To avoid collective disasters, it is better to commit to a flawed AI than to commit the errors ourselves

Inês Terrucha[a,b;*], **Elias Fernández Domingos**[b,c,d], **Pieter Simoens**[a] **and Tom Lenaerts**[b,c,d,e]

[a]IDLab, Ghent University-IMEC
[b]AILab, Vrije Universiteit Brussel
[c]Machine Learning Group, Université Libre de Bruxelles
[d]FARI Institute, Université Libre de Bruxelles-Vrije Universiteit Brussel
[e]Center for Human-Compatible AI, UC Berkeley
ORCiD ID: Inês Terrucha https://orcid.org/0000-0003-2086-1644, Elias Fernández Domingos https://orcid.org/0000-0002-4717-7984, Pieter Simoens https://orcid.org/0000-0002-9569-9373, Tom Lenaerts https://orcid.org/0000-0003-3645-1455

Abstract. Humans make mistakes. Even when a strategy is perfectly crafted to address a problem in hand, the implementation of such a strategy can still be plagued by execution errors if conducted by a human. The noise associated with human execution is one of the main contributors to the growth of the AI industry: autonomous artificial agents are expected to execute the strategies that they are programmed to implement without such noise. However, because the designers of such agents are human, errors may occur on the programming of such agents. This might lead to an AI agent that perfectly executes the strategy it was programmed with, but the strategy is actually misaligned with the intended goals of the human who configured it, a problem of AI alignment. In this work, we explore, by means of an evolutionary game-theoretical model, how errors in the configuration of artificial agents (or in the choice of an artificial delegate) changes the outcome of a collective risk dilemma (CRD). We find that for high risk situations, errors decrease the success rate in comparison with the case of perfect execution. However, it is better to delegate and commit to a flawed strategy executed perfectly by an autonomous agent, than to make execution errors ourselves.

1 Introduction

Over the past years, human society has come to rely more and more on artificially intelligent (AI) applications to facilitate processes of decision-making or the execution of physical tasks. One of the reasons humans might choose to delegate any task [19], even in the human-to-human scenario, is to obtain a better result than they would have on their own, even if it is simply a matter of saving time. Up until now, AI applications have been incredibly successful in elevating human performance in an array of different activities, to name a few: game playing [40, 10], document translation [6, 44], car parking [45], writing text and code [14, 11] or even navigation [27]. Perhaps due to its proven success in these areas, humans now appear to also seek the help of AI to improve their social skills and interpersonal development [26]. It therefore appears that human society is ready to also start using these AI applications in a more social context, which in turn has spurred a growing body of literature in the fields of cooperative, social and affective AI [46, 4, 17].

Most of our social interactions are actually mixed motive scenarios, where different humans might display varied preferences in how to act with regards to others, and these are not necessarily always in conflict but neither are they in full concordance either. Methods such as Inverse Reinforcement Learning [2] have been developed to extract human preferences in a set of different tasks by learning their goals from their actions when acting on their own. However, humans themselves are far from perfect and often make mistakes [41, 34, 28, 25], so that these execution errors might result in a course of action that deviates from their intended preferences. And even if the true goals of the human principal are known, another problem related to how to unambiguously encode the human's preferences into the agent's code arises, which constitutes another part of the AI value alignment problem [36, 16]. As it was first warned by Norbert Wiener in 1960 [49], if we choose to delegate our actions to automated machines whose course of action we might not be able to interrupt, we "better be quite sure that the purpose put into the machine is the purpose which we really desire". With the scale deployment of different language models and other AI applications, we will be gradually observing more and more concerning [43][1] (or sometimes just funny [18]) results of a competent - but misaligned - AI performing a given task. Provided the intention to develop AI applications to aid in the decision-making process of social dilemma situations, for example policy-making for wealth redistribution and climate change action[22, 12], we believe it is of the utmost importance to understand how such AI applications can influence the evolution of human behavior in the long-term.

Similarly to other works concerned with the evolution of humans in the presence of AI [38, 37, 20], we turn to an evolutionary game theoretical approach [39, 33, 9] to tackle this issue. However, we believe to be the first that distinguish between delegated and non-delegated action merely through the moment when mistakes in the strategy occur. In this work, we assume that when humans choose

* Corresponding Author. Email: ines.terrucha@ugent.be.

[1] This event allegedly never happened, but we still consider it to be a very plausible outcome of such an AI application as a result of reward-hacking[42].

or program an agent to act in their place, they make mistakes. This way, a human principal might end up committing to a strategy that - although close enough (assuming a low error rate) - does not fully correspond to their own preferences. However, they will be unable to correct their mistakes during their agent's course of action and are therefore bounded to the strategy of an AI misaligned with their own intentions. In contrast, we assume that when humans do not delegate in our model, they follow their preferences, although not perfectly, so that mistakes can happen when executing their strategies at each round of interaction (similarly to the trembling hand mentioned in other works that consider execution errors in iterated game-play [32, 34, 35, 29, 28]).

Inspired by recent experimental work in delegation to autonomous agents within the context of the collective-risk dilemma[12] (CRD), we also set our model within that realm. As we will further elaborate in the section Methodology, different human preferences arise when playing the CRD, a game where they are confronted with a mixed motive situation where they must coordinate behavior to avoid collective tragedy [31, 7, 12], making it a great tool to abstract human behavior when tackling pressing global issues such as the climate change. With the model we are here proposing on delegation to autonomous agents, we intend to pinpoint where our theoretical assumptions deviate from the experimental findings and where they corroborate the results. Moreover, even if the participants of the experiment do not appear to be keen on delegating again, we want to determine whether delegation within the context of the collective-risk dilemma could prove itself the dominant strategy in the long-term.

