Rethinking Safe Control in the Presence of Self-Seeking Humans

Anonymous submission

Abstract

Safe control methods are often intended to behave safely even in worst-case human uncertainties. However, humans may exploit such safety-first systems, which results in greater risk for everyone. Despite their significance, no prior work has investigated and accounted for such factors in safe control. In this paper, we leverage an interaction-based payoff structure from game theory to model humans' short-sighted, self-seeking behaviors and how humans change their strategies toward machines based on prior experience. We integrate such strategic human behaviors into a safe control architecture. As a result, our approach achieves better safety and performance trade-offs when compared to both deterministic worst-case safe control techniques and equilibrium-based stochastic methods. Our findings suggest an urgent need to fundamentally rethink the safe control framework used in human-technology interaction in pursuit of greater safety for all.

1 Introduction

This paper focuses on the safety-critical interactions of human agents and autonomous agents in the mixture of self-seeking and altruistic behaviors. Many safe control methods intend to behave safely in worst-case human uncertainties. The uncertainties can be great due to the complexity and difficulty of modeling human behaviors, which forces these methods to maintain a large safety margin and behave conservatively. However, when we assume humans are self-seeking actors, the standard safe methods might elicit the opposite effect. For example, when a human driver realizes that autonomous vehicles (AVs) always yield their right of way, the driver may cut in and change lanes aggressively toward AVs, which would pose greater risks for everyone. Safe control technology could be improved by incorporating how humans behave in response to autonomous systems.

In this work, we investigate the impact of strategic human behaviors in such interactions using an integration of control- and game-theoretic models. We use the principles of evolutionary game to characterize how rational humans change their strategy and behaviors over time and their impact on the performance and safety of the interaction. We classified diverse scenarios into four qualitatively different types and studied when safety is intended for worst-case human uncertainties, denoted as deterministic worst-case safe control (DWSC), and when equilibrium-based stochastic strategies, denoted as mixed strategy Nash equilibrium (MSNE), are chosen. Interestingly, the deterministic safe control discourages collaborative human behaviors, resulting in more risky interactions (Corollary 1, Figure 1).

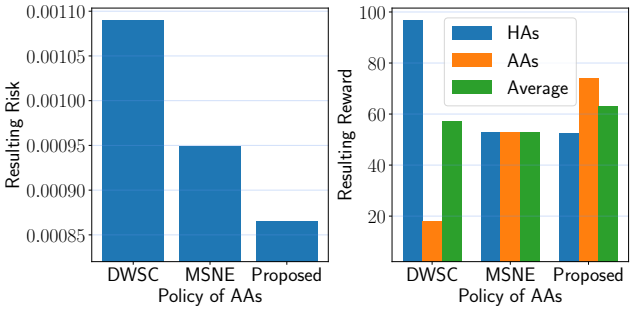

Figure 1: **Left:** Resulting risk. **Right:** Resulting reward. Both when Autonomous Agents use different policies in Type A interaction. Types of interactions are defined in Section 4.1.

Building on these insights, we then propose a method to encourage human to behave in an optimum way such that the safety of the overall system is maximized (Theorem 1, Figure 1, Figure 4). The proposed policy has better safety and performance trade-offs when compared to both deterministic worst-case safe control methods and equilibrium-based stochastic methods in the presence of strategic human behaviors (Figure 1, Figure 5). The advantage of the proposed method in risk management compared to MSNE and DWSC for one of the types is shown in Figure 2.

2 Related Work

Safe Control Many safe control methods exist for the design of autonomous systems that interact with humans. Some model human behaviors as uncertainties and noises and use stochastic safe control and multi-agent control (Ahmadi et al. 2019; Luo, Sun, and Kapoor 2020; Lyu, Luo, and Dolan 2021; Cheng et al. 2020). Others use various human models (Kulic and Croft 2007; Ding et al. 2011; Kelley et al. 2008; Ravichandar, Kumar, and Dani 2018; Koppula and Saxena 2015) to design control policies in a variety of tasks: *e.g.,* robotic swarm control (Atman et al. 2018; Diaz-Mercado, Lee, and Egerstedt 2017), manipulation tasks (Er-

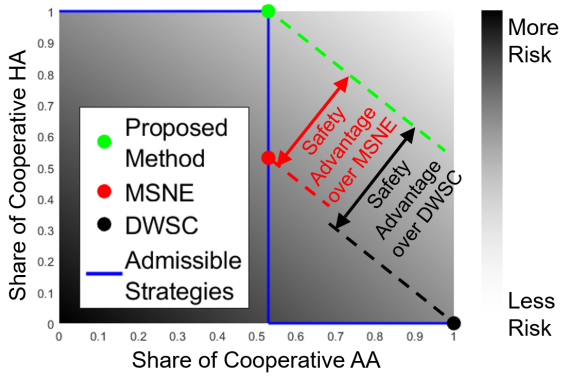

Figure 2: Risk map in Type A interaction. The blue line shows the strategies achievable by the proposed method. The dotted lines are level sets for the risk. The proposed method results in less risk compared to MSNE and DWSC. Types of interactions are defined in Section 4.1.

hart and Hirche 2016; Peng, Carabis, and Wen 2018) and autonomous vehicle control (Cummings et al. 2011). These methods are often designed to accommodate human behaviors and act with large safety margins with the intent to reduce tail (risk) events in the mixture of autonomous and human-driven cars. Since people use social information about others in coordinated movements (Faria, Krause, and Krause 2010), however, human drivers can develop risky behavior when they learn their counterpart is "playing the coward." In short, the cooperative policies of autonomous systems can be simply exploited by non-cooperative people (*e.g.,* the people who pursue self-interests) (Ishowo-Oloko et al. 2019; Dawes 1980; Shirado and Christakis 2020) and, as a result, may lead to greater risks for the entire system. To avoid such unintended consequences and improve the safety of both humans and machines, we might need to account for social interactions between them (Chen et al. 2017) in the safe control framework, which is the focus of this paper.

