Rethinking Safe Control in the Presence of Self-Seeking Humans

Zixuan Zhang*, Maitham AL-Sunni*, Haoming Jing*, Hirokazu Shirado, Yorie Nakahira

Carnegie Mellon University
{zzhang4, malsunni, haomingj}@andrew.cmu.edu, {shirado, yorie}@cmu.edu

Abstract

Safe control methods are often designed to behave safely even in worst-case human uncertainties. Such design can cause more aggressive human behaviors that exploit its conservatism and result in greater risk for everyone. However, this issue has not been systematically investigated previously. This paper uses an interaction-based payoff structure from evolutionary game theory to model humans' short-sighted, self-seeking behaviors. The model captures how prior human-machine interaction experience causes behavioral and strategic changes in humans in the long term. We then show that deterministic worst-case safe control techniques and equilibrium-based stochastic methods can have worse safety and performance trade-offs than a basic method that mediates human strategic changes. This finding suggests an urgent need to fundamentally rethink the safe control framework used in human-technology interaction in pursuit of greater safety for all.

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.

Motivated by this potential, this work investigates how safe control methods change human behavior and the safety of the overall system in the long term. The human-machine interaction is modeled using game and control theory as follows. The principles of evolutionary games are used to approximate the causal influence of the machines' policy on strategic human behaviors, *i.e.,* how rational humans change their strategy over time based on past experience. Control theory is used to characterize the impact of such changes on the performance and safety of the overall human-in-loop system. 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). Surprisingly, the deterministic safe control discourages collaborative human behaviors, resulting in more risky interactions (Lemma 1, Figure 1).

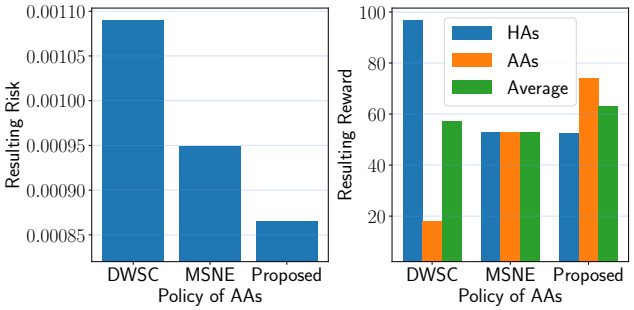

Figure 1: Resulting risk (left) and reward (right) when Autonomous Agents use different safe control methods in a Type A interaction. Types of interactions are defined in the Problem Statement.

Building on these insights, we then introduce a basic method that encourages humans to behave in a way that improves the safety of the overall system (Theorem 1, Figure 1, Figure 4). Interestingly, existing safe control methods can have worse safety and performance levels when compared to the proposed method that mediates human strategic behavior (Figure 1, Figure 5). A comparison between risk management in existing methods and the proposed method is shown in Figure 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

*These authors contributed equally.

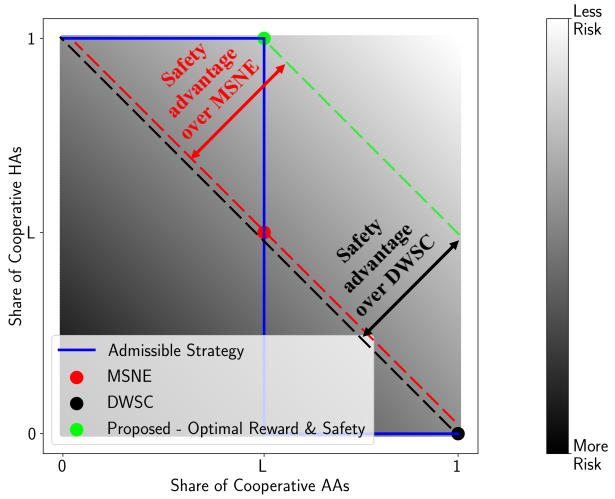

Figure 2: Risk map in a Type A interaction. The blue line shows the strategies achievable by the proposed method. The dashed lines are level sets for the risk. The proposed method results in less risk compared to MSNE and DWSC. Types of interactions are defined in the Problem Statement.

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; Jing and Nakahira 2022). 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 (Erhart 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. show 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.

