

# The art of compensation: how hybrid teams solve collective risk dilemmas

Inês Terrucha  
IDLab, Ghent University – imec &  
Artificial Intelligence Lab, Vrije  
Universiteit Brussel  
Gent & Brussels, Belgium  
ines.terrucha@ugent.be

Elias Fernández Domingos  
ULB Machine Learning Group  
Brussels, Belgium  
elias.fernandez.domingos@ulb.be

Francisco C. Santos  
INESC-ID & Instituto Superior  
Técnico, Universidade de Lisboa  
Porto Salvo, Portugal  
franciscocsantos@tecnico.ulisboa.pt

Pieter Simoens  
IDLab, Ghent University – imec  
Gent, Belgium  
pieter.simoens@ugent.be

Tom Lenaerts  
ULB Machine Learning Group  
Brussels, Belgium  
tom.leanerts@ulb.be

## ABSTRACT

It is widely known how the human ability to cooperate has influenced the thriving of our species. However, as we move towards a hybrid human-machine future, it is still unclear how the introduction of AI agents in our social interactions will affect this cooperative capacity. Within the context of the one-shot collective risk dilemma, where enough members of a group must cooperate in order to avoid a collective disaster, we study the evolutionary dynamics of cooperation in a hybrid population made of both adaptive and fixed-behavior agents. Specifically, we show how the first learn to adapt their behavior to compensate for the behavior of the latter. The less the (artificially) fixed agents cooperate, the more the adaptive population is motivated to cooperate, and vice-versa, especially when the risk is higher. By pinpointing how adaptive agents avoid their share of costly cooperation if the fixed-behavior agents implement a cooperative policy, our work hints towards an unbalanced hybrid world. On one hand, this means that introducing cooperative AI agents within our society might unburden human efforts. Nevertheless, it is important to note that costless artificial cooperation might not be realistic, and more than deploying AI systems that carry the cooperative effort, we must focus on mechanisms that nudge shared cooperation among all members in the hybrid system.

## KEYWORDS

Coordination, Cooperation, Evolution and co-evolution of multi-agent systems

## 1 INTRODUCTION

As Artificial Intelligence (AI) systems are making more and more decisions autonomously, we are relinquishing decision control, for example by allowing intelligent machines to accomplish some of our goals independently or alongside us (e.g., using Google translate to enable business opportunities across different languages [14]), within the context of hybrid human-machine socio-technical systems (e.g., sharing the road with self-driving cars [24]). Given the extraordinary difficulties humans have demonstrated when

trying to overcome global crises, such as the COVID-19 pandemic [18] or climate change [23], the question can be raised on how AI agents may help to resolve the problems in coordinating the efforts in those and similar mixed-motive situations.

Even though many different works have advocated for the introduction of beneficial AI to promote human prosociality [13, 27, 30], others have pointed that humans may be keen to exploiting this benevolent AI behavior in their own favor [3, 4, 17, 21]. Thus, before flooding society with AI applications with the promise that they could solve some of our most pressing issues, it is worth asking: What behavioral response can be expected in the presence of AI agent partners? How may decision-making potentially be affected? Will hybrid groups involving AI with predefined decision-processes actually achieve greater collective success?

We frame here these questions within the context of the Collective Risk Dilemma (CRD) [23], a game that abstracts the conflict between helping the group to achieve a future goal at a personal cost, or free ride on the efforts of others and just collect the gains associated with achieving the goal. CRD is a public goods game with a delayed reward that is associated with societal problems like pandemic control through vaccination, climate negotiation to achieve CO<sub>2</sub> reduction and energy-grid usage by prosumers. Many experiments to assess human behavior have been performed [2, 6, 22, 23, 34], where [11] found that replacing all human participants by AI agents has a positive impact on the success rate of human groups. Yet, they also showed that in hybrid groups of humans and AI agents this success is again reduced to the level of only human participants. Within this context, this manuscript aims to unravel in more detail the dynamics in hybrid human-agent groups, providing thus knowledge that allows one to design novel experiments to further this line of research on human-AI interactions in mixed-motive and competitive scenarios.

Using the two-action one-shot CRD as defined in [31] (see also Methods), the current study aims to show which strategic responses can be expected (i.e., fraction of cooperative strategies) in groups consisting of AI agents with predefined stochastic behaviors and whether these learned responses are sufficient for success in reaching the goals. Hence two types of participants are considered, i.e. adaptive individuals that can change their behavior over time based on the outcome of their interactions (a proxy for potential human

responses) and artificial agents that have a probabilistic behavior that does not change as a result of the interactions (a proxy for average AI agent behavior). We apply a social learning approach to alter the strategy of the first type of individuals. They can switch between the possible actions in function of their success in the interactions; both when there are other adapting individuals or stochastic AI agents in groups of a given size. Such social adaptation can be achieved in different ways (e.g. Roth-Erev learning [19], Q-learning [10]), but here an evolutionary game theoretical approach is considered wherein strategic behaviors change in the adaptive individuals population by imitating those individuals that are performing the best [15, 16, 25, 35, 36].

In our model, automation does not assume that every AI agent will be continuously learning while acting; rule-based systems are used in AI products and they are hard-wired in the systems as learning on the fly might be costly or even dangerous. Thinking about real-world AI applications, one should always consider that producers of AI products want to give guarantees on what the product does (also on what are its limitations, which is why we use a stochastic behavior that includes errors), and that allowing for extensive adaptation while in use may be very risky. It is important to note that in this work, we are not considering the AI designer, neither the dynamics involved behind them. We are simply probing: If we consider this space of behaviors for AI agents, what kind of human behaviors emerge given constant hybrid interactions? Even though CRD scenarios are used to model very high risk events like a pandemic or the climate change, the same kind of non-linearity could be observed within any industrial or software hybrid team, which if the project is not delivered might suffer the consequences of losing their bonuses or even their jobs. This is especially interesting to probe since most teams are already hybrid if we consider the extensive productivity softwares that are already available in the market.