2 Methodology

2.1 Evolutionary game theory

As mentioned before, we use an evolutionary game theoretical framework [39, 33, 9, 21, 23, 47, 15, 24] to understand how delegation to autonomous agents affects human behavior in collective social dilemmas in the long term. In summary, such an approach consists of having a population of Z agents that play a game between themselves according to their randomly assigned strategies. At the end of a generational round of play, during which they have accrued payoff in accordance to their strategy and the strategies of their co-players, one agent is randomly chosen to update their strategy. They do this either by imitation learning or mutation. In the case of imitation learning, another individual is randomly selected from the population, and their strategy is more likely to be imitated if their generational payoff was higher than the current strategy of the focal agent chosen to update it. This imitation process is more influential to the evolution of the population the higher the selection strength (parameter $beta$) associated with it. The method is therefore borrowed from the Darwinian competition idea of the survival of the fittest [30], where the fittest strategy would be the one to consistently gather higher payoffs in any population configuration until it takes over the entire population over many rounds of evolution.

2.2 Playing the collective-risk dilemma game

In order for an individual's strategy to propagate within the population, it must harness more fitness than the others. A strategy's fitness is calculated based on how much payoff it collects by playing a game against all the other strategies within the same population. Indeed, the fitness corresponds to the expected payoff a certain strategy gets if it plays a game within all possible different group configurations (given all the other strategies present) within that population. The payoff of each group interaction depends on the game being played, and is dependent not only on the strategy itself, but also on the strategies it is interacting with while playing the game.

We use the collective-risk dilemma game [31, 7, 12]. The game is played in groups of N players, each starting with an endowment of E monetary units. The players have r rounds to contribute to a public account a value between 0 and E/r (and this action space S might be more or less granular [1]). If the group is able to collectively accrue at least $E \times 2$ within the public account by the end of the last round, everyone is able to keep whatever is left from their endowments. Otherwise, there is a risk probability p of everyone losing it all.

Given that the reward is not directly proportional to the contributed effort, but rather dependent on the reaching of a certain threshold, the collective-risk dilemmas belongs to the class of threshold public goods games. On top of this, participants are conflicted with the choice of costly contributions in every round, but are only able to connect a reward in the future end of all rounds. These features associated with risky outcome rather than a certain loss in case of failure to meet the collective target, make it a good game abstraction of some of the most pressing issues faced by our current society, for example the climate change [31]. Which is why, we specifically choose this game to analyze the question of how delegation to autonomous agents might influence the outcome of collective dilemmas. Previous works have turned to automata to analyze how the construction and commitment to a certain strategy over a long number of rounds might influence the outcome of the iterated dilemma [5, 32, 34, 35, 29, 28]. However, few have looked into the construction of a strategy to deal with a one-shot game, with only one reward, but whose outcome is dependent on the many rounds play [8, 12]. We consider this to be a very interesting feature from the point of view of delegation to AI as a commitment device, because even if given the opportunity to adjust the behavior of their agents, one can only evaluate their success (or lack of thereof) at the end of the game, when it might be too late to steer its behavior.

2.3 Defining the strategies

Since this work focuses on the effect of delegation to AI from an evolutionary game theoretical perspective, the definition of the competing strategies within the evolving population must reflect the difference between delegated and non-delegated behavioral profiles. We consider the specific case where a human delegate is able to convey their specific behavior preferences to an artificial agent, for example, if they were the ones programming the agent themselves as in [12]. Within such a scenario, classical game theory would predict no difference between the direct of the principal and the strategy implemented by a program [3], however empirical differences are often observed in the experimental context [3, 12].

We assume the behavioral diversity present within the population of individuals to be the same as the one in the delegation treatment as presented to the participants in [12], where 5 different strategies are made available to play the collective-risk dilemma: the always-0, always-2, always-4, the reciprocal and the compensatory. The first 3 behavioral profiles correspond to fixed strategies where the agents contribute in every round 0, 2 or 4, respectively. The last two behavioral profiles correspond to conditional behavior, where the reciprocal strategy corresponds to contributing 0, 2 and 4 given that the rest of the group members have contributed on average in the previous round 0, 2 or 4, hence reciprocating their behavior in the following round; while the compensatory strategy on the contrary contributes 4,

2 and 0 if the others have contributed 0, 2 or 4 on the previous round, therefore compensating their contributions. Both conditional behavioral profiles are assumed to contribute 2 in the first round when there is no previous round to condition its behavior and to stop contributing once the collective target is achieved.

The individuals within the evolving population are assumed to make mistakes with a certain probability ϵ while executing their strategies. The way we distinguish between delegated and non-delegated action lies precisely in the moment on which the individuals make those mistakes. In our model, we consider that an individual playing on their own might deviate from their intended strategy in each round of play, committing what we may call *execution errors*. If on the contrary, the individual delegated the strategy to an autonomous agent, the latter is assumed to not commit any execution errors during the game-play. However, the principal is considered to be human, and is therefore bounded by the same probability ϵ of committing mistakes. So in the case of the delegated action, we assume the mistakes are committed in the choice or the programming of the autonomous agent. For example, if the protocol is for the human to choose from a set of 5 agents and they wrongly choose an agent that actually does not correspond to their behavior profile (they are reciprocal but choose always-2 by mistake), we call this a *delegate error*. Following [12], another possibility is for the human principal to program their own agents, in which case a human principal who is reciprocal might commit a *program error* and code their agent to contribute 4 (instead of 0) if the others in the group have contributed 0 in the previous round.