Problem Statement

System Model

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 a 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 the 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, the performance, and the safety of interactions 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

$$R^h = \mathbb{E}[\rho^h(x, u)], \tag{5}$$
$$R^a = \mathbb{E}[\rho^a(x, u)], \tag{6}$$

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.,*

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

The risks we consider here are the types of risks that are not the major decision factors of humans, such as the long-term future risk, which are usually downplayed by humans against the immediate reward (Shirado, Crawford, and Christakis 2020). The risks that are major decision factors of humans are incorporated into the reward.

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, *i.e.,* subscript $[k]$ for interaction time and subscript t for strategy update time. Each agent will have 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.

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

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

We model human behaviors based on the widely accepted framework of myopic and bounded rationality in a distributed coordination, where HAs choose whether to cooperate based on the expectation of self-interests in the short term, *i.e.,* the individual reward R^h. In this setting, humans are well modeled as "conditional cooperators" theoretically and empirically (Hilbe, Nowak, and Sigmund 2013; Nowak and Sigmund 2005). In this setting, there exists a few common ways from existing literature that models the strategy update (dynamics) of HAs in evolutionary game theory (Hofbauer and Sigmund 1998; Nowak 2006). These human dynamics model the causal relationship between what humans experience and how they change their strategies or behaviors. To account for a wide range of possibilities, we adopt 3 of such models and consider an update rule consisting of a mixture of these models. The human strategy update rule for each of the models are defined below.

- Replicator Dynamics (Taylor and Jonker 1978).

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

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

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

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

- Smith Dynamics (Smith 1984).

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

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

The mixed dynamics update rule is given by

$$\begin{aligned}\dot{\pi}^h_t =& w_r f_r(\pi^h_t, \pi^a_t) + w_b f_b(\pi^h_t, \pi^a_t) + w_s f_s(\pi^h_t, \pi^a_t) \\ :=& f_m(\pi^h_t, \pi^a_t),\end{aligned} \tag{12}$$

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

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

Here, we make a reasonable assumption that $\pi^h_0 \neq 0$ and $\pi^h_0 \neq 1$.

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, we assume it is chosen such that (14) is feasible, and $\mathcal{A}$ is the admissible strategy set, which is defined below.

Definition 1 (Admissible Strategy Set)**.** *The admissible strategy set is the set of all strategies π that make HAs' strategy remain static, i.e., $\dot{\pi}^h = 0$.*

Here, the admissible strategy is not equivalent to 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 over, 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).

Reward and Risk Categorization

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, we require a few reasonable assumptions:

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

$$R_{DD} < R_{CD}, \tag{16}$$

$$R_{CC} < R_{DC}. \tag{17}$$

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 risk are ordered as follows.

$$\text{Reward case 1: } 2R_{CC} > R_{CD} + R_{DC} > 2R_{DD}. \tag{18}$$

$$\text{Reward case 2: } R_{CD} + R_{DC} > 2R_{CC} > 2R_{DD}. \tag{19}$$

$$\text{Risk case 1: } W_{CC} < W_{CD} < W_{DD}. \tag{20}$$

$$\text{Risk case 2: } W_{CD} < W_{CC} < W_{DD}. \tag{21}$$

The above cases create qualitatively different types of interactions. To characterize such differences, we classify the interactions as below, along with example scenarios in autonomous driving.

- 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 (both case 1 or both case 2), there exist strategies that simultaneously maximize the reward and minimize the risk. In this scenario, the optimizer of (14) is such a 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.

The above model accounts 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 are self-seeking while AAs are designed to optimize the safety and reward of the whole system. In the next sections, we will understand the influence of short-sighted self-seeking humans, and study how to account for such propensity to improve the safety and efficacy of collective movements.

Case Studies

Autonomous Driving Simulation

We use some typical autonomous driving settings to study the types of interactions. 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 havoc in traffic 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 a higher total reward by eliminating its yield time.
- Type D: In a high-speed lane-changing scenario. Different from Type C, yielding and reducing speed when many vehicles are driving at high speed may lead to more likelihood of crashing.