We will show that the adaptive individuals in this type of hybrid teams - where humans are informed about their artificial counterparts use and limitations - respond by exploiting the benevolence of the AI agents, by avoiding to contribute with cooperative efforts when the latter are already meeting the threshold needed, as previously hinted to in [17]. On the contrary, when the AI agents added to each group are associated with a lower capacity to contribute for the collective endeavor, and if the risk is high, the adaptive population cooperation levels are boosted to compensate. What is thus observed, is that in the presence of a set of pre-specified AI behaviors in a hybrid group of "humans" that can adapt their behavior in function of their gains, one obtains compensating behavior in both directions, requiring thus additional mechanisms to align the efforts of both types of participants in the hybrid group dynamic. So, adding cooperative or pro-social agents into a group decision-making process with humans may increase the success of achieving the task, but it will not necessarily promote pro-social behavior.

## 2 RELATED WORK

In [3] it is pointed out that more experimental research is needed to really understand how human strategic decision-making changes when interacting with autonomous agents. Following on this, [21]

compiles a review of more than 90 experimental studies that have made use of computerized players. Its main conclusions validate that indeed, human behavior changes when some of the other players are artificial, and furthermore, the behavior deviates to become more rational (or in other words, selfish), where humans are observed to actually try to exploit the artificial players.

This last conclusion was both supported by [4] and [17]. The first finds that humans cheat more against machines than against other humans, and thus prefer to play with machines, in an experiment that tested honesty in opposition to the possibility of higher financial gains. The latter recently published an experimental study that concludes that humans are keen on exploiting benevolent AI in various different classical social dilemma games. Within the context of the CRD used for the present work, [11] groups participants in hybrid teams with AI agents. Even though 3 out of the 6 group members were AI agents that were successful in avoiding the risky outcome in previous treatments, the hybrid groups were not more successful than only human groups. Looking closer at the results, one can see that the average payoff of the humans in hybrid teams actually increases. These experimental results already hint towards the adoption of a compensatory behavior on part of the human members of the group once they are informed about the addition of somewhat collaborative agents to the group.

In contrast with aforementioned works, [13] and [30] point towards the possibility of engineering prosociality in human behaviour through the use of pro-social AI agents. In the pursuit of this idea, [27] assembles a comprehensive review on the use of robots and virtual agents to trigger pro-social behaviour. Out of 23 studies included, 52% reported positive effects in triggering such cooperative behavior. However, 22% were inconclusive and 26% reported mixed results. Moreover, while recent experimental works show that programming autonomous agents [8] that include emotion [7] or some form of communication [5] may positively impact human cooperation, it is still unclear what are the mechanisms facilitating this effect.

More directly related to our theoretical study, there are different works on the dynamics of how evolving populations adapt their behavioral profile given the introduction of agents with a fixed behavior (usually cooperative) either at the group level or at the population level [20, 26, 32, 33, 38]. With our research questions, we also aim at understanding how the introduction of agents with a fixed behavior, not necessarily cooperative, affects the evolution of cooperation.

## 3 METHODS

### 3.1 The one-shot Collective Risk Dilemma (CRD)

In this manuscript we adopt the  $N$  person one-shot CRD [1, 9, 10, 12, 29, 31, 31, 37]. Here, a group of  $N$  individuals must each decide whether to Cooperate ( $C$ ), by contributing a fraction  $c$  of their initial endowment  $b$ , or to Defect ( $D$ ) and contribute nothing. If the group contains at least  $M$   $C$  players, i.e., the group contributes in total  $Mcb$  ( $M \leq N$ ) to the public good, then each player may keep whatever is left of their initial endowment. Otherwise, there is a probability  $r$  that all players will lose all their savings and receive a payoff of 0,

hence the dilemma. Thus, the expected payoff of a  $D$  and a  $C$  player can be defined in functions of the number of  $C$ s in the group,  $j$ :

$$\pi_D(j) = b(1 - r + r\theta(j - M)) \quad (1)$$

$$\pi_C(j) = \pi_D - cb, \quad (2)$$

where  $\theta(x)$  is the Heaviside unit step function, with  $\theta(x) = 0$  if  $x < 0$  and  $\theta(x) = 1$  otherwise.

### 3.2 CRD with hybrid interactions

We consider a population  $H$  of  $Z$  adaptive agents which are randomly sampled into groups of size  $N - a$  to play the CRD with  $a$  agents from population  $A$  (whose individuals display a fixed averaged behavior). This allows us, as explained in the section below, to investigate the population dynamics of this dilemma. When engaging in group interactions, each adaptive agent can either cooperate  $C$  or defect  $D$ . The state of the population is then defined by the number of cooperators  $k \in [0, Z]$ . The behavior of the fixed agents is defined by their probability of cooperating in each interaction,  $p \in [0, 1]$ , thus, they implement a stochastic (or mixed) strategy. In each group we can calculate the expected payoff of  $D$ s or  $C$ s in function of the number of cooperators from the adaptive population,  $i$ , the number of fixed agents  $a$  and the payoff of a  $D$  ( $C$ )  $\pi_{D(C)}$ :

$$\Pi_{D(C)}(i, a, p) = p\pi_{D(C)}(j = i + a) + (1 - p)\pi_{D(C)}(j = i) \quad (3)$$

The behavioral dynamics exhibited by the population of adaptive agents are governed by a social learning mechanism, where two randomly chosen individuals compare their fitness and imitate the one who is more successful within their social environment [15, 16, 25, 35, 36]. Their fitness is the measure of the success of their current strategy (their payoff) averaged over all different group interactions. It can be defined as a function of the aforementioned variables by taking into account the population state and the payoffs given by Eqs. 1 and 2. Following on this, the fitness equations for cooperative ( $C$ ) and defective ( $D$ ) strategies, can be written as:

$$f_C = \binom{Z-1}{N-a-1}^{-1} \sum_{i=0}^{N-a-1} \binom{k-1}{i} \binom{Z-k}{N-a-1-i} \Pi_C(i+1, a, p) \quad (4)$$

$$f_D = \binom{Z-1}{N-a-1}^{-1} \sum_{i=0}^{N-a-1} \binom{k}{i} \binom{Z-k-1}{N-a-1-i} \Pi_D(i, a, p). \quad (5)$$

