our proposed framework and will affect its performance in runtime. Finally, in our final example with MTR, we only evaluate our framework in one intersection with hand-crafted target and failure sets. However, modern trajectory predictor models, such as MTR, have already generalized to various scenarios given compatible map information. One potential future work is to leverage latent information learned from the trajectory prediction model and generalize the *Belief Game* safety analysis to general scenes.

6 Conclusion

In this paper, we present the *Belief Game*, a novel safety analysis framework that closes the loop between the robot's prediction-planning-control pipeline and its runtime learning process by *jointly* reasoning over agents' physical states and the robot's belief states. Our approach builds upon a game theoretic formulation and can scale up to high-dimensional problems and handle implicit learning dynamics via model-free adversarial reinforcement learning. Our experiment results indicate that the proposed method works with different types of interaction uncertainty, can be well combined with off-the-shelf prediction models, and plan efficient and safe trajectories.

References

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

[2] S. Shi, L. Jiang, D. Dai, and B. Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 2022.

[3] K. Nakamura and S. Bansal. Online update of safety assurances using confidence-based predictions. *arXiv preprint arXiv:2210.01199*, 2022.

[4] A. Bajcsy, S. Bansal, E. Ratner, C. J. Tomlin, and A. D. Dragan. A robust control framework for human motion prediction. *IEEE Robotics and Automation Letters*, 6(1):24–31, 2020.

[5] A. Muthali, H. Shen, S. Deglurkar, M. H. Lim, R. Roelofs, A. Faust, and C. Tomlin. Multi-agent reachability calibration with conformal prediction. *arXiv preprint arXiv:2304.00432*, 2023.

[6] I. M. Mitchell, A. M. Bayen, and C. J. Tomlin. A time-dependent hamilton-jacobi formulation of reachable sets for continuous dynamic games. *IEEE Transactions on automatic control*, 50 (7):947–957, 2005.

[7] R. Tian, L. Sun, A. Bajcsy, M. Tomizuka, and A. D. Dragan. Safety assurances for human-robot interaction via confidence-aware game-theoretic human models. *CoRR*, abs/2109.14700, 2021. URL https://arxiv.org/abs/2109.14700.

[8] Z. Zhang and J. F. Fisac. Safe Occlusion-Aware Autonomous Driving via Game-Theoretic Active Perception. In *Robotics: Science and Systems XVII*. Robotics: Science and Systems Foundation. ISBN 978-0-9923747-7-8. doi:10.15607/RSS.2021.XVII.066. URL http://www.roboticsproceedings.org/rss17/p066.pdf.

[9] C. Packer, N. Rhinehart, R. T. McAllister, M. A. Wright, X. Wang, J. He, S. Levine, and J. E. Gonzalez. Is anyone there? learning a planner contingent on perceptual uncertainty. In *Conference on Robot Learning*, pages 1607–1617. PMLR, 2023.

[10] A. Bajcsy, A. Siththaranjan, C. J. Tomlin, and A. D. Dragan. Analyzing human models that adapt online. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2754–2760. IEEE, 2021.

[11] S. Bansal and C. J. Tomlin. Deepreach: A deep learning approach to high-dimensional reachability. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1817–1824. IEEE, 2021.

[12] K.-C. Hsu, V. Rubies-Royo, C. J. Tomlin, and J. F. Fisac. Safety and liveness guarantees through reach-avoid reinforcement learning. In *Proceedings of Robotics: Science and Systems*, Held Virtually, July 2021. doi:10.15607/RSS.2021.XVII.077.

[13] K.-C. Hsu, D. P. Nguyen, and J. F. Fisac. Isaacs: Iterative soft adversarial actor-critic for safety. In *Proceedings of the 5th Conference on Learning for Dynamics and Control*, 2023.

[14] S. Bansal, M. Chen, S. Herbert, and C. J. Tomlin. Hamilton-jacobi reachability: A brief overview and recent advances. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 2242–2253. IEEE, 2017.

[15] R. Isaacs. Differential games i: Introduction. Technical report, RAND CORP SANTA MONICA CA SANTA MONICA, 1954.

[16] E. A. Hansen, D. S. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. In *AAAI*, volume 4, pages 709–715, 2004.

[17] A. A. Feldbaum. Dual control theory. i. *Avtomatika i Telemekhanika*, pages 1240–1249, 1960. URL http://mi.mathnet.ru/at12624.

[18] A. Mesbah. Stochastic model predictive control with active uncertainty learning: A survey on dual control. *Annual Reviews in Control*, 45:107–117, 2018. doi:10.1016/j.arcontrol.2017.11.001.

[19] J. F. Fisac, M. Chen, C. J. Tomlin, and S. S. Sastry. Reach-avoid problems with time-varying dynamics, targets and constraints. In *Proceedings of the 18th international conference on hybrid systems: computation and control*, pages 11–20, 2015.

[20] T. Zrnic, E. Mazumdar, S. Sastry, and M. Jordan. Who leads and who follows in strategic classification? In *Advances in Neural Information Processing Systems*, volume 34, pages 15257–15269, 2021. URL https://proceedings.neurips.cc/paper/2021/file/812214fb8e7066bfa6e32c626c2c688b-Paper.pdf.

[21] M. Bui, G. Giovanis, M. Chen, and A. Shriraman. Optimizeddp: An efficient, user-friendly library for optimal control and dynamic programming. *arXiv preprint arXiv:2204.05520*, 2022.

[22] L. Peters, D. Fridovich-Keil, C. J. Tomlin, and Z. N. Sunberg. Inference-based strategy alignment for general-sum differential games. In *19th International Conference on Autonomous Agents and Multi Agent Systems*, pages 1037–1045, 2020. URL https://dl.acm.org/doi/10.5555/3398761.3398882.

[23] J. Hardy and M. Campbell. Contingency planning over probabilistic obstacle predictions for autonomous road vehicles. *IEEE Transactions on Robotics*, 29(4):913–929, 2013.

[24] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, Z. Yang, A. Chouard, P. Sun, J. Ngiam, V. Vasudevan, A. McCauley, J. Shlens, and D. Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9710–9719, October 2021.

[25] D. González, J. Pérez, V. Milanés, and F. Nashashibi. A review of motion planning techniques for automated vehicles. *IEEE Transactions on intelligent transportation systems*, 17(4):1135–1145, 2015.

[26] L. Peters, A. Bajcsy, C.-Y. Chiu, D. Fridovich-Keil, F. Laine, L. Ferranti, and J. Alonso-Mora. Contingency games for multi-agent interaction. *arXiv preprint arXiv:2304.05483*, 2023.

[27] H. Hu, D. Isele, S. Bae, and J. F. Fisac. Active uncertainty reduction for safe and efficient interaction planning: A shielding-aware dual control approach. *arXiv preprint arXiv:2302.00171*, 2023.

[28] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.