We use a narrow road driving example illustrated in Figure 3 to simulate realistic reward and risk values that represent 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 on 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 ν_2 such that they can both drive on a

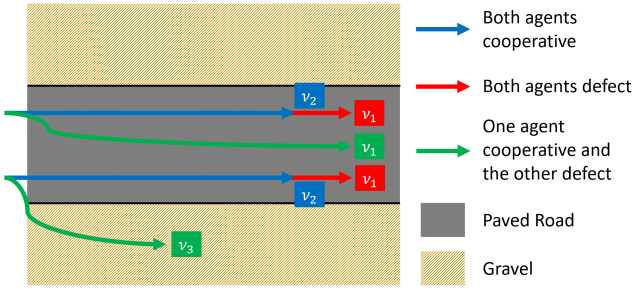

Figure 3: Simulated driving scenario.

paved road without violating the safety distance. If the human driver chooses to defect, continuing driving with speed ν_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 ν_3. If both vehicles choose to defect, they will drive the paved road with their original speed ν_1. The detailed specifications of the simulation are provided in the extended version of the paper (Zhang et al. 2022). 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.

Existing Methods

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.

Lemma 1. *Assuming* (16) *and* (17). *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 extended version of this paper (Zhang et al. 2022).

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

$$\begin{aligned}\mathbb{E}[W|\pi_t] = W_{CC}\pi_t^h\pi_t^a + W_{CD}\pi_t^h(1-\pi_t^a) \\ + W_{CD}(1-\pi_t^h)\pi_t^a + W_{DD}(1-\pi_t^h)(1-\pi_t^a).\end{aligned} \tag{22}$$

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

$$\mathbb{E}[W|\pi_t] = W_{CD}. \tag{23}$$

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 the Autonomous Driving Simulation. 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, the risk level goes higher. As shown in Figure 4, in the beginning, DWSC starts at a certain level of safety but risk increases as HAs cooperation decrease 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.

Alternatively, MSNE methods are designed to drive both AAs' and HAs' policies to the Mixed Strategy Nash Equilibrium (MSNE), which is defined below.

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*

$$L = \frac{R_{DD} - R_{CD}}{R_{CC} + R_{DD} - R_{DC} - R_{CD}}. \tag{24}$$

From (16) and (17), we have

$$L \in (0, 1). \tag{25}$$

In MSNE methods, risks and rewards are restricted to those that correspond to the equilibrium $\pi^a = \pi^h = L$. These methods cannot achieve Pareto-optimality between performance and safety because it is restricted to the equilibrium strategy. Risk resulting from AAs adopting MSNE methods can be calculated using (22) as

$$\begin{aligned}\mathbb{E}[W|\pi_t] = L^2(W_{CC} + W_{DD} - 2W_{CD}) \\ + 2L(W_{CD} - W_{DD}) + W_{DD}.\end{aligned} \tag{26}$$

Hence, no safety guarantee in MSNE methods, as its resulting risks mainly depend on different environmental variables. Figure 4 shows safety and reward levels for MSNE methods compared with other methods in a Type A interaction.

Proposed Algorithm

In this section, we present our results and the proposed algorithm with their theoretical guarantees.

Let

$$\mathcal{A}_0 = \{(\pi^h, \pi^a) : \pi^h = 0, \pi^a \in (L, 1]\}, \tag{27}$$

$$\mathcal{A}_1 = \{(\pi^h, \pi^a) : \pi^h = 1, \pi^a \in [0, L)\}, \tag{28}$$

$$\mathcal{A}_d = \{(\pi^h, \pi^a) : \pi^h \in [0, 1], \pi^a = L\}. \tag{29}$$

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

$$\mathcal{A} = \mathcal{A}_0 \cup \mathcal{A}_1 \cup \mathcal{A}_d. \tag{30}$$

In Lemma 2, we show that the set $\mathcal{A}$ is indeed the admissible strategy as defined in Definition 1.

Lemma 2. *Consider a system where AAs are adopting a policy π_t^a, HAs are following the mixed dynamics* (12)*. Assuming* (16) *and* (17)*, then,*

$$\pi_t := (\pi_t^h, \pi_t^a) \in \mathcal{A} \Leftrightarrow \dot{\pi}_t^h = 0. \tag{31}$$

The proof is provided in the extended version of this paper (Zhang et al. 2022).