Each agent in the population of adaptive agents may change its strategy profile at a given evolutionary step in the following way: an agent with a  $D$  ( $C$ ) strategy is randomly selected from the population  $H$  to adapt. With probability  $\mu$  it will mutate into a  $C$  ( $D$ ) strategy, otherwise, with probability  $1 - \mu$ , it will compare its fitness with another randomly selected agent (assuming the newly selected agent has a different strategy) [15, 16, 25, 28, 31, 35, 36]. In case imitation is selected, a  $D$  ( $C$ ) strategy will turn into a  $C$  ( $D$ ) with a probability

$$P(D \rightarrow C) = \frac{1}{1 + e^{-\beta(f_C - f_D)}}, \quad (6)$$

described by the Fermi function. This changes the state of the population  $H$  of adaptive agents from  $k$  to  $k + 1$ . This probability becomes higher with a larger difference between the fitness of the two agents,  $f_C - f_D$ , or with a larger selection strength of the process,  $\beta$ .

The transition probabilities that regulate the stochastic dynamics of population  $H$ , by defining the probability of increasing (+) or decreasing (-) the number of cooperators within a population are given by:

$$T^+(k) = \frac{Z-k}{Z} \left( (1-\mu) \frac{k}{Z-1} P(D \rightarrow C) \right) \quad (7)$$

$$T^-(k) = \frac{k}{Z} \left( (1-\mu) \frac{Z-k}{Z-1} P(C \rightarrow D) \right), \quad (8)$$

where  $P(C \rightarrow D)$  is obtained by replacing  $C$  with  $D$ , and  $D$  with  $C$  in Eq. 6.

From these equations, we can construct the complete Markov chain of the  $Z + 1$  different states that fully describe the evolutionary process of the population  $H$ . From this Markov Chain we can compute the stationary distribution  $P(k)$ , the average cooperation level  $\bar{C}$  and the average group success  $\bar{s}_G$  of each population configuration.

To compute the stationary distribution  $P(k)$ , we retrieve the eigenvector corresponding to the eigenvalue 1 of the tridiagonal transition matrix  $S = [p_{ij}]^T$  [15, 25, 35]. The values  $p_{ij}$  are defined by the equations:

$$p_{k,k\pm 1} = T^\pm(k) \quad (9)$$

$$p_{k,k} = 1 - p_{k,k-1} - p_{k,k+1} \quad (10)$$

where the formulas that define  $T^\pm(k)$  can be consulted in Eqs. 7 and 8.

From this it follows that the cooperation level  $\bar{C}$  of population  $H$  (for a given set of parameters  $N$ ,  $M$ ,  $r$ ,  $a$  and  $p$ ) by averaging the fraction of cooperators in each population state,  $k/Z$ , over the stationary distribution of states  $P(k)$  as given by:

$$\bar{C} = \sum_{k=0}^Z P(k) \frac{k}{Z}. \quad (11)$$

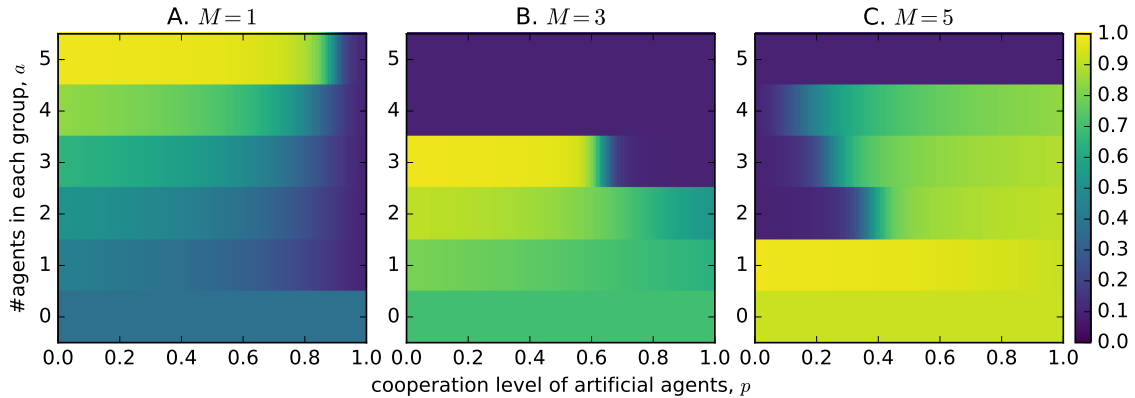
As already mentioned, within the context of the CRD [1, 9, 10, 12, 23, 29, 31, 37] another relevant quantity to derive is the probability of success of each group in reaching the threshold necessary of  $M$  cooperators to avoid the collective risk.

At the population level, we compute the fraction of groups in each population state that are successful by resorting to the multivariate hypergeometric sampling, as follows

$$s_G(k) = \binom{Z}{n-a}^{-1} \sum_{h=0}^{N-a} \binom{k}{h} \binom{Z-k}{N-h-a} \times \\ \times (p\theta(h+a-M) + (1-p)\theta(h-M)) \quad (12)$$

where  $\theta(x)$  is the Heaviside unit step-function as in Eqs. 1 and 2 of the main text.