In summary our model considers 5 different behavioral profiles and 3 different error modes. The 5 different behavioral profiles follow a program with 4 different settings: one setting to define the action at the first round, and 3 settings to define how much to contribute if others have contributed 0, 2 or 4 in the previous round (we assume at each round, participants can either contribute 0, 2 or 4). The 3 different error modes represent non-delegation action (through *execution errors*) and two different modes of delegated action (*delegate error* and *program error*). Even though we consider a universe of 15 different actions, depending on the specific research question, the evolving population might not consider all the 15 competing against each other.

2.4 Extracting relevant metrics for analysis

The evolutionary process described allows us to calculate the stationary distribution of the competing strategies within an evolving population playing the collective-risk dilemma. The stationary distribution will mirror the long term success of the strategies relative to each other within this dilemma's context, informing us about which ones will prevail if enough time is provided for evolution to reach a stable solution. Given that we will always have at least 5 competing strategies within a population we take the small mutation limit[15, 24, 9] approach to facilitate our analysis. In this case, we assume that the probability of an individual to adopt another strategy through mutation (rather than imitation) is so small that the population spends most of its time in a monomorphic state. When a new strategy appears through mutation, the new strategy either takes over the population or disappears long before another mutation appears. This way, the evolutionary dynamics can be approximated through a Markov chain with the number of states equal to the number of strategies. Our implementation follows version v0.1.12 of the recently published hybrid C++/Python library EGTtools [9, 13], whose methods allows us to easily retrieve the stationary distribution provided the game payoffs associated with each strategy and group composition and the relevant population parameters such as the selection strength $beta$ and the number of individuals Z. Given the noisy nature of the model in analysis, the game payoffs have to be previously estimated through a series of simulated play within each group composition before being used as input to the small mutation limit methods described in [9], so that the number of simulations $\#sim$ used to make this estimation is also a relevant parameter to reproduce the results exhibited in the section Results and Discussion.

With the stationary distribution, we are then able to calculate game-related metrics that are relevant for our analysis in a time-independent way. For example, we can calculate the probability of a group made of N players of each strategy achieving success in the collective-risk dilemma. A weighted average of that quantity with each strategy's stationary distribution will then return the average success rate associated with a population where those strategies are present, similarly to what was done in [48, 7]. By calculating this quantity for both a population where only no-delegation is allowed and a population where only delegation is allowed, we can then infer which would be the best long-term solution for the collective-risk dilemma for any combination of other parameters, for example our parameter of interest, the probability of error (which is precisely what we will demonstrate in the Results section).

If instead of populations where only delegation or no-delegation strategies are considered, we can instead explore a hybrid population where both delegation and no-delegation strategies are possible. In this case, we are also able to define a delegation rate, by summing over the stationary distributions associated with delegation strategies only to examine the prevalence of delegation within such a hybrid population. With this delegation vs no-delegation analysis we are also able to define the average fitness associated with delegation and no-delegation strategies: through a weighed average between the payoffs associated with each strategy in their monomorphic state and their respective stationary distribution for each case, delegation and no-delegation. Such a method allows us to understand when does delegation become dominant in a population where it is optional, as shown in the section Results.

3 Results and Discussion

3.1 When the error probability is small, groups are more successful in the CRD when they delegate

This first section of Results is dedicated to answering the question: does delegation increase the success rate of groups of individuals tackling a CRD? To answer this question we compare the success rates achieved by 3 different populations, one where individuals do not delegate and two where individuals delegate through different methods. As previously discussed in the section Methodology, we assume that when individuals do not delegate, they commit *execution* errors, whereas when they delegate they might commit *delegate* errors - by choosing the wrong agent from a group of pre-programmed agents - or *program* errors by wrongly programming the settings of an agent. These errors are viewed as deviations from their true preferences, which we assume are limited to 5 different strategies: **R**eciprocal (players start by contributing 2 and then proceed to contribute 0, 2 or 4 if others contributed in the previous round 0, 2 or 4, respectively), **C**ompensatory (players start by contributing 2 and then proceed to contribute 4, 2 or 0 if others contributed in the previous round 0, 2 or 4, respectively), always-**0** (players always contribute 0), always-**2** (players always contribute 2) and always-**4** (players always contribute 4).

The success rates obtained for each evolving population (*execution*, *program* and *delegate*) with regards to different values of error probability are shown in Fig. 1. As is shown in the figure, success rates for a population with *execution* errors (no-delegation case) decrease rapidly for small values of error probabilities, resulting in the lowest success rates between the 3 represented cases when error probability ≤ 0.15, at which point it returns higher success rates than the case where individuals program their own agent (*program*), but still lower than when they must choose an agent from a group of pre-set delegates (*delegate*) for error probabilities ≤ 0.4. For this figure and the following analysis we have chosen to work with a selection strength $beta = 0.05$, however similar results (delegation being more successful in avoiding collective risk than no-delegation for small error probabilities) can be obtained for other values of $beta$ as we show in the Supplemental Information (SI). We find these results to be especially interesting since, for small values of error probability, they corroborate the trends observed experimentally [12] using the same CRD parameters as this model ($p = 0.9$, $N = 6$, $r = 10$, $S = 0, 2, 4$, $E = 40$). Indeed, the region of small error probability is also the most relevant one for designing future experiments as it is unlikely that human participants would commit errors with frequency higher than 0.5, or 1 error for every 2 moves.