Human-Machine Cooperation Game theory is one of the major theoretical frameworks to examine complex social interactions. Using the framework, researchers have studied how cooperation can emerge from rational actors (Axelrod 1984). Cooperation is actually challenging because it creates a social dilemma (also known as the free-rider problem) (Dawes 1980). A group does well if individuals cooperate, but each individual is tempted to defect (Olson 1965). Even if one individual cooperates with others, the others could have an easy life by exploiting the first individual's benevolent effort (Nowak 2006). To overcome such cooperation dilemmas, a large body of work has explored broader, institutional approaches, such as punishment (Fehr and Gachter 2002), group dynamics (Shirado et al. 2013), and the establishment of a central authority (Ostrom 1990).

The cooperation problem also occurs in mixed groups of humans and machines. For example, Shirado and Christakis have introduced preprogrammed autonomous agents (bots) into a network of people to examine which bot strategies can facilitate cooperation in human groups (Shirado and Christakis 2020). In the study, the bots that always cooperated with humans were simply exploited by them, and most people eventually chose defection with the cooperative bots. Ishowo-Oloko et al. shows that people do not cooperate, especially when they realize that they are interacting with autonomous systems (Ishowo-Oloko et al. 2019). As theoretical and empirical evidence suggests the importance of accounting for self-seeking behaviors in cooperation, machines need to consider such human nature to facilitate cooperative human-machine systems (Paiva, Santos, and Santos 2018; Dafoe et al. 2021; Rahwan et al. 2019). This paper explores this implication in the safe control framework.

3 Problem Statement

We apply an interaction model of evolutionary game theory that two types of agents, Human Agents (HAs) and Autonomous Agents (AAs) interact with each other based on their payoffs for infinite periods (Hofbauer and Sigmund 1998; Nowak 2006). Specifically, the interactions between HAs and AAs are modeled as follows. We assume the existence of infinitely many HAs and AAs, and focus on the interaction between an HA and an AA. Their decision models are composed of interaction strategy $\pi = (\pi^h, \pi^a)$ and control policy $\phi = (\phi^h, \phi^a)$. Throughout this paper, we use superscript h to denote HAs and superscript a to denote AAs.

At each interaction of a HA and an AA, their intention I is decided based on their strategies π as follows:

$$\pi^h = \mathbb{P}(I^h = C) = 1 - \mathbb{P}(I^h = D) \quad (1)$$

$$\pi^a = \mathbb{P}(I^a = C) = 1 - \mathbb{P}(I^a = D), \quad (2)$$

The intentions can either be conservative (denoted as C for cooperative) or aggressive (denoted as D for defect), *i.e.,* $I^h, I^a \in \{C, D\}$.

Given the intention, they use the control policy

$$\mathbb{P}(u^h_{[k]}|I^h, x_{[k]}) \quad (3)$$

$$\mathbb{P}(u^a_{[k]}|I^a, x_{[k]}), \quad (4)$$

where $k \in \{0, 1, \cdots, K\}$, to generate the control action $u = (u^h, u^a)$ based on the state of both agents. Here, we use subscript $[k]$ to denote the discrete time point $k\Delta t$, where Δt is the sampling interval. Here, $u = \{u_{[k]} = [u^h_{[k]}, u^a_{[k]}]^T, k \in \{0, 1, \cdots, K\}\}$ and $x = \{x_{[k]} = [x^h_{[k]}, x^a_{[k]}]^T, k \in \{0, 1, \cdots, K\}\}$ are the control action and state of the HA and the AA at time k, and $\{0, 1, \cdots, K\}$ is the duration of interaction. We assume the control policy is identical among the population.

The system dynamics is characterized by the conditional transition probability $\mathbb{P}(x_{[k+1]}|x_{[k]}, u_{[k]}), k \in \{0, 1, \cdots, K\}$, which is identical among the population[1]. We will quantify the reward of each agent and the performance and safety of an interaction as follow. When the states and control actions end up being x and u, the reward of the HA and the AA is given by $\rho^h(x, u)$ and $\rho^a(x, u)$. Let

$$R^h = \mathbb{E}[\rho^h(x, u)] \quad (5)$$

$$R^a = \mathbb{E}[\rho^a(x, u)]. \quad (6)$$

[1]The system dynamics is assumed to be uniform for both the HA and the AA.

denote the expected reward received by the HA and the AA. The expected performance of the interaction is given by

$$R = R^h + R^a. \tag{7}$$

In addition to the reward, a latent risk is also present with the interaction. The risk W is quantified by the probability of the occurrence of some undesirable risk event, denoted as $\mathcal{U}$, *i.e.,*

$$W = \mathbb{E}[\mathbb{P}(\mathcal{U}|x, u)]. \tag{8}$$

HAs and AAs use different sets of information about the outcome of past interactions to change their strategies based on the outcome. We assume that the strategy update is performed at a sufficient slower timescale than individual interactions, so the strategy update can use accurate statistics of the outcomes associated with the past and current strategy[2]. As a result, we use different notations for the interaction time and the strategy update time. Each agent will have the information about the expected reward they receive under certain intentions. The AAs will have the information of the total reward R of the system as well. Unlike reward, only the AAs are able to calculate the latent risk W. This information asymmetry models the following two factors: the strategy of an AA can use the aggregate information of all other AAs; in contrast, HAs may not have a good estimate of the rare event probability such as crashes based on their experience, and HAs who get into accidents may exit from the population. The goal of the human agents is to achieve higher individual reward R^h. There exists a few common ways from existing literature that models the strategy update of HAs in evolutionary game theory (Hofbauer and Sigmund 1998; Nowak 2006). We adopt 3 of such models and consider an update rule consisting a mixture of these models. The human strategy update rule for each of the models are defined below.