Finally, similarly to what was done in Eq. 11, we calculate the averaged group success by weighing the group success of each population state of Eq. 12 over the stationary distribution of the



**Figure 1: Average cooperation  $\bar{C}$  (Eq. 11) when interacting in groups of size  $N = 6$  with  $a$  AI agents that cooperate with probability  $p$ . In order to avoid the risk of losing all the initial endowment with probability  $r = 0.9$  (value set for this figure),  $M = 1$  cooperators per group are needed in A.,  $M = 3$  in B. and  $M = 5$  in C.. When the threshold  $M$  is low, the addition of enough highly cooperative agents discourages the need for increased more cooperation of the adaptive agents (see A. and B.). However, when the threshold  $M$  is high, the addition of many highly cooperative AI agents turns the collective effort attainable and boosts cooperation (see C.). Other parameters used for this figure were:  $Z = 100$ ,  $\mu = 0.01$ ,  $\beta = 2$ ,  $b = 1$ ,  $c = 0.1$ .**

evolutionary process  $P(k)$ :

$$\bar{s}_G = \sum_{k=0}^Z P(k) s_G(k). \quad (13)$$

## 4 RESULTS AND DISCUSSION

### 4.1 Non-cooperative AI agents require the adapting population to cooperate to be successful

What level of cooperation can be expected from the population of adapting individuals? Specifically, we want to know how average cooperation  $\bar{C}$  (see Eq. 11) varies in relation to four parameters: the number of AI agents ( $a$ ) in each group, the probability of them cooperating ( $p$ ), the probability of risking complete loss of initial endowment ( $r$ ), and the number of cooperators needed per group to avoid that risk ( $M$ ).

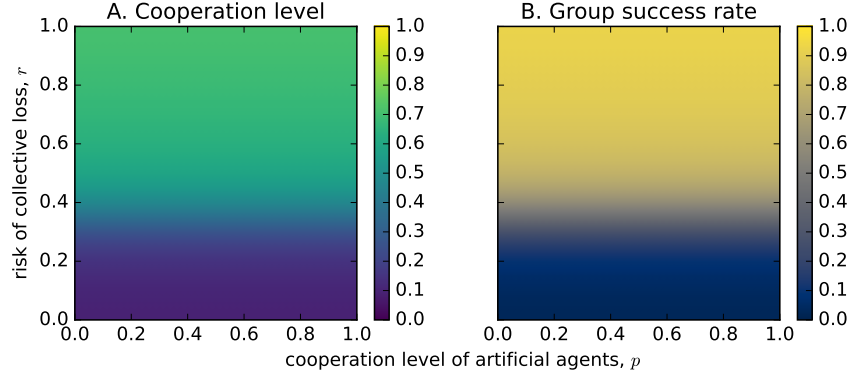
In Fig. 1, the variation in average cooperation  $\bar{C}$  is shown in terms of the number of AI agents per group  $a$  and their probability of cooperation  $p$ . Here, the group size is set to  $N = 6$ , and the collective risk of loss to  $r = 0.9$  if the group is unable to reach a threshold of at least  $M$  cooperators.  $M = 1$  (or 3 or 5) thus means that at least 1 (or 3 or 5) participant needs to cooperate to avoid disaster. It is observed in Fig. 1 that when the AI agents are at a fully cooperative behavior, i.e.  $p = 1$ , there is no motivation for human cooperation once the number of agents is equal to the threshold value,  $a = M$ , as they already contribute with all the effort necessary to avoid collective risk of loss. On the contrary, when AI agents that contribute with probability  $p = 0$  to the avoidance of that risk are introduced, then the  $N - a$  adaptive agents in the group are pressured to cooperate in order to reach  $M$  by themselves. Still, if the threshold is high enough (in comparison to  $N$ ), the introduction of at least somewhat cooperative agents makes the threshold  $M$  more

attainable, especially when we consider low costs of contribution as we do here with  $c = 0.1b$  (see Fig. 1C).

Prior work on the CRD [1, 9, 10, 12, 29, 31, 31, 37], here shown through Fig. 2 where we consider no presence of artificial agents, has interestingly shown that as the risk of disaster when not reaching the target increases, average cooperation also increases, while full cooperation (i.e. a population consisting of only  $C$ -strategists) is not achieved even for high risk situations (see panel A in Fig. 2). In contrast, our findings portrayed in Fig. 3 shows that one can now observe that by adding exogenous AI agents (from a population  $A$ ) of increasing cooperativeness (higher  $p$ ), the average cooperation level of the adapting population decreases, especially for higher risk (i.e.,  $r$ ). Yet, for this region of high risk, when the added agents are non-cooperative, one can observe a higher level of cooperation from the adaptive population.

This trend becomes clearly visible when increasing the number of AI agents in each interacting group, i.e.  $a$ , as can be observed when comparing panels A-C in Fig. 3: The steepness of the boundary between full and almost null  $\bar{C}$  is increased in relation to  $p$ . This transition in the cooperation level of the adapting population is maximized for  $a = M$ , as is clearly illustrated in Fig. 1. So when adding AI agents, the average investment in cooperation by the adapting population may decrease for both low risk and high-risk situations where the AI agents are mostly cooperative, but it has to be maximised when the AI agents are not that much into cooperation.

Overall, the findings in Fig. 1 are consistent with recent experimental work of [17] where algorithm exploitation is proven to be the main driver of the lack of cooperation with AI. Humans were found to act selfishly, by leaving the AI agents less well-off although not out of a competitive wish to end up better off than the machine. With our work we do indeed find that whenever the threshold is met completely by  $a = M$  fully cooperative AI agents, the behavior selected for is less likely be cooperative. However,



**Figure 2: Control treatments with  $a = 0$  that accompany Figs. 4 and 3. A. Average cooperation  $\bar{C}$  (Eq. 9) of the population with adaptive individuals who engage in a collective risk dilemma only with members of the same population ( $a = 0$ ). B. Average group success  $\bar{s}_S$  (Eq. 9) in avoiding the risk of collective loss for the population of adaptive individuals when interacting in groups without agents of fixed behavior ( $a = 0$ ). To produce these figures, we used:  $N = 6$ ,  $M = 3$ ,  $Z = 100$ ,  $\mu = 0.01$ ,  $\beta = 2$ ,  $b = 1$ ,  $c = 0.1$ .**

when the threshold is high enough, the introduction of benevolent AI agents selects for cooperative behavior so that the risk of losing all the endowment is avoided.