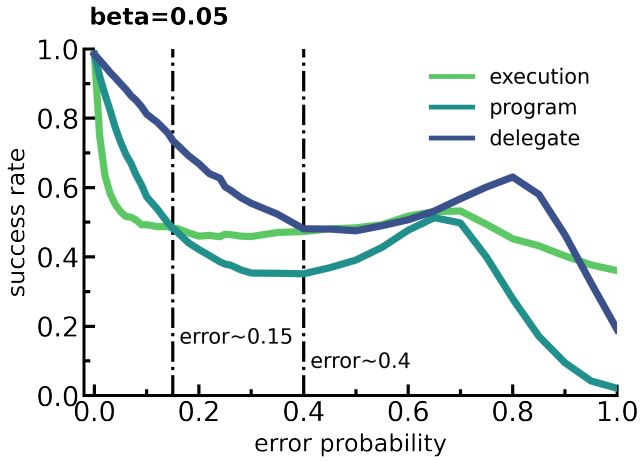

Figure 1. Success rate observed in terms of error probability for 3 different evolving populations: one where individuals do not delegate and therefore only commit *execution* errors; one where they delegate to an agent they choose from a group of pre-set delegates, possibly committing *delegate* errors; and one where they program themselves their agent and can therefore commit *program* errors. Each different line corresponds to a population, colored according to the legend on the top right corner of the figure. Two dotted-dashed lines indicate when (at error$\simeq 0.15$) programming your own agent and (at error$\simeq 0.4$) choosing a delegate stops being more successful than not delegating (and just playing by themselves). Selection strength is indicated on top left of the Figure, and corresponds to $beta = 0.05$. The other parameters used to reproduce this image are: $p = 0.9$, $r = 10$, $E = 40$, $S = 0, 2, 4$, $N = 6$, $Z = 100$, $\#sim = 1000$.

In order to better assess how committing errors influences the success rate obtained by the different populations, an analysis of the transitions observed between each population monomorphic states at error equal to 0 is conducted. This scenario is identical for all the different error cases - *execution*, *program* and *delegate* - since it corresponds to the case where no errors are committed and individuals play exactly with the strategies that they intended to use, independently of whether they delegate or not. Figure 2 represents the Markov chain that illustrates the transitions between the 5 different monomorphic states, one for each strategy competing within the population. As we can see, for a risk probability p as high as 0.9 (see [8] for a deeper analysis on the CRD), strategies *2*, *R* and *C* each occupy around 33% of the stationary distribution, together fully dominating the population, which justifies the very high values of success rate obtained for error$= 0$ in Fig. 1. Within this figure, arrows represent invasion relationships between strategies when the fixation probability of a mutant strategy in the monomorphic state of another is higher than random drift (strategy connected to arrow-tail node is invaded by strategy connected to arrow-head node). Strategy *4* is observed to be invaded by all strategies, while no strategy invades any of the others. Moreover, a random drift triad is formed around strategies *2*, *R* and *C*, since they all accrue the same fitness when playing among each other in a group, only random drift determines which one takes over the population after enough time.

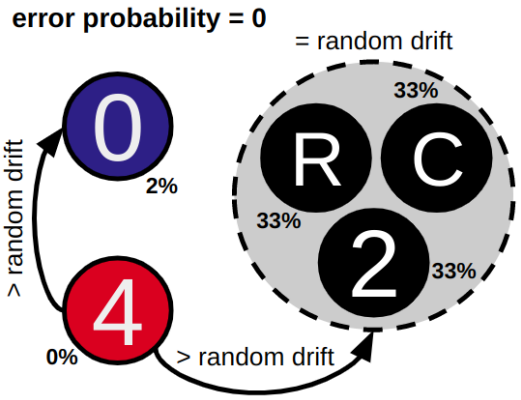

Figure 2. Representation of the Markov chain that illustrations the transitions between the monomorphic states of the population when individuals commit no errors when implementing their strategies. Nodes *R*, *C*, *0*, *2* and *4* stand for the strategies Reciprocal, Compensatory, always-0, always-2 and always-4, respectively. The stationary distribution of each strategy within the evolutionary process is indicated with a number (approximated to the integer) and the $\%$ sign next to each node. Arrows represent transitions where a mutant from a strategy (arrow head) is able to invade a monomorphic population of another (arrow tail) with a fixation probability higher than random drift. A dashed circle around *R*, *C* and *2* represents the mutual random drift like fixation that these strategies exhibit with one another. The values in the image are obtained for any of three previously mentioned error type populations (*execution*, *program* or *delegate*) but with error probability equal to 0. The other parameters used to reproduce this image are: $p = 0.9$, $r = 10$, $E = 40$, $S = 0, 2, 4$, $N = 6$, $beta = 0.05$, $Z = 100$, $\#sim = 1000$.