- Replicator Dynamics (Taylor and Jonker 1978).

$$\begin{aligned}\dot{\pi}_t^h =& \pi_t^h(\mathbb{E}[R^h|\pi_t^h = 1, \pi_t^a] - \mathbb{E}[R^h|\pi_t^h, \pi_t^a]) \\ :=& f_r(\pi_t^h, \pi_t^a).\end{aligned} \tag{9}$$

- Brown-Nash-von Neumann Dynamics (Brown and Von Neumann 1950).

$$\begin{aligned}\dot{\pi}_t^h =& [\mathbb{E}[R^h|\pi_t^h = 1, \pi_t^a] - \mathbb{E}[R^h|\pi_t^h, \pi_t^a]]_+ \\ & - \pi_t^h([\mathbb{E}[R^h|\pi_t^h = 1, \pi_t^a] - \mathbb{E}[R^h|\pi_t^h, \pi_t^a]]_+ \\ & - [\mathbb{E}[R^h|\pi_t^h = 0, \pi_t^a] - \mathbb{E}[R^h|\pi_t^h, \pi_t^a]]_+) \\ :=& f_b(\pi_t^h, \pi_t^a).\end{aligned} \tag{10}$$

Here, $[q]_+ = \max(0, q)$.

[2]Here, we assume that humans collect sufficient information (interaction samples) before they change their behaviors. For example, if an HA meet with an AA showing conservative behaviors, it will not exploit this behavior. However, if the HA meets with the AA for sufficiently many times that it can confirm the AA will always act conservatively, it will starts to exploit the conservativeness with aggressive behaviors.

- Smith Dynamics (Smith 1984).

$$\begin{aligned}\dot{\pi}_t^h =& (1 - \pi_t^h)[\mathbb{E}[R^h|\pi_t^h = 1, \pi_t^a] - \mathbb{E}[R^h|\pi_t^h = 0, \pi_t^a]]_+ \\ & - \pi_t^h[\mathbb{E}[R^h|\pi_t^h = 0, \pi_t^a] - \mathbb{E}[R^h|\pi_t^h = 1, \pi_t^a]]_+ \\ :=& f_s(\pi_t^h, \pi_t^a).\end{aligned} \tag{11}$$

Here, $[q]_+ = \max(0, q)$.

The mixed dynamics update rule is given by

$$\begin{aligned}\dot{\pi}_t^h =& w_r f_r(\pi_t^h, \pi_t^a) + w_b f_b(\pi_t^h, \pi_t^a) + w_s f_s(\pi_t^h, \pi_t^a) \\ :=& f^a(\pi_t^h, \pi_t^a),\end{aligned} \tag{12}$$

where w_r, w_b and w_s are the weights for Replicator Dynamics, Brown-Nash-von Neumann Dynamics and Smith Dynamics, respectively, and

$$w_r + w_b + w_s = 1. \tag{13}$$

Our objective is to optimize the performance of the interaction while controlling the risk to be within a tolerable range. Toward this goal, we will design the strategy update rules of the AAs, which in turn influence the strategy of HAs, for optimizing the outcome of the AA-HA interactions, which depends on the strategies of both AAs and HAs. This objective is formally stated below.

$$\begin{aligned}\pi^\star = \arg\max_{\pi \in \mathcal{A}} \mathbb{E}[R|\pi] \\ \text{subject to } \mathbb{E}[W|\pi] \leq \epsilon.\end{aligned} \tag{14}$$

Here, ϵ is the tolerable risk and $\mathcal{A}$ is the set of admissible strategies, at which HAs have no incentive to deviate (Definition 1). We assume ϵ is chosen such that (14) is feasible.

The above model account for typical characteristics in human-machine interactions and has the following distinct factors from conventional game-theoretic models. First, there is a safety (or a latent risk) factor in addition to the rewards. Second, the available information is asymmetric: HAs can only estimate the expected rewards, while AAs can estimate both expected rewards and risks. Third, HAs is self-seeking while AAs is designed to optimize the safety and reward of the whole system. In the next section, we will understand the influence of short-sighted self-seeking humans, and study how to account for such propensity for improving the safety and efficacy in collective movements.

4 Case Studies

4.1 Categorizing Reward and Risk

We use a notation system such that R_{XY} denotes the expected reward when an agent chooses strategy X and its confronting agent chooses strategy Y. For simplicity, we consider a symmetric reward table. Likewise, for risk, we consider the same notation system. However, unlike reward, here we have one risk quantity for each interaction. Also, for simplicity, we make the following assumption:

$$W_{CD} = W_{DC}. \tag{15}$$

The simplified reward and risk table is given in Table 1.

	C	D
C	(R_{CC}, R_{CC});W_{CC}	(R_{CD}, R_{DC});W_{CD}
D	(R_{DC}, R_{CD});W_{CD}	(R_{DD}, R_{DD});W_{DD}

Table 1: Reward and risk table.

We typically observe scenarios whose reward and risks are ordered as follows.

$$\text{Reward case 1: } 2R_{CC} > R_{CD} + R_{DC} > 2R_{DD} \quad (16)$$

$$\text{Reward case 2: } R_{CD} + R_{DC} > 2R_{CC} > 2R_{DD} \quad (17)$$

$$\text{Risk case 1: } W_{CC} < W_{CD} < W_{DD} \quad (18)$$

$$\text{Risk case 2: } W_{CD} < W_{CC} < W_{DD}. \quad (19)$$

The interactions of the above case is qualitatively different. To better characterize such differences, we classify the interactions as below, along with example scenarios in the autonomous driving settings.