## 4.2 Coordination and success in the CRD is achieved through compensation

When studying a CRD, the average group success of that population in avoiding the collective risk,  $\bar{s}_G$  (see Eq. 13), is even more essential, and of course deeply connected to the cooperation level observed in such a population. Fig. 4 visualizes  $\bar{s}_G$  for  $N - a$  adaptive individuals and  $a$  AI agents sampled from population  $A$  that engage in a CRD with  $N = 6$ ,  $M = 3$  and varying  $r$ . Two regions are identified as responsible for high frequencies of group success: First, in Fig. 4A-C, we show that for high risk  $r$  and low probability of cooperation ( $p$ ) by the AI agents introduced in each group, the average success is high (with a frequency  $> 0.9$ ), which is consistent with the findings related to the cooperation level previously studied. Indeed, Fig. 4 appears as result of the superposition between the cooperation level of the adaptive population shown in Fig. 3 and vertical lines that would represent the boundary where high cooperative effort from the fixed agents is sufficient to achieve group success.

Within the boundaries marked by low  $p$ , the boost in the average success rate is due to the evolved behavior in the population alone, since within this boundary the contribution of the AI agents to reach the threshold  $M$  is minimum. Note also that the range for which cooperation is preferred is larger than what one would expect in a scenario where there are no AI agents (see Fig. 2B). Second, for high values of  $p$ , the increase in average group success is especially justified by the introduction of fully cooperative AI agents ( $p = 1$ ). This effect is notably observable for values of  $a$  closer to  $M$ , as one can see from Fig. 4C. With the increment of  $a$ , the average group success also becomes increasingly more dependent on  $p$ , especially in groups where  $N - a < M$  and  $a \geq M$ , since the number of humans per group cannot change in any way the group outcome and the

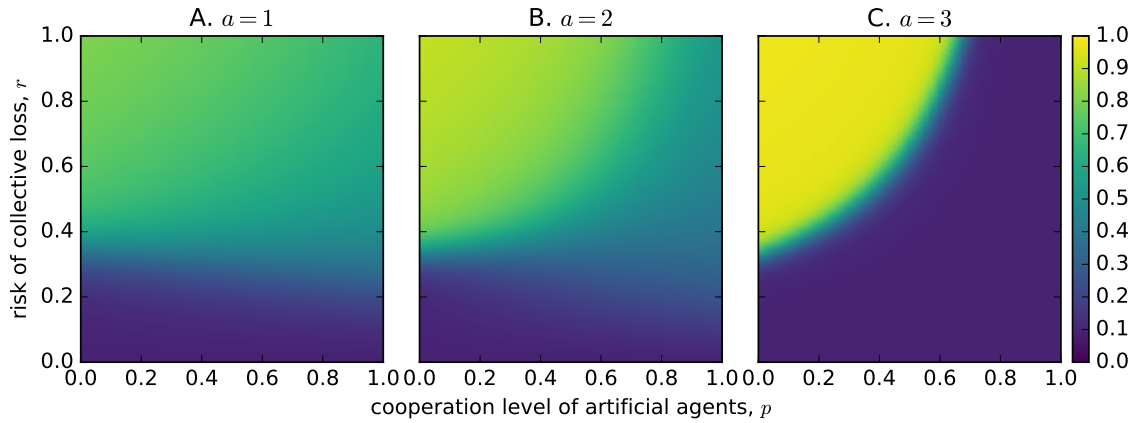
agents are already responsible for most of the possible cooperative effort.

## 4.3 The effect of the number of AI agents $a$ , the group size $N$ and the threshold $M$

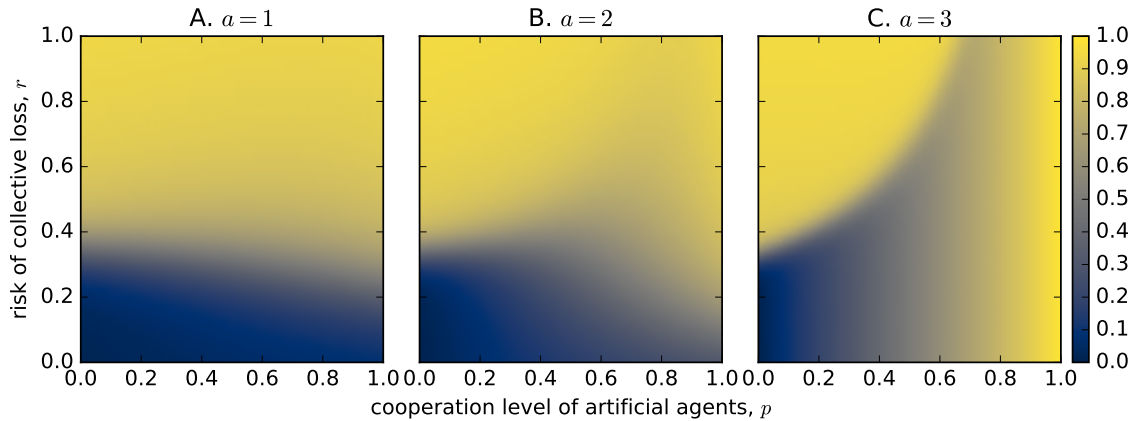
The addition of AI agents to each group corresponds to a change in the game environment itself. Which is why the adaptive population is able to exploit such game transformation, benefiting from having information about the cooperation level of the added AI agents. Here we show how exactly the addition of AI agents with associated probability to cooperate ( $p$ ) affects the stationary distribution of the population dynamics engaging in these hybrid group interactions.