Figure 3 represents the stationary distribution obtained for each different strategic group in terms of increasing error probabilities for both a delegation (*program* errors are made) and the no-delegation case (*execution* errors occur). Although the random drift triad as represented in Fig. 2 encompassing strategies 2, R and C is broken when errors are present (these perturb the random drift relationship allowing for small variations in the stationary distribution of these strategies), their stationary distributions follow similar trends on average with regards to increasing error probability for the delegation case. Therefore, to simplify our analysis, in Fig. 3 only the average stationary distribution of these strategies is shown, so that the colored legend only indicates the 3 different strategic groups *0*, $\overline{2RC}$ and *4*. In the figure, full lines represent the stationary distribution for the delegation case (here, *program*) and dashed lines for the no-delegation case (*execution*). As in Fig. 1, a dotted-dashed line de-

termines the boundary below which delegation is always more successful than no-delegation strategies in avoiding collective tragedy in a CRD. This boundary appears to almost coincide with a change in the stationary distributions of the triad $\overline{2RC}$ and *0* within the *program* errors population, where the the stationary distribution appears to start decreasing with increasing error probability, contrarily to *0*, whose stationary distribution starts increasing at this point. In the *execution* errors population no great changes occur at this point, from which we infer that the change in success rate at this boundary is mainly caused by the lack of robustness of delegation strategies in the presence of higher error probabilities rather than any big changes caused by higher error rates for no-delegation strategies. For further detail, the figure corresponding to the stationary distribution of the 5 different strategies can be found in the SI and shows that indeed the three strategies within the $\overline{2RC}$ group change similarly with regards to the error probability, allowing us to make such a simplification so that the results can be more easily visualized. Also, note that in Fig. 3 the focus is on the delegation case where people program their agents and might commit *program* errors, simply because it centers the analysis on the lower boundary at error$\simeq 0.15$, below which delegation always results in higher success rates for the population. As it is shown in the SI, a comparable study can be done to the case where people choose a delegate agent and can commit *delegate* errors in this process (although in this case, the growing of the stationary distribution of *0* within delegation alone is not sufficient to explain the change in the success rate trends, one must also consider the slight decrease of the presence of *0* at this point within the no-delegation case).

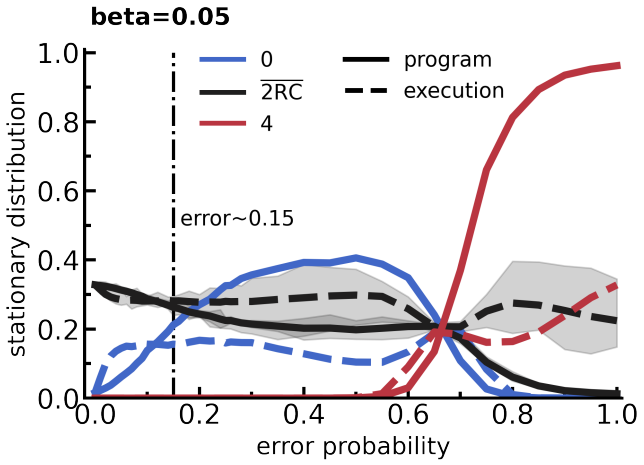

Figure 3. Stationary distribution of the three strategy groups represented in Fig. 2 in terms of error probability for the case of *execution* and *program* errors. The strategy groups are *0*, $\overline{2RC}$ and *4* as represented by the colored legend. The line used to represent the group $\overline{2RC}$ indicates the averaged values of the stationary distribution of the three strategies *2*, *R* and *C* and the shadow filling around it the 95% confidence interval around that average. The figure shows how the stationary distribution of no-delegation strategies, represented through *execution* errors (full lines), differs from delegation strategies, here focused on *program* errors (dashed lines). A dotted-dashed line at error probability $\simeq 0.15$ marks where the success rate of a population of *execution* errors starts to surpass the success of a population of *program* errors. Selection strength is indicated on top left of the Figure, and corresponds to $beta = 0.05$. The other parameters used to reproduce this image are: $p = 0.9$, $r = 10$, $E = 40$, $S = 0, 2, 4$, $N = 6$, $Z = 100$, $\#sim = 1000$.

3.2 *In the long term, delegation rates are higher when agents are pre-set rather than programmable*

In this section of Results, we try to pinpoint which delegation methods are more likely to be adopted by a hybrid population in the long-term. In order to answer this question, two hybrid populations are compared: one where delegation strategies where individuals program their own agents and might commit *program* errors compete with no-delegation strategies (with *execution* errors); and another where the competition is between no-delegation strategies and delegation strategies where individuals have to choose a delegate from a pre-set group of agents (associated with *delegate* errors). Again, within each population we will consider that there are 5 different strategic behaviors in competition, although in this case each delegation itself, or the lack of thereof, is also part of the strategy taken by each individual; existing therefore 10 different competing strategies within each hybrid population.