Next, we present our proposed algorithm. Let

$$\mathcal{B}_0 = \left[\frac{\epsilon - W_{DD}}{W_{CD} - W_{DD}}, \infty\right), \tag{32}$$

$$\mathcal{F}_0 = \{(\pi^h, \pi^a) : \pi^h = 0, \pi^a \in (L, 1] \cap \mathcal{B}_0\}, \tag{33}$$

$$\mathcal{B}_1 = \begin{cases} \left(-\infty, \frac{\epsilon - W_{CD}}{W_{CC} - W_{CD}}\right], & W_{CC} > W_{CD} \\ \left[\frac{\epsilon - W_{CD}}{W_{CC} - W_{CD}}, \infty\right), & W_{CC} < W_{CD}, \end{cases} \tag{34}$$

$$\mathcal{F}_1 = \{(\pi^h, \pi^a) : \pi^h = 1, \pi^a \in [0, L) \cap \mathcal{B}_1\}, \tag{35}$$

$$\mathcal{B}_d = \begin{cases} \left(-\infty, \frac{\epsilon - W_{CD}L - W_{DD}(1-L)}{W_{CC}L + W_{CD} - 2W_{CD}L - W_{DD}(1-L)}\right], \\ \quad W_{CC}L + W_{CD} > 2W_{CD}L + W_{DD}(1-L) \\ \left[\frac{\epsilon - W_{CD}L - W_{DD}(1-L)}{W_{CC}L + W_{CD} - 2W_{CD}L - W_{DD}(1-L)}, \infty\right), \\ \quad W_{CC}L + W_{CD} < 2W_{CD}L + W_{DD}(1-L), \end{cases} \tag{36}$$

$$\mathcal{F}_d = \{(\pi^h, \pi^a) : \pi^h \in [0, 1] \cap \mathcal{B}_d, \pi^a = L\}. \tag{37}$$

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

$$\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \mathcal{F}_d. \tag{38}$$

Then, the optimal strategy is given by

$$\pi^\star := (\pi^{h\star}, \pi^{a\star}) = \underset{(\pi^h, \pi^a) \in \mathcal{F}}{\arg\max}\, \mathbb{E}[R | \pi^h, \pi^a]. \tag{39}$$

The proposed policy is given by

$$\pi_t^a = \begin{cases} L - G(\pi^{h\star} - \pi_t^h), & \pi^\star \in \mathcal{F}_d \\ \pi^{a\star}, & \text{otherwise} \end{cases} := f^a(\pi_t^h). \tag{40}$$

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

Theorem 1. *Consider a system where AAs are adopting the policy defined in* (32) *to* (40)*, and HAs are following the mixed dynamics* (12)*. Assuming* (16) *and* (17)*, then,* $(\pi_t^h, \pi_t^a) \xrightarrow{t \to \infty} (\pi^{h\star}, \pi^{a\star})$*, which is the solution to* (14)*.*

The proof is provided in the extended version of this paper (Zhang et al. 2022).

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 (27) to (30).
2: Compute $\mathcal{F}$ using (32) to (38).
3: Compute $\pi^\star$ using (39).
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**

Numerical Simulations

The experiments are divided into two parts. First, we study the existing methods (DWSC and MSNE) and show that DWSC is not the safest and can result in riskier situations. Then, we demonstrate the advantage of the proposed method and show 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.

Settings

We perform numerical simulations in Python with both ODE-based update rules and Monte Carlo-based update rules. We simulate the evolution of autonomous-human interactions with 4 types of interaction scenarios introduced in the Reward and Risk Categorization with reward and risk table obtained in the Autonomous Driving Simulation. 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 sub-populations (RD, BNN, and SD). Each individual will change their intention following the update rule of their sub-population but will consider the whole population when making updates rather than their own sub-population.

For AAs' update rules, we compare the proposed method with DWSC and MSNE approaches. As discussed before, AAs with DWSC will always be conservative and will always account for the worst-case, which means $\pi^a = 1$ for all time. For MSNE approaches, since AAs aim to drive the system to equilibrium, we simulate the state of MSNE for reference.