By introducing  $a$  AI agents within a group, the group size  $N$  is transformed into  $N - a$ , reducing thus the number of adapting individuals that can be introduced in each group, and - depending on whether the AI agents are cooperative with  $p = 1$  (or defective with  $p = 0$ ) - one also changes the number of those individuals needed to avoid the risk as  $M \rightarrow M - a$  (or  $M \rightarrow M$ ). This effect can be observed by looking at Figs. 5A. and B. In Fig. 5A, the values  $N - a$  and  $M$  were fixed to be able to observe these trends. Indeed, when adding  $a$  AI agents that contribute nothing to reach threshold  $M$  (with  $p = 0$ ) one can observe that it is the same as adding no AI agents, since  $N - a > M$  and there is still the possibility for the cooperative members of the population to reach the threshold themselves. Furthermore, we find that for bigger values of  $N - a$ , cooperation becomes harder as the stationary distribution shifts to less cooperative states (similarly to the observed effect of increasing  $N$  [31]).

When the added AI agents are cooperative, the transformation  $N \rightarrow N - a$  is also valid for changing the threshold  $M \rightarrow M - a$ . As it is shown in Fig. 5B., the addition of a cooperative AI agents corresponds to both limiting the available places within the group and to lowering the threshold by that number of agents, who are contributing to the cooperative efforts themselves.



**Figure 3:** Average cooperation  $\bar{C}$  (Eq. 11) of an evolving population with  $Z$  individuals who engage in a CRD where  $a$  group members (AI agents) cooperate with probability  $p$ . In order to avoid the collective risk  $r$  of loss, there must be at least  $M$  cooperators, either adaptive  $h$  and/or artificial  $a$ , in a group with size  $N$ . To obtain these results  $N = 6$  and  $M = 3$  were used. For A.  $a = 1$ , for B.  $a = 2$  and for C.  $a = M = 3$ . AI agents are added in each group. Cooperation is highest when risk is high, the number of AI agents not contributing in the group is high ( $N - M = a$ ), so that  $M$  can only be achieved through human cooperators  $M = h$ . Other parameters used for this figure were:  $Z = 100$ ,  $\mu = 0.01$ ,  $\beta = 2$ ,  $b = 1$ ,  $c = 0.1$ .



**Figure 4:** Average group success  $\bar{s}_S$  (Eq. 13) in avoiding the risk of collective loss for each population configuration. In other words, probability of each group of size  $N$  (here,  $N = 6$ ) having at least  $M$  cooperators (here,  $M = 3$ ), averaged over every population state. In A.  $a = 1$ , in B.  $a = 2$  and in C.  $a = M = 3$  AI agents are added in each group. D.  $a = M = 3$  artificial members are added in each group, which corresponds to the threshold of cooperators needed to avoid the collective risk. In C., two boundaries can be clearly identified to achieve group success: the combination of a low AI agent cooperativeness and high risk of collective loss portray the upper left corner semi circle, and the presence of enough cooperative agents to achieve  $M$  cooperators per group on their own draw a vertical line on the right of the graph. Overall, we can see that the introduction of AI agents increases the area of  $r \times p$  where the threshold is met at least 90% (painted in yellow) of the times. For this figure,  $Z = 100$ ,  $\mu = 0.01$ ,  $\beta = 2$ ,  $b = 1$ ,  $c = 0.1$ .

However, when adding AI agents with a fixed non-deterministic behavior, i.e. for  $0 < p < 1$ , the resulting stationary distribution cannot be simply accounted for by a transformation in  $N$  or in  $M$ , as is evinced for example in Fig. 5A. Both the number of AI agents in each group and their cooperation level contribute to shape the resulting stationary distribution of the adaptive population. To support these conclusions one may look at Fig. 5C., where we

show the resulting stationary distributions for the same  $N$ ,  $M$  and  $E = \sum p a$ , where the latter should account for the cooperative effort employed by the addition of agents to the group. We find that even though the  $E$  is fixed at 1, different curves are obtained.

Overall, Fig. 5 exacerbates the relevance of this study by pointing towards the complexity of adapting a population simply by fixing the behavior of some group members present in each interaction.

Indeed, even though we are simply transforming the game environment, the resulting dynamics of the adaptive population are not trivial.

## 5 CONCLUSIONS

In this work we investigated what behavior is selected by social learning in the context of the one-shot CRD, when interactions occur in hybrid groups made of adapting agents (a proxy for human decision-making) and AI agents with fixed probabilistic behaviors (a proxy for average AI behavior). This model is used as a thought experiment to reason about the behavior one could expect in hybrid groups of humans and AI agents. It focuses on mixed-motive situations where there is a conflict between individual and common interests, as well as a risk that of not getting any benefits when not achieving the common goal, which affects the group as a whole. The model allowed us to understand under what conditions the introduction of AI agents in interacting groups actually contributes to an increase in their cooperation and their success rate in reaching a common goal. We have disentangled how the changes in success rate are related to the effort produced by the AI agents: whenever the latter are perceived as highly cooperative, the behavior of the adaptive population evolves to exploit the AI agents' benevolence. This effect was also shown by [17]. However, when the AI agents are low contributors, the adaptive population shifts to compensate those low contributions, as success can only be achieved when the goal is reached. Yet, this is only true for higher risk levels. In general, our model appears to indicate that the success rate of hybrid human-agents teams may be higher for a larger variety of settings than for human-only groups, or even for AI agents alone if they are not able to make the full contribution needed (lower  $p$ ). Thus, our research suggests that there is potential benefit of using AI to nudge cooperation in human groups, which will need to be verified experimentally. Nevertheless, our findings also point towards an unbalanced future for human and AI teams. By working alongside cooperative AI, humans will eventually adapt to relax their own cooperative efforts. Hence, we must either identify AI policies that avoid this scenario and still promote cooperation to avoid collective risks or promote other modes of interaction in-between hybrid teams.