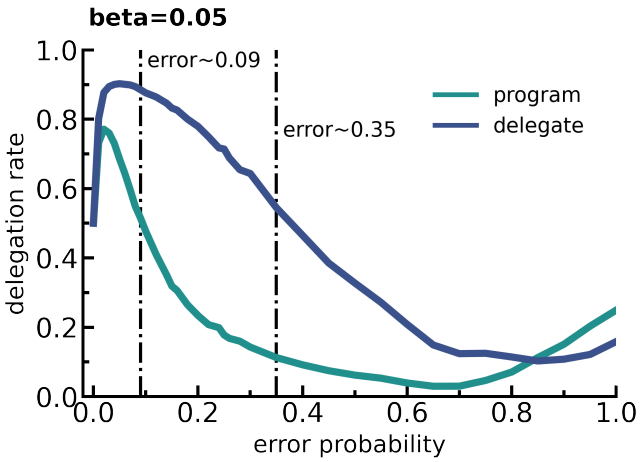

Figure 4. Delegation rates in terms of error probability for two different hybrid populations where individuals can either use a delegation or a no-delegation strategy. In one population the delegation strategy is represented by *program* errors, in the other by *delegate* errors, and these are distinguished by color following the legend on the top right corner of the figure. The no-delegation strategies are represented by *execution* errors in both cases. Within each population there are therefore 10 competing strategies: *R*, *C*, 0, 2 and 4 for both delegation and the no-delegation case. With a dotted-dashed line we represent when delegation stops being the most adopted strategies within the population (the summed stationary distribution of delegated strategies decreases to ≤ 0.5 with increasing error): at $error \simeq 0.09$ for *program* errors and at $error \simeq 0.35$ for *delegate* errors. Selection strength is indicated on top left of the Figure, and corresponds to $beta = 0.05$. The other parameters used to reproduce this image are: $p = 0.9$, $r = 10$, $E = 40$, $S = 0, 2, 4$, $N = 6$, $Z = 100$, $\#sim = 1000$.

Figure 4 shows the variation of delegation rates obtained within each hybrid population (one where delegation is given by *program* errors and the other where it is represented by *delegate* choice errors). With a dashed line we denote when the delegation rate, or in other words, the summed stationary distributions of all delegation strategies within a population, goes below 50%, therefore marking the point where delegation stops being the most adopted strategy within the population. In the experimental study conducted in [12] (with similar game parameters), participants have answered that they would rather delegate again if they were given the opportunity to program their agents, rather than choose a delegate from a group of pre-set ones. Interestingly, this model predicts that delegation rates

will be higher when delegation is made through a pre-set agent for almost all error probabilities (and all the plausible ones, as an error probability above 0.8 is unlikely in a real-world context).

With Fig. 5 we take a closer look into the hybrid population where delegation happens through the programming of an agent (although a similar analysis can be taken for the case of choosing a delegate, see SI for the analogous image). Fig. 5 illustrates the changes in the stationary distribution of the different strategies (10 in total, although we reduce the visualization to 6 by once again grouping the triad composed by the strategies *2*, *R* and *C* within both delegation and no-delegation strategic groups). The dashed line at error probability $\simeq 0.09$ denotes the boundary where delegation stops being the most adopted group of strategies within the population, following Fig. 4. It may be observed that this boundary coincides with the moment when the no-delegation triad $\overline{2RC}$ starts to increase their averaged stationary distribution, contrarily to its delegation homologue that starts decreasing. In the SI it is possible to consult the stationary distribution of all the 10 individual strategies for both the aforementioned hybrid populations, from which similar conclusions may be drawn.

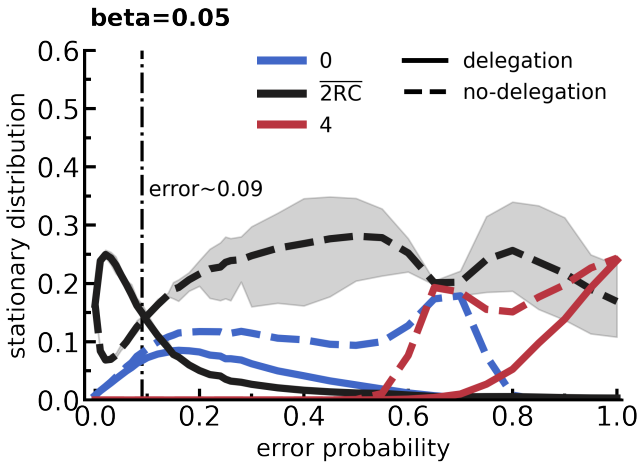

Figure 5. Stationary distribution of the three strategy groups represented in Fig. 2 in terms of error probability of both delegation and no-delegation strategies in a hybrid population where delegation is represented by *program* errors. The colored legend indicates the group (following Fig. 2 notation) to which the strategy belongs: *0*, $\overline{2RC}$ or *4*. Delegation is represented with full lines and no-delegation with dashed lines. The line used to represent the group $\overline{2RC}$ indicates the averaged values of the stationary distribution of the three strategies *2*, *R* and *C* and the shadow filling around it the 95% confidence interval around that average. The dotted-dashed line at error probability$\simeq 0.09$ indicates where delegation rate decreases to ≤ 0.5 within the hybrid population as observed in Fig. 4. Selection strength is indicated on top left of the Figure, and corresponds to $beta = 0.05$. The other parameters used to reproduce this image are: $p = 0.9$, $r = 10$, $E = 40$, $S = 0, 2, 4$, $N = 6$, $Z = 100$, $\#sim = 1000$.

4 Conclusion

With this modelling work we are able to draw interesting conclusions with regards to the success rate and delegation rate obtained for different delegation methods used to tackle the collective-risk dilemma. We find that for small error probabilities our results corroborate previous experimental work [12] in finding that groups made of individuals that delegate achieve higher success rates in avoiding the collective disaster. Moreover, we find that delegation is not only more successful in solving the dilemma, in the long term it is also the most adopted strategy within a hybrid population where delegation and no-delegation strategies compete with each other.

Relating specifically to the problem of solving a CRD with high risk, with this work we find that introducing errors perturbs the random drift triad composed by the three strategies responsible for holding high success rates - reciprocal, compensatory and always-2. All the interesting conclusions, such as the identification of the boundaries when delegation is more successful in solving the dilemma, or when delegation is more adopted within the population, relate to the changes observed in the average stationary distribution of these three strategies. Previous work [7, 8, 12] had already identified the importance of these three strategies to reach high levels of success, and this work adds to this literature by showing how the presence of errors perturbs this triad and immediately lowers the expected success rates even for low error probabilities. However, in real-world scenarios mistakes happen, and our work shows that it is better to commit them through an artificial delegate than when implementing the strategy ourselves.