- Type A: reward case 1 and risk case 1.
- Type B: reward case 1 and risk case 2.
- Type C: reward case 2 and risk case 1.
- Type D: reward case 2 and risk case 2.

Intuitively, when the reward and risk belong to the same case, there exist strategies that simultaneously maximizes the reward and minimizes the risk. In this scenario, the optimizer of (14) is such strategy. When the reward and risk belong to different cases, the two objectives compete. In this scenario, the optimizer of (14) for varying risk tolerance ϵ characterizes a set of pareto-optimal strategies.

4.2 Autonomous Driving Simulation

Typical scenarios for each type in the autonomous driving setting are as follows:

- Type A: At a stop-sign controlled intersection, where the aggressive behavior of one vehicle leads to less total reward (causing a havoc in passing order that takes time to be resolved) and more risk (crashing into vehicles passing normally).
- Type B: Driving on a narrow road that constantly poses hazards to the vehicles on it (*e.g.,* falling rock). The aggressive behavior of one vehicle helps reduce the risk by deciding the passing order in a time shorter than the time needed for 2 cooperative vehicles to negotiate the order.
- Type C: In a lane changing scenario, the aggressive behavior of one vehicle creates more risk since it is more likely to crash by not yielding. On the other hand, it gives better total reward by eliminating its yield time.
- Type D: In a high speed lane changing scenario. Different from the previous scenario, yielding and reducing speed when many vehicles are driving at high speed may leads to more likelihood for crashing.

We use a narrow road driving example illustrated in Figure 3 to simulate the realistic reward and risk values that represents all 4 types. The example considers two vehicles, one is driven by a human and the other is an AV, on a narrow road with gravel outside of the paved road. When driving

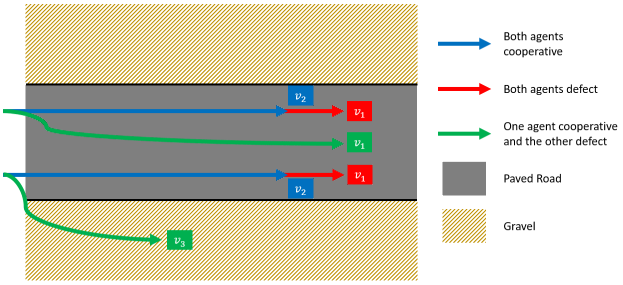

Figure 3: Simulated driving scenario.

on the gravel, the vehicles will have a crash probability. The AV has the option to act cooperatively by decelerating and adopting a DWSC approach. This approach keeps a safety distance whose length depends on the speed of the other vehicle. It also has the option to defect by ignoring the safety distance. If both vehicles choose to cooperate, they will both decelerate to a speed v_2 such that they can both drive on paved road without violating the safety distance. If the human driver choose to defect, continuing driving with speed v_1 without decelerating, and the AV adopts DWSC, the AV will be forced onto gravel due to safety distance constraints and only drive at a much more reduced speed v_3. If both vehicles choose to defect, they will drive the paved road with their original speed v_1. The detailed specifications for the simulation is provided in the Appendix. Table 2 shows the generated rewards and risks for all cases.

	Type A	Type B	Type C	Type D
R_{CC}	65.51	53.53	60.29	56.13
R_{CD}	17.93	-0.05	40.79	49.87
R_{DC}	96.8	68.7	95.28	88.24
R_{DD}	-69.23	-264.59	40.31	43.49
W_{CC}	0.00078	0.00147	0.00058	0.00057
W_{CD}	0.00109	0.00134	0.00073	0.00044
W_{DD}	0.00147	0.00172	0.0015	0.00077

Table 2: Reward and risk values for different types.

4.3 Deterministic Worst-Case Safe Control (DWSC)

In DWSC, AAs are designed to be always collaborative (*i.e.,* they adopt the time-invariant policy $\pi^a = 1$). The intention is to make AAs always ready for worst-case scenarios. However, such designs overlook the fact that in the presence of self-seeking HAs, the risk of the interaction might be higher than other not very conservative control policies.

To investigate the safety of DWSC, we need to see the convergence of HAs' policy under DWSC.

Corollary 1. *Assuming (23)–(25). If HAs are following the mixed dynamics* (12)*, an AA that will always choose* $\pi^a = 1$ *will result in* $\pi_t^h \xrightarrow{t\to\infty} 0$.

The proof is provided in the Appendix.

Based on corollary 1, we can calculate the expected risk as

$$\mathbb{E}[W|\pi_t] = W_{CC}\pi_t^h\pi_t^a + W_{CD}\pi_t^h(1-\pi_t^a) + W_{CD}(1-\pi_t^h)\pi_t^a + W_{DD}(1-\pi_t^h)(1-\pi_t^a). \tag{20}$$

As $t \to \infty$, we have $\pi_t^a = 1$ and $\pi_t^h = 0$. Hence,

$$\mathbb{E}[W|\pi_t] = W_{CD}. \tag{21}$$

Given the above, we can see that DWSC will always give $\mathbb{E}[W|\pi_t] = W_{CD}$. However, W_{CD} is not necessarily the minimum among $\{W_{CC}, W_{CD}, W_{DD}\}$. This also can be seen in the cases simulated in Section 4. This suggests that DWSC will not provide the safest behaviors in Type A and Type C interactions in the presence of strategic HAs' behaviors. In these types, DWSC will always start at a certain level of safety (given the underlying HAs cooperation distribution); however, when HAs start to exploit AAs cooperation, risk level goes higher. As shown in Figure 4, at the beginning, DWSC starts at a certain level of safety but risk increases as HAs cooperation decreases with time (HAs distribution changes toward a non-favorable manner w.r.t. safety). With this, we see how DWSC achieves short-sighted safety. Accordingly, people should be careful about that.