## ACKNOWLEDGMENTS

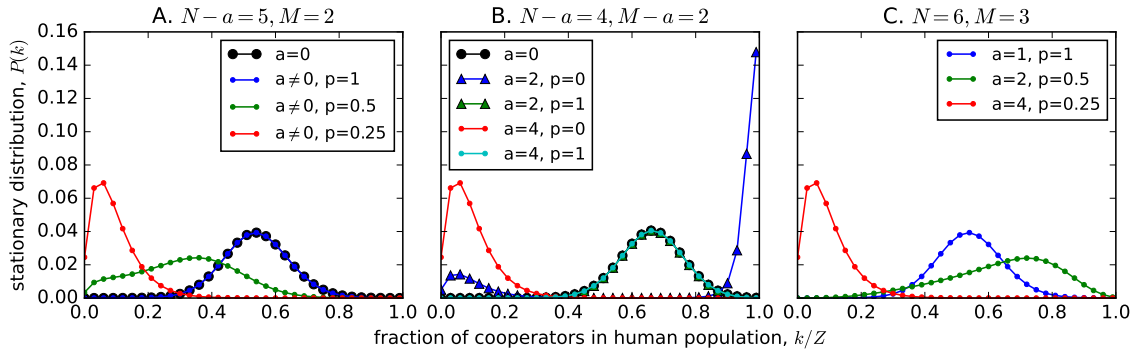
This research was funded by FWO-Vlaanderen under grant agreement no. G054919N. T.L. benefits from the support by the Flemish Government through the AI Research Program and F.C.S. and T.L. by TAILOR, a project funded by EU Horizon 2020 research and innovation program under GA No 952215. F.C.S. acknowledges support from FCT Portugal (UIDB/50021/2020, PTDC/MAT-APL/6804/2020 and PTDC/CCI-INF/7366/2020).

## REFERENCES

- [1] Maria Abou Chakra and Arne Traulsen. 2012. Evolutionary Dynamics of Strategic Behavior in a Collective-Risk Dilemma. *PLoS Computational Biology* 8, 8 (08 2012), 1–7. <https://doi.org/10.1371/journal.pcbi.1002652>
- [2] Charles Bram Cadsby and Elizabeth Maynes. 1999. Voluntary provision of threshold public goods with continuous contributions: experimental evidence. *Journal of Public Economics* 71, 1 (1999), 53–73.
- [3] Colin F. Camerer. 2019. Artificial Intelligence and Behavioral Economics. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press,

- 587–608. <http://www.nber.org/chapters/c14013>
- [4] Alain Cohn, Tobias Gesche, and Michel André Maréchal. 2018. *Honesty in the digital age*. CESifo Working Paper 6996. Center for Economic Studies and the ifo institute. 1–26 pages.
- [5] Jacob W Crandall, Mayada Oudah, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A Goodrich, Iyad Rahwan, et al. 2018. Cooperating with machines. *Nature communications* 9, 1 (2018), 1–12.
- [6] Astrid Dannenberg, Andreas Löschel, Gabriele Paolacci, Christiane Reif, and Alessandro Tavoni. 2015. On the provision of public goods with probabilistic and ambiguous thresholds. *Environmental and Resource Economics* 61, 3 (2015), 365–383.
- [7] Celso M de Melo, Peter Khooshabeh, Ori Amir, and Jonathan Gratch. 2018. Shaping cooperation between humans and agents with emotion expressions and framing. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 2224–2226.
- [8] Celso M de Melo, Stacy Marsella, and Jonathan Gratch. 2019. Human cooperation when acting through autonomous machines. *Proceedings of the National Academy of Sciences* 116, 9 (2019), 3482–3487.
- [9] Elias Fernández Domingos, Jelena Grujić, Juan C Burguillo, Georg Kirchsteiger, Francisco C Santos, and Tom Lenaerts. 2020. Timing uncertainty in collective risk dilemmas encourages group reciprocation and polarization. *iScience* 23, 12 (2020), 101752.
- [10] Elias Fernández Domingos, Jelena Grujić, Juan C Burguillo, Francisco C Santos, and Tom Lenaerts. 2021. Modeling behavioral experiments on uncertainty and cooperation with population-based reinforcement learning. *Simulation Modelling Practice and Theory* 109 (2021), 102299.
- [11] Elias Fernández Domingos, Inês Terrucha, Rémi Suchon, Jelena Grujić, Juan C Burguillo, Francisco C Santos, and Tom Lenaerts. 2021. Delegation to autonomous agents promotes cooperation in collective-risk dilemmas. *arXiv e-prints* (2021), arXiv–2103.
- [12] Kristin Hagel, Maria Abou Chakra, Benedikt Bauer, and Arne Traulsen. 2016. Which risk scenarios can drive the emergence of costly cooperation? *Scientific reports* 6, 1 (2016), 1–9.
- [13] John Harvey, David Golightly, and Andrew Smith. 2014. HCI as a means to prosociality in the economy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2955–2964.
- [14] Muhammad Hasyim, Ade Yolanda Latjuba, Andi Muhammad Akhmar, Kaharuddin Kaharuddin, and Noer Jihad Saleh. 2021. Human-Robots And Google Translate: A Case Study Of Translation Accuracy In Translating French-Indonesian Culinary Texts. *Turkish Journal of Computer and Mathematics Education* (2021).
- [15] Laura Hindersin, Bin Wu, Arne Traulsen, and Julian Garcia. 2019. Computation and simulation of evolutionary game dynamics in finite populations. *Scientific reports* 9, 1 (2019), 1–21.
- [16] Josef Hofbauer, Karl Sigmund, et al. 1998. *Evolutionary games and population dynamics*. Cambridge university press.
- [17] Jurgis Karpus, Adrian Krüger, Julia Tovar Verba, Bahador Bahrami, and Ophelia Deroy. 2021. Algorithm exploitation: Humans are keen to exploit benevolent AI. *iScience* 24, 6 (2021), 102679. <https://doi.org/10.1016/j.isci.2021.102679>
- [18] Jana Kunicova. 2020. Driving the COVID-19 Response from the Center: Institutional Mechanisms to Ensure Whole-of-Government Coordination (English). *Governance and Institutions Responses to COVID-19 Washington, D. C.* (2020).
- [19] Michael W Macy and Andreas Flache. 2002. Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences* 99, suppl 3 (2002), 7229–7236.
- [20] Andrew Mao, Lili Dworkin, Siddharth Suri, and Duncan J Watts. 2017. Resilient cooperators stabilize long-run cooperation in the finitely repeated prisoner's dilemma. *Nature communications* 8, 1 (2017), 1–10.
- [21] Christoph March. 2019. *The Behavioral Economics of Artificial Intelligence: Lessons from Experiments with Computer Players*. CESifo Working Paper 7926. Center for Economic Studies and the ifo institute, Munich. 1–40 pages. <https://ssrn.com/abstract=3485475>
- [22] Manfred Milinski, Torsten Röhl, and Jochem Marotzke. 2011. Cooperative interaction of rich and poor can be catalyzed by intermediate climate targets. *Climatic change* 109, 3 (2011), 807–814.
- [23] Manfred Milinski, Ralf D Sommerfeld, Hans-Jürgen Krambeck, Floyd A Reed, and Jochem Marotzke. 2008. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proceedings of the National Academy of Sciences* 105, 7 (2008), 2291–2294.
- [24] Gopindra S. Nair and Chandra R. Bhat. 2021. Sharing the road with autonomous vehicles: Perceived safety and regulatory preferences. *Transportation Research Part C: Emerging Technologies* 122 (2021), 102885. <https://doi.org/10.1016/j.trc.2020.102885>
- [25] Martin A Nowak. 2006. *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press.
- [26] Barbara Oakley, Ariel Knafo, Guruprasad Madhavan, and David Sloan Wilson. 2011. *Pathological altruism*. Oxford University Press.
- [27] Raquel Oliveira, Patricia Arriaga, Fernando P Santos, Samuel Mascarenhas, and Ana Paiva. 2021. Towards prosocial design: A scoping review of the use of robots