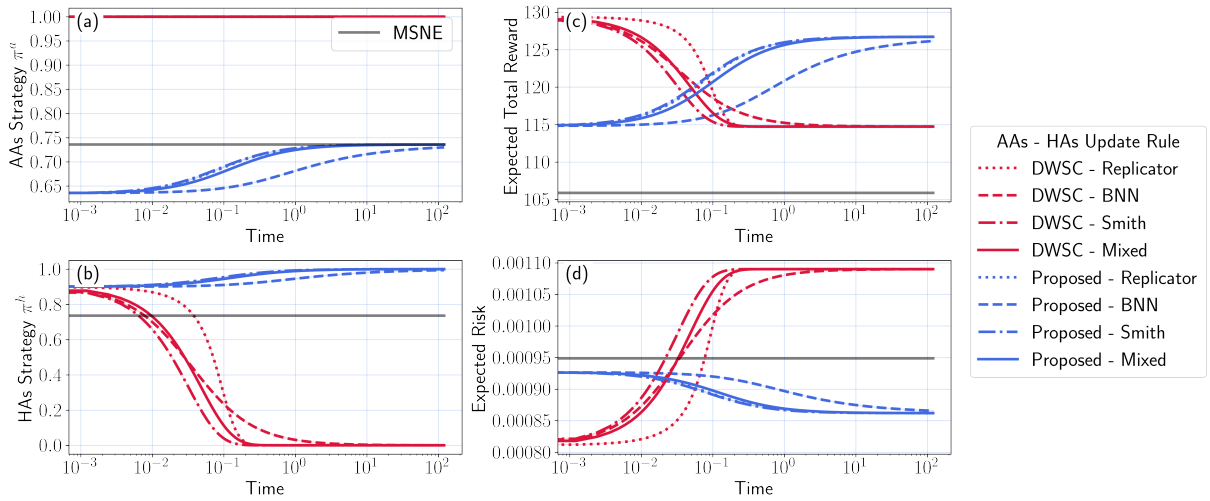

Figure 4: Results in a Type A interaction with different AAs-HAs update rules ($\pi_0^h = 0.9$, $\epsilon = 9e-4$, $G = 1$ and $w_r = w_b = w_s = \frac{1}{3}$). In plot (b), HAs' strategies converge to $\pi^h = 0$ if AAs take DWSC. Meanwhile, the expected total reward in plot (c) decreases and the expected risk in plot (d) increases as $\pi_t^h \to 0$. 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.

Results

We first compare the evolutionary trajectories of different AAs' update rules against different HAs' update rules. As shown in Figure 4, taking DWSC and always being conservative results in relatively safe interactions at the beginning. Then, HAs' strategies converge to $\pi^h = 0$ as time evolves, which corresponds to Lemma 1. However, as $\pi^h \to 0$, 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 strategy in (30) and compare that with DWSC and MSNE in Figure 5. Both DWSC and MSNE show unsatisfactory behaviors in some types of 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 faces a situation similar to Type A where DWSC will result in a sub-optimal safety level. All of these results signify the drawbacks of DWSC techniques and MSNE methods.

Then, we show that the proposed method can control risks within a tolerable range in the long time scale and achieve a better trade-off between safety and reward than DWSC and MSNE. Figure 6 shows that the proposed method can reach the entire Pareto frontier between optimal reward and optimal safety in Type B interactions. It also shows that DWSC and MSNE only form two points on the graph that are not always optimal in all 4 types. On the other hand, in Type A interaction, 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. The comparison of results of different methods is summarized below.

- DWSC: It cannot achieve the safest strategy in Type A and C, and cannot achieve the highest reward in Type A and B.
- MSNE: It cannot guarantee reward and safety levels in all types.
- Proposed method: 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 create a trade-off between the best safety and the highest reward.

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 the highest performance. Namely, we proved that when humans exploit the cooperative behavior of autonomous agents originating from conventional safe control methods, the overall safety of the human-machine interaction decreases. In addition, we proposed a basic policy for autonomous agents that can encourage self-seeking humans to behave in a way that optimizes the safety and performance of interactions. We have shown that existing