5 Proposed Algorithm

Before discussing our technical results, we list two definitions which are important preliminaries.

Definition 1 (Admissible Strategy)**.** *An admissible strategy is a strategy π which make HAs' strategy remain static, i.e., $\dot{\pi}^h = 0$.*

Here, the admissible strategy is not equivalent to the equilibrium. In fact, an equilibrium is an admissible strategy, but not every admissible strategy is an equilibrium. An equilibrium is a point where both π^h and π^a do not move. On the other hand, an admissible strategy only requires π^h to stop evolving, since π^a is something we have control on, it can be either static or dynamic, and it is decided by any control policy. As stated in (14), our design objective is to achieve an optimum policy within the set of admissible strategy. We consider this limitation because it can be difficult to design a control policy when HAs are changing strategies. In fact, most applied control policies are designed based on a training distribution (i.e., a certain HAs environment).

Definition 2 (Mixed Strategy Nash Equilibrium)**.** *In an interaction between AAs and HAs, although we have both rewards and risks, we define the mixed strategy Nash equilibrium (MSNE) only based on rewards as*

$$L = \frac{R_{DD} - R_{CD}}{R_{CC} + R_{DD} - R_{DC} - R_{CD}}. \tag{22}$$

Also, we require a few reasonable assumptions, which are satisfied in all the types presented in Section 4:

$$L \in (0,1) \tag{23}$$

$$R_{DD} < R_{CD} \tag{24}$$

$$R_{CC} < R_{DC} \tag{25}$$

Now, in lemma 1, we present the set of admissible strategies. Let

$$\mathcal{A}_0 = \{(\pi^h, \pi^a) : \pi^h = 0, \pi^a \in (L, 1]\} \tag{26}$$

$$\mathcal{A}_1 = \{(\pi^h, \pi^a) : \pi^h = 1, \pi^a \in [0, L)\} \tag{27}$$

$$\mathcal{A}_d = \{(\pi^h, \pi^a) : \pi^h \in [0, 1], \pi^a = L\}. \tag{28}$$

The set of admissible strategy $\mathcal{A}$ is given by

$$\mathcal{A} = \mathcal{A}_0 \cup \mathcal{A}_1 \cup \mathcal{A}_d. \tag{29}$$

Lemma 1. *Consider a system where AAs are adopting a policy π_t^a, HAs are following the mixed dynamics* (12)*. Assuming (23)–(25), then,*

$$\pi_t := (\pi_t^h, \pi_t^a) \in \mathcal{A} \Rightarrow \dot{\pi}_t^h = 0. \tag{30}$$

The proof is provided in the Appendix.

Next, we present our proposed algorithm. Let

$$\mathcal{B}_0 = \left[\frac{\epsilon - W_{DD}}{W_{CD} - W_{DD}}, \infty\right) \tag{31}$$

$$\mathcal{F}_0 = \{(\pi^h, \pi^a) : \pi^h = 0, \pi^a \in (L, 1] \cap \mathcal{B}_0\} \tag{32}$$

$$\mathcal{B}_1 = \begin{cases} \left(-\infty, \frac{\epsilon - W_{CD}}{W_{CC} - W_{CD}}\right], & W_{CC} > W_{CD} \\ \left[\frac{\epsilon - W_{CD}}{W_{CC} - W_{CD}}, \infty\right), & W_{CC} < W_{CD} \end{cases} \tag{33}$$

$$\mathcal{F}_1 = \{(\pi^h, \pi^a) : \pi^h = 1, \pi^a \in [0, L) \cap \mathcal{B}_1\} \tag{34}$$

$$\mathcal{B}_d = \begin{cases} \left(-\infty, \frac{\epsilon - W_{CD}L - W_{DD}(1-L)}{W_{CC}L + W_{CD} - 2W_{CD}L - W_{DD}(1-L)}\right], \\ \quad W_{CC}L + W_{CD} - 2W_{CD}L - W_{DD}(1-L) > 0 \\ \left[\frac{\epsilon - W_{CD}L - W_{DD}(1-L)}{W_{CC}L + W_{CD} - 2W_{CD}L - W_{DD}(1-L)}, \infty\right), \\ \quad W_{CC}L + W_{CD} - 2W_{CD}L - W_{DD}(1-L) < 0 \end{cases} \tag{35}$$

$$\mathcal{F}_d = \{(\pi^h, \pi^a) : \pi^h \in [0, 1] \cap \mathcal{B}_d, \pi^a = L\}. \tag{36}$$

We define the feasible set of strategies under the constraints of (14):

$$\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \mathcal{F}_d. \tag{37}$$

Then, the optimal strategy is given by

$$\pi^* := (\pi^{h*}, \pi^{a*}) = \underset{(\pi^h, \pi^a) \in \mathcal{F}}{\arg\max}\, \mathbb{E}[R|\pi^h, \pi^a]. \tag{38}$$

The proposed policy is given by

$$\pi_t^a = \begin{cases} L - G(\pi^{h*} - \pi_t^h), & \pi^* \in \mathcal{F}_d \\ \pi^{a*}, & \text{otherwise} \end{cases} := f^a(\pi_t^h). \tag{39}$$

Here, $G \in \mathbb{R}$ is a strictly positive constant.

Theorem 1. *Consider a system where AAs are adopting the policy defined in* (31) *to* (39)*, and HAs are following the mixed dynamics* (12)*. Assuming (23)–(25), then, $(\pi_t^h, \pi_t^a) \xrightarrow{t \to \infty} (\pi^{h*}, \pi^{a*})$, which is the solution to* (14)*.*

The proof is provided in the Appendix.