To our knowledge, it is also the first time that a delegation mechanism is formalized within the context of evolutionary game theory without having to consider different behavioral strategies, associated delegation costs or self-interested agents that specifically play a delegation game. In itself, we find this to be an interesting contribution to game-theoretical problems involving delegation to autonomous agents as often in the experimental settings or real-world applications, the main difference between delegated and non-delegated action lies in the moment when the commitment is made, which in this modelling work we connect with the moment when errors might be committed.

Acknowledgements

I.T. received funding from FWO under grant agreement no. G054919N. E.F.D. is supported by an F.R.S-FNRS (Fonds de la Recherche Scientifique) Chargé de Recherche grant (nr. 40005955). T.L. is supported by an F.W.O. project (grant nr. G054919N) and two F.R.S.-FNRS PDR (grant numbers 31257234 and 40007793). E.F.D. and T.L. are supported by Service Public de Wallonie Recherche under grant n° 2010235–ariac by digitalwallonia4.ai. T.L. and P.S. acknowledge the support by the Flemish Government through the AI Research Program. P.S. acknowledges the support of F.W.O. grant nr. G054919N. T.L. acknowledges the support by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

[1] Ricardo Arlegi, Juan M Benito-Ostolaza, and Nuria Osés-Eraso, 'Participation in and provision of public goods: Does granularity matter?', *Journal of Economic Interaction and Coordination*, **16**, 265–285, (2021).

[2] Saurabh Arora and Prashant Doshi, 'A survey of inverse reinforcement learning: Challenges, methods and progress', *Artificial Intelligence*, **297**, 103500, (2021).

[3] Jordi Brandts and Gary Charness, 'The strategy versus the direct-response method: a first survey of experimental comparisons', *Exp Econ*, **14**, 375–398, (2011).

[4] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel, 'Cooperative ai: machines must learn to find common ground', *Nature*, **593**(7857), 33–36, (2021).

[5] Pedro Dal Bó and Guillaume R Fréchette, 'Strategy choice in the infinitely repeated prisoner's dilemma', *American Economic Review*, **109**(11), 3929–52, (2019).

[6] Erik De Vries, Martijn Schoonvelde, and Gijs Schumacher, 'No longer lost in translation: Evidence that google translate works for comparative bag-of-words text applications', *Political Analysis*, **26**(4), 417–430, (2018).

[7] Elias Fernández Domingos, Jelena Grujić, Juan C Burguillo, Georg Kirchsteiger, Francisco C Santos, and Tom Lenaerts, 'Timing uncertainty in collective risk dilemmas encourages group reciprocation and polarization', *Iscience*, **23**(12), 101752, (2020).

[8] Elias Fernández Domingos, Jelena Grujić, Juan C Burguillo, Francisco C Santos, and Tom Lenaerts, 'Modeling behavioral experiments on uncertainty and cooperation with population-based reinforcement learning', *Simulation Modelling Practice and Theory*, **109**, 102299, (2021).

[9] Elias Fernández Domingos, Francisco C Santos, and Tom Lenaerts, 'Egttools: Evolutionary game dynamics in python', *Iscience*, **26**(4), (2023).

[10] Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al., 'Human-level play in the game of diplomacy by combining language models with strategic reasoning', *Science*, **378**(6624), 1067–1074, (2022).

[11] Yunhe Feng, Sreecharan Vanam, Manasa Cherukupally, Weijian Zheng, Meikang Qiu, and Haihua Chen, 'Investigating code generation performance of chat-gpt with crowdsourcing social data', in *Proceedings of the 47th IEEE Computer Software and Applications Conference*, pp. 1–10, (2023).

[12] Elias Fernández Domingos, Inês Terrucha, Rémi Suchon, Jelena Grujić, Juan C Burguillo, Francisco C Santos, and Tom Lenaerts, 'Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma', *Scientific reports*, **12**(1), 1–12, (2022).

[13] Elias Fernández Domingos. Egttools: Toolbox for evolutionary game theory, July 2023.

[14] Tira Nur Fitria, 'Artificial intelligence (ai) technology in openai chatgpt application: A review of chatgpt in writing english essay', in *ELT Forum: Journal of English Language Teaching*, volume 12 (1), pp. 44–58, (2023).

[15] Drew Fudenberg and Lorens A Imhof, 'Imitation processes with small mutations', *Journal of Economic Theory*, **131**(1), 251–262, (2006).

[16] Iason Gabriel, 'Artificial intelligence, values, and alignment', *Minds and machines*, **30**(3), 411–437, (2020).

[17] Riccardo Gervasi, Federico Barravecchia, Luca Mastrogiacomo, and Fiorenzo Franceschini, 'Applications of affective computing in human-robot interaction: State-of-art and challenges for manufacturing', *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, **237**(6-7), 815–832, (2023).

[18] David Ha, 'Reinforcement learning for improving agent design', *Artificial life*, **25**(4), 352–365, (2019).

[19] John R Hamman, George Loewenstein, and Roberto A Weber, 'Self-interest through delegation: An additional rationale for the principal-agent relationship', *American Economic Review*, **100**(4), 1826–1846, (2010).

[20] The Anh Han, Cedric Perret, and Simon T Powers, 'When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games', *Cognitive Systems Research*, **68**, 111–124, (2021).