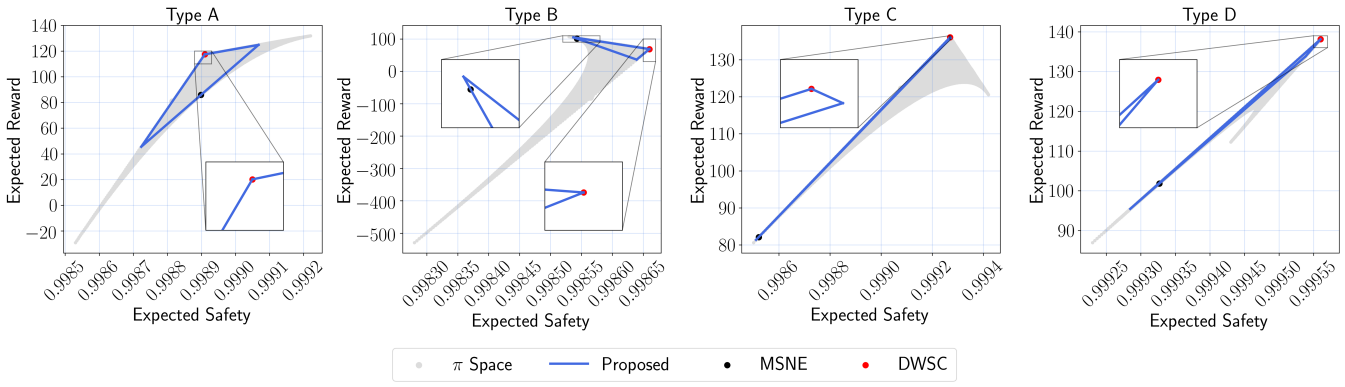

Figure 5: Illustration of achievable rewards and safety of different methods. The gray area labeled as 'π Space' denotes all possible interaction strategies. Note that the outcome of DWSC and MSNE methods is always achievable by the proposed method.

safe control techniques can result in worse safety and performance levels compared to the basic policy we proposed. As a result, we suggest that rethinking the existing safe control framework used in autonomous agents interacting with humans is an urgent need.

Limitations and Future Work The main limitations of our work are as follows. First, we assume that the intentions of agents do not change within each episode, which is not always the case. We plan to consider intention changes within single interaction episodes in the future to further analyze the safety of interaction and improve our method. Second, the intention of agents is limited to either cooperate or defect. To extend the applicability of our method to even more general situations, we will consider other kinds of intentions, such as some more neutral intentions between the aggressive intention and conservative intention. Third, it is natural that the environment changes over time and the rewards and risks

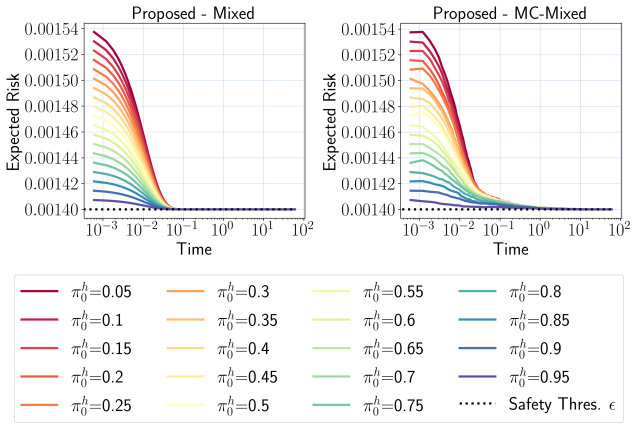

Figure 6: Robustness of the proposed method controlling risks within a tolerable range against mixed HAs' update rules (Left: ODE, Right: Monte Carlo) given different initial conditions in a Type B interaction ($\epsilon = 1.4e-3$, $w_r = w_b = w_s = \frac{1}{3}$ and $G = 1$).

may also change. Therefore, we will investigate how self-seeking humans affect safety when we have time-varying interaction environments. Furthermore, our method and framework mainly focus on the long-term causal effect of the strategic behaviors of human agents. In some interactions, the short-term effect at the beginning of the dynamic evolution also has a significant impact on safety. In the future, we will quantitatively evaluate the impact of human agent and autonomous agent strategies on safety for both the short-term and long-term dynamic evolution. Lastly, it is worthwhile to mention that in this work, we did not evaluate the proposed method with low-level human agents and autonomous agents dynamics; however, we presented the proposed method and showed the validity of it using evolutionary games principles to highlight the possible room for improvement that can be built on top of the existing methods.

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.

Hilbe, C.; Nowak, M. A.; and Sigmund, K. 2013. Evolution of extortion in iterated prisoner's dilemma games. *Proceedings of the National Academy of Sciences*, 110(17): 6913–6918.

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.

Jing, H.; and Nakahira, Y. 2022. Probabilistic Safety Certificate for Multi-agent Systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, 5343–5350. IEEE.

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.

Nowak, M. A.; and Sigmund, K. 2005. Evolution of indirect reciprocity. *Nature*, 437(7063): 1291–1298.

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.; Crawford, F. W.; and Christakis, N. A. 2020. Collective communication and behaviour in response to uncertain 'Danger'in network experiments. *Proceedings of the Royal Society A*, 476(2237): 20190685.

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.

Zhang, Z.; AL-Sunni, M.; Jing, H.; Shirado, H.; and Nakahira, Y. 2022. Rethinking Safe Control in the Presence of Self-Seeking Humans. arXiv:2212.00295.