The proposed algorithm is given in Algorithm 1.

Algorithm 1: Proposed algorithm.

Input: Rewards and risks (Table 1), the tolerable risk ϵ, the constant G.

1: Compute $\mathcal{A}$ using (26) to (29).
2: Compute $\mathcal{F}$ using (31) to (37).
3: Compute π^* using (38).
4: **while** $t > 0$ **do**
5: Observe HA strategy π_t^h.
6: $\pi_t^a \leftarrow f^a(\pi_t^h)$.
7: **end while**

6 Numerical Simulations

The experiments are divided into two parts. First, we study the existing methods from DWSC to equilibrium-based methods and aims to show that DWSC is not the safest and can result in riskier situations. Then we demonstrates the benefit of the proposed method that it can control risks within a tolerable range in the long time scale and achieve a better trade-off between safety and reward than both DWSC and MSNE.

6.1 Settings

We performed numerical simulations in Python with both ODE-based update rules and Monte Carlo based update rules. We simulated the evolution of autonomous-human interactions with 4 types of interaction scenarios introduced in Section 4.1 with reward and risk table obtained in Section 4.2. For HA update rules, we adopt both ODE based and Monte Carlo based versions of 4 dynamics, including Replicator Dynamics, Brown-Nash-von Neumann dynamics, Smith Dynamics, and a mixture of them. For ODE based update rules, we update the states based on dynamics of the form (9) to (12) by the Runge–Kutta method (Runge 1895) over 1×10^5 time steps. In Monte Carlo simulations, we consider an infinite population of HAs and an infinite population of AAs interacting with each other. We randomly pick $N = 1 \times 10^3$ pairs and update their intentions over 1×10^5 time steps. The proportion of cooperators in each sampled population is interpreted as the strategy π for that population. Each individual updates its intention as follows,

- Replicator Dynamics (RD): Each HA i will randomly pick another HA j with probability $\frac{1}{N-1}$, and change I_i^h to I_j^h with a probability proportional to the excess part of j's reward over its own.
- Brown-Nash-von Neumann Dynamics (BNN): Each HA with intention C will switch to D with a probability proportional to the excess part of D's expected reward over HAs' average reward , and vice versa.
- Smith Dynamics (SD): Each HA with intention C will switch to D with a probability proportional to the excess part of D's expected reward over C's expected reward, and vice versa.
- Mixed Dynamics: It consists of three HA subpopulations that take RD, BNN, and SD respectively. Each individual will change their intention following the update rule of their subpopulation but considering the whole population when making updates rather than their own subpopulation.

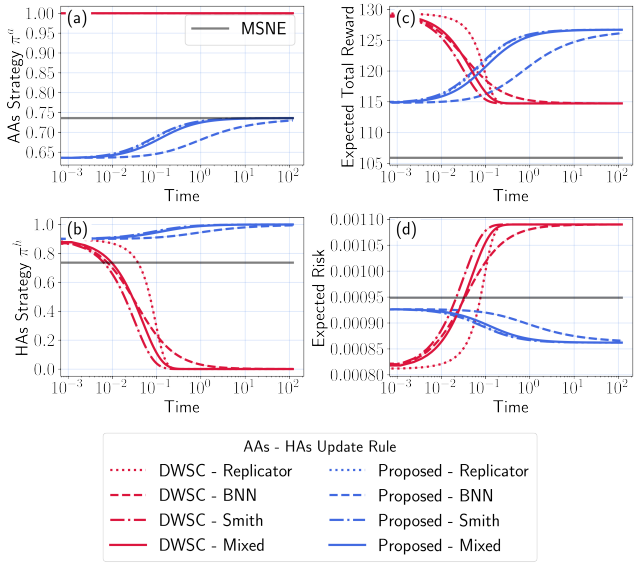

Figure 4: Results in Type A interaction with different AAs-HAs Update Rules ($\pi^h(0) = 0.9$, $\epsilon = 9e-4$, $G = 1$ and $w_r = w_b = w_s = \frac{1}{3}$). In plot (b), HAs' strategies converge to $\pi^h = 1$ if AAs takes DWSC. Meanwhile, the expected total reward in plot (c) decreases and the expected risk in plot (d) increases as $\pi^h \to 1$. Even though the initial rewards and risks of DWSC are quite desirable, it cannot remain static and it degrades quickly. In this case, the proposed method achieves the highest reward and lowest risk at the same time, while both the reward and risk that MSNE and DWSC achieve are sub-optimal.

For update rules of AAs, we compared the proposed methods with DWSC and equilibrium-based approaches. As discussed before, AA with DWSC will always be conservative for the worst-case, which means $\pi^a = 1$ for all time. For equilibrium-based approaches, since they aims to achieve MSNE, we simulate the state of MSNE for reference.

6.2 Results

We first compared the evolutionary trajectories of different AAs' update rules against different HAs' update rules. As Figure 4 shows, taking DWSC and always being conservative results in relatively safe interactions at the beginning. Then, HAs' strategies converge to $\pi^h = 1$ as time evolves, which corresponds to Corollary 1. However, as $\pi^h \to 1$, the expected total reward decreases and the expected risk increases, and it eventually ends up in riskier interactions. That demonstrates the incapability of DWSC to achieve safety in the long term.