**Figure 5: Stationary distribution in terms of fraction of evolving cooperators for different values of  $N$ ,  $M$ ,  $a$  and  $p$ .** In A. we set  $N - a = 5$  and  $M = 2$ , with results holding for any values of  $N$  and  $a$  that satisfy that condition. One can see that for fixed values of  $N - a$  and  $M$ , the addition of AI agents with  $p = 0$  is equivalent to adding no AI agents, i.e. for  $N - a = N$ . In B. we fix  $N - a = 4$  and  $M - a = 2$ . We find that fixing both  $N - a$  and  $M - a$  when  $a \neq 0$  and  $p = 1$  matches the results obtained for the case of  $a = 0$  with same  $N - a$  and  $M - a$ . In C. we show how setting  $a$  and  $p$ , so that  $E = \sum pa$  is maintained, does not produce equivalent results. Specifically we plot the curves for  $(a = 1, p = 1)$ ,  $(a = 2, p = 0.5)$  and  $(a = 4, p = 0.25)$ , for groups of  $N = 6$  and  $M = 3$ . Other parameters used were:  $Z = 100$ ,  $\mu = 0.01$ ,  $\beta = 2$ ,  $b = 1$ ,  $c = 0.1$  and  $r = 0.5$ .

- and virtual agents to trigger prosocial behaviour. *Computers in Human Behavior* 114 (2021), 106547.
- [28] Jorge M Pacheco, Francisco C Santos, Max O Souza, and Brian Skyrms. 2009. Evolutionary dynamics of collective action in N-person stag hunt dilemmas. *Proceedings of the Royal Society B: Biological Sciences* 276, 1655 (2009), 315–321.
- [29] Jorge M Pacheco, Vitor V Vasconcelos, and Francisco C Santos. 2014. Climate change governance, cooperation and self-organization. *Physics of life reviews* 11, 4 (2014), 573–586.
- [30] Ana Paiva, Fernando P Santos, and Francisco C Santos. 2018. Engineering prosociality with autonomous agents. In *32nd AAAI Conference on Artificial Intelligence*. 7994–7999.
- [31] Francisco C Santos and Jorge M Pacheco. 2011. Risk of collective failure provides an escape from the tragedy of the commons. *Proceedings of the National Academy of Sciences* 108, 26 (2011), 10421–10425.
- [32] Fernando P Santos, Jorge M Pacheco, Ana Paiva, and Francisco C Santos. 2019. Evolution of collective fairness in hybrid populations of humans and agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6146–6153.
- [33] Hirokazu Shirado and Nicholas A Christakis. 2017. Locally noisy autonomous agents improve global human coordination in network experiments. *Nature* 545, 7654 (2017), 370–374.
- [34] Alessandro Tavoni, Astrid Dannenberg, Giorgos Kallis, and Andreas Loschel. 2011. Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences* 108, 29 (2011), 11825–11829.
- [35] Arne Traulsen, Christoph Hauert, Hannelore De Silva, Martin A Nowak, and Karl Sigmund. 2009. Exploration dynamics in evolutionary games. *Proceedings of the National Academy of Sciences* 106, 3 (2009), 709–712.
- [36] Arne Traulsen, Jorge M Pacheco, and Martin A Nowak. 2007. Pairwise comparison and selection temperature in evolutionary game dynamics. *Journal of theoretical biology* 246, 3 (2007), 522–529.
- [37] Vitor V Vasconcelos, Francisco C Santos, and Jorge M Pacheco. 2013. A bottom-up institutional approach to cooperative governance of risky commons. *Nature Climate Change* 3, 9 (2013), 797–801.
- [38] Vitor V Vasconcelos, Francisco C Santos, Jorge M Pacheco, and Simon A Levin. 2014. Climate policies under wealth inequality. *Proceedings of the National Academy of Sciences* 111, 6 (2014), 2212–2216.