[21] Laura Hindersin, Bin Wu, Arne Traulsen, and Julian García, 'Computation and simulation of evolutionary game dynamics in finite populations', *Scientific reports*, **9**(1), 6946, (2019).

[22] Arend Hintze and Peter T Dunn, 'Whose interests will ai serve? autonomous agents in infrastructure use', *Journal of Mega Infrastructure & Sustainable Development*, **2**(sup1), 21–36, (2022).

[23] Josef Hofbauer, Karl Sigmund, et al., *Evolutionary games and population dynamics*, Cambridge university press, 1998.

[24] Lorens A Imhof, Drew Fudenberg, and Martin A Nowak, 'Evolutionary cycles of cooperation and defection', *Proceedings of the National Academy of Sciences*, **102**(31), 10797–10800, (2005).

[25] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein, *Noise: a flaw in human judgment*, Hachette UK, 2021.

[26] Alyson Krueger, 'We need to talk - just as soon as i consult chatgpt', *New York Times*, (2023).

[27] Wildi Kusumasari, Friska M Ilmi, Estiara Ellizar, and Yos Y Rabung, 'Evaluating the use of google maps as navigation application by identifying hazards and assessing risks using hira matrix', in *IOP Conference Series: Earth and Environmental Science*, volume 1065 (1), p. 012046. IOP Publishing, (2022).

[28] Luis A Martinez-Vaquero, The Anh Han, Luís Moniz Pereira, and Tom Lenaerts, 'Apology and forgiveness evolve to resolve failures in cooperative agreements', *Scientific reports*, **5**(1), 10639, (2015).

[29] Luis A Martinez-Vaquero, Francisco C Santos, and Vito Trianni, 'Signalling boosts the evolution of cooperation in repeated group interactions', *Journal of the Royal Society Interface*, **17**(172), 20200635, (2020).

[30] Smith J Maynard, 'The logic of animal conflicts', *Nature*, **246**, 15–18, (1973).

[31] Manfred Milinski, Ralf D Sommerfeld, Hans-Jürgen Krambeck, Floyd A Reed, and Jochem Marotzke, 'The collective-risk social dilemma and the prevention of simulated dangerous climate change', *Proceedings of the National Academy of Sciences*, **105**(7), 2291–2294, (2008).

[32] Martin Nowak and Karl Sigmund, 'A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game', *Nature*, **364**, 56–58, (1993).

[33] Martin A Nowak, *Evolutionary dynamics: exploring the equations of life*, Harvard university press, 2006.

[34] Martin A Nowak, Karl Sigmund, and Esam El-Sedy, 'Automata, repeated games and noise', *Journal of Mathematical Biology*, **33**(7), 703–722, (1995).

[35] Flavio L Pinheiro, Vitor V Vasconcelos, Francisco C Santos, and Jorge M Pacheco, 'Evolution of all-or-none strategies in repeated public goods dilemmas', *PLoS computational biology*, **10**(11), e1003945, (2014).

[36] Stuart Russell, 'Human-compatible artificial intelligence', *Human-like machine intelligence*, 3–23, (2021).

[37] Fernando P Santos, Jorge M Pacheco, Ana Paiva, and Francisco C Santos, 'Evolution of collective fairness in hybrid populations of humans and agents', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33 (01), pp. 6146–6153, (2019).

[38] Hirokazu Shirado and Nicholas A Christakis, 'Locally noisy autonomous agents improve global human coordination in network experiments', *Nature*, **545**(7654), 370–374, (2017).

[39] Karl Sigmund, 'The calculus of selfishness', in *The Calculus of Selfishness*, Princeton University Press, (2010).

[40] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al., 'A general reinforcement learning algorithm that masters chess, shogi, and go through self-play', *Science*, **362**(6419), 1140–1144, (2018).

[41] Herbert A Simon, 'Bounded rationality', *Utility and probability*, 15–18, (1990).

[42] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger, 'Defining and characterizing reward gaming', *Advances in Neural Information Processing Systems*, **35**, 9460–9471, (2022).

[43] Guardian staff, 'Us colonel retracts comments on simulated drone attack 'thought experiment'', *The Guardian*, (2023).

[44] Yosuke Takakusagi, Takahiro Oike, Katsuyuki Shirai, Hiro Sato, Kio Kano, Satoshi Shima, Keisuke Tsuchida, Nobutaka Mizoguchi, Itsuko Serizawa, Daisaku Yoshida, et al., 'Validation of the reliability of machine translation for a medical article from japanese to english using deepl translator', *Cureus*, **13**(9), (2021).

[45] Diya Thomas and Binsu C Kovoor, 'A genetic algorithm approach to autonomous smart vehicle parking system', *Procedia Computer Science*, **125**, 68–76, (2018).

[46] Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle CM Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, et al., 'Ai for social good: unlocking the opportunity for positive impact', *Nature Communications*, **11**(1), 2468, (2020).

[47] Arne Traulsen, Christoph Hauert, Hannelore De Silva, Martin A Nowak, and Karl Sigmund, 'Exploration dynamics in evolutionary games', *Proceedings of the National Academy of Sciences*, **106**(3), 709–712, (2009).

[48] Vítor V Vasconcelos, Francisco C Santos, Jorge M Pacheco, and Simon A Levin, 'Climate policies under wealth inequality', *Proceedings of the National Academy of Sciences*, **111**(6), 2212–2216, (2014).

[49] Norbert Wiener, 'Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers.', *Science*, **131**(3410), 1355–1358,

(1960).