To have a global view of what each method is able to achieve, we depict the admissible strategies in (29) and compared that with DWSC and MSNE in Figure 5. Both DWSC and MSNE show unsatisfactory behaviors in some types of

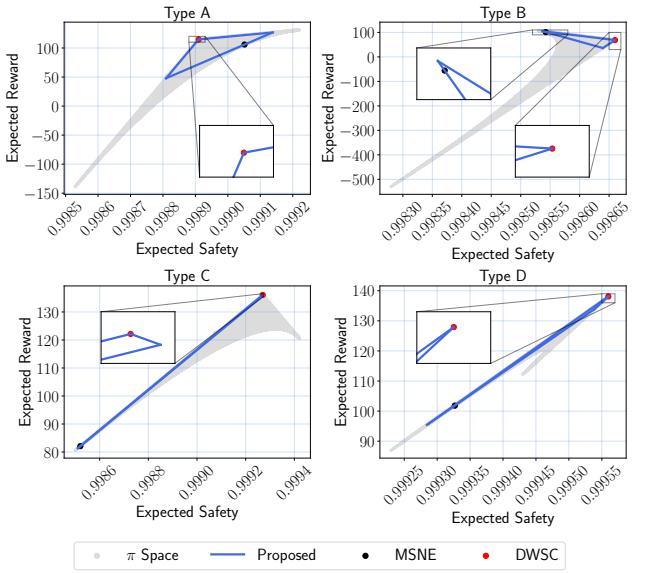

Figure 5: Illustration of achievable rewards and safety of different methods. The gray area labeled as 'π Space' denotes all possible interaction strategies. The red dots represent DWSC, and DWSC is always on the lines formed by the proposed method.

interactions and cannot guarantee either safety or reward. In Type A interactions, the plot matches the result in Figure 4. MSNE can achieve better safety than DWSC while DWSC generates better reward, but none of these two is optimal in either safety or reward. In Type B interactions, DWSC reaches the safest strategy and MSNE will bring a higher reward (less than the proposed method) at the cost of safety. While MSNE generates the lowest rewards and the highest risks among all AAs' update rules in Type C and D interactions, DWSC achieves the highest reward there. However, for DWSC in Type C, it will face a situation similar to Type A that DWSC will result in sub-optimal safety. All of these results signify the drawbacks of DWSC techniques and MSNE methods.

Then, we show that the proposed method that can control risks within a tolerable range in the long time scale and achieve a better trade-off between safety and reward than others. Figure 6 shows that the proposed method managed to control risk within ϵ in Type B interactions where we have no optimal value and need to make a trade-off. As shown in Figure 5, DWSC and MSNE only form two points that are not always optimal in all 4 types. However, in Type A and D interactions, the proposed method outperforms all others, and can even achieve the optimal strategy with the highest reward and safety. Moreover, the proposed method reaches a set of Pareto-optimal strategies, where it can make the trade-off between safety and rewards, in Type B and C interactions. Note that in Type A and C, the admissible strategy does not exist when ϵ less than a certain value, so we cannot be static there and the proposed method is actually the otimal. Although the exact shape of the plot depends on the

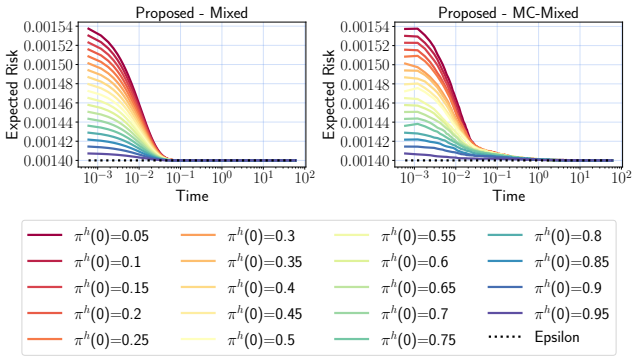

Figure 6: Robustness of the proposed methods controlling risks within a tolerable range against mixed HAs' update rules (**Left**: ODE, **Right**: Monte Carlo) given different initial conditions in Type B game ($\epsilon = 1.4e-3$, $w_r = w_b = w_s = \frac{1}{3}$ and $G = 1$).

interaction scenario's reward table and risk table, the relative relations between the proposed method and others are the same across each type. The comparison results of different methods is summarized below.

- DWSC: It can achieve the safest strategy in Type B, D and achieve the highest reward in Type C, D.
- MSNE: It cannot not guarantee to be the strategy with either the lowest risk or the highest reward in all types.
- Proposed: In Type A and D it can achieve the optimal strategy that is both the safest and has the highest reward. In Type B and C, it can make a trade-off between the best safety and the highest reward.

All above illustrates that the proposed method is optimal and can always achieve at least the same performance as, if not better than, DWSC methods and MSNE methods.

7 Conclusion

Summary We investigated conventional safe control methods in the presence of self-seeking humans. Our results showed that such control methods are actually not always the safest and the ones with highest performance. Namely, we proved that when humans exploit the cooperative behaviour of autonomous agents originated from conventional safe control methods, the overall safety of the human-machine interaction decreases. Also, we proposed a policy for autonomous agents that can encourage self-seeking humans to behave in a way that optimizes safety and performance of the interaction. Moreover, our method has a better design trade-offs than existing methods like deterministic worst-case safe control, and equilibrium-based stochastic strategies.

Future Work Currently, we assume the reward and risk table is fixed for different scenarios. However, it is natural that the environment changes over time and the interaction will change between different types. In the future, we will investigate how self-seeking human will affect safety when we have time-varying interaction environments.

References

Ahmadi, M.; Singletary, A.; Burdick, J. W.; and Ames, A. D. 2019. Safe policy synthesis in multi-agent POMDPs via discrete-time barrier functions. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 4797–4803. IEEE.

Atman, M. W. S.; Hatanaka, T.; Qu, Z.; Chopra, N.; Yamauchi, J.; and Fujita, M. 2018. Motion synchronization for semi-autonomous robotic swarm with a passivity-short human operator. *International Journal of Intelligent Robotics and Applications*, 2(2): 235–251.

Axelrod, R. 1984. *The evolution of cooperation*. Basic Books.

Brown, G. W.; and Von Neumann, J. 1950. Solutions of games by differential equations. *Contributions to the Theory of Games I*, 73–79.

Chen, Y. F.; Everett, M.; Liu, M.; and How, J. P. 2017. Socially aware motion planning with deep reinforcement learning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1343–1350. IEEE.

Cheng, R.; Khojasteh, M. J.; Ames, A. D.; and Burdick, J. W. 2020. Safe multi-agent interaction through robust control barrier functions with learned uncertainties. In *2020 59th IEEE Conference on Decision and Control (CDC)*, 777–783. IEEE.

Cummings, M. L.; How, J. P.; Whitten, A.; and Toupet, O. 2011. The impact of human–automation collaboration in decentralized multiple unmanned vehicle control. *Proceedings of the IEEE*, 100(3): 660–671.

Dafoe, A.; Bachrach, Y.; Hadfield, G.; Horvitz, E.; Larson, K.; and Graepel, T. 2021. Cooperative AI: machines must learn to find common ground. *Nature*, 593(7857): 33–36.

Dawes, R. M. 1980. Social Dilemmas. *Annual Review of Psychology*, 31(1): 169–193.

Diaz-Mercado, Y.; Lee, S. G.; and Egerstedt, M. 2017. Human–swarm interactions via coverage of time-varying densities. *Trends in Control and Decision-Making for Human–Robot Collaboration Systems*, 357–385.

Ding, H.; Reißig, G.; Wijaya, K.; Bortot, D.; Bengler, K.; and Stursberg, O. 2011. Human arm motion modeling and long-term prediction for safe and efficient human-robot-interaction. In *2011 IEEE International Conference on Robotics and Automation*, 5875–5880. IEEE.

Erhart, S.; and Hirche, S. 2016. Model and analysis of the interaction dynamics in cooperative manipulation tasks. *IEEE Transactions on Robotics*, 32(3): 672–683.

Faria, J. J.; Krause, S.; and Krause, J. 2010. Collective behavior in road crossing pedestrians: the role of social information. *Behavioral Ecology*, 21(6): 1236–1242.

Fehr, E.; and Gachter, S. 2002. Altruistic punishment in humans. *Nature*, 425: 137–140.

Hofbauer, J.; and Sigmund, K. 1998. *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Ishowo-Oloko, F.; Bonnefon, J.-F.; Soroye, Z.; Crandall, J.; Rahwan, I.; and Rahwan, T. 2019. Behavioural evidence for a transparency–efficiency tradeoff in human-machine cooperation. *Nature Machine Intelligence*, 1(11): 517–521.

Kelley, R.; Tavakkoli, A.; King, C.; Nicolescu, M.; Nicolescu, M.; and Bebis, G. 2008. Understanding human intentions via hidden markov models in autonomous mobile robots. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, 367–374.

Koppula, H. S.; and 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.

Kulic, D.; and Croft, E. A. 2007. Affective state estimation for human–robot interaction. *IEEE transactions on robotics*, 23(5): 991–1000.

Luo, W.; Sun, W.; and Kapoor, A. 2020. Multi-robot collision avoidance under uncertainty with probabilistic safety barrier certificates. *Advances in Neural Information Processing Systems*, 33: 372–383.

Lyu, Y.; Luo, W.; and Dolan, J. M. 2021. Probabilistic safety-assured adaptive merging control for autonomous vehicles. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 10764–10770. IEEE.

Nowak, M. A. 2006. *Evolutionary Dynamics*. Harvard University Press.

Olson, M. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press.

Ostrom, E. 1990. *Governing the Commons*. Cambridge University Press.

Paiva, A.; Santos, F. P.; and Santos, F. C. 2018. Engineering Pro-Sociality With Autonomous Agents. *The Thirty-Second AAAI Conference on Artifical Intelligence*, 7994–7999.

Peng, Y.-C.; Carabis, D. S.; and Wen, J. T. 2018. Collaborative manipulation with multiple dual-arm robots under human guidance. *International Journal of Intelligent Robotics and Applications*, 2(2): 252–266.

Rahwan, I.; Cebrian, M.; Obradovich, N.; Bongard, J.; Bonnefon, J.-F.; Breazeal, C.; Crandall, J. W.; Christakis, N. A.; Couzin, I. D.; Jackson, M. O.; Jennings, N. R.; Kamar, E.; Kloumann, I. M.; Larochelle, H.; Lazer, D.; McElreath, R.; Mislove, A.; Parkes, D. C.; Pentland, A. s.; Roberts, M. E.; Shariff, A.; Tenenbaum, J. B.; and Wellman, M. 2019. Machine behaviour. *Nature*, 568(7753): 477–486.

Ravichandar, H. C.; Kumar, A.; and Dani, A. 2018. Gaze and motion information fusion for human intention inference. *International Journal of Intelligent Robotics and Applications*, 2(2): 136–148.

Runge, C. 1895. Über die numerische Auflösung von Differentialgleichungen. *Mathematische Annalen*, 46(2): 167–178.

Shirado, H.; and Christakis, N. A. 2020. Network Engineering Using Autonomous Agents Increases Cooperation in Human Groups". *iScience*, 23(9).

Shirado, H.; Fu, F.; Fowler, J. H.; and Christakis, N. A. 2013. Quality versus quantity of social ties in experimental cooperative networks. *Nature Communications*, 4.

Smith, M. J. 1984. The stability of a dynamic model of traffic assignment—an application of a method of Lyapunov. *Transportation science*, 18(3): 245–252.

Taylor, P. D.; and Jonker, L. B. 1978. Evolutionary stable strategies and game dynamics. *Mathematical biosciences*, 40(1-2): 145–156.