9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Delegation to autonomous agents: a key to overcome past failure and focus on the collective target ahead

Keywords: delegation, autonomous agents, behavioral experiments, collective-risk dilemma, surprise restart

Extended Abstract

Background and Impact

Even though we know that **delegation to autonomous agents** might affect the outcome of our decision-making, it is still unknown in which situations it will have a positive or a negative impact in comparison to what people would do when acting on their own. However, due to the digital transformation we have been witnessing in many sectors, it is only a matter of time before strategic **individual decisions that impact collective goods** will also have to be made virtually through the use of autonomous agents [1]. It is therefore critical to understand how effective is this delegation mechanism in promoting the provision of the public goods when global issues put them at risk. In [1], it is hypothesized that delegation to these algorithmic agents can act as a commitment device: by being asked to program an artificial agent to act in their place in a public goods game, individuals appear to commit to the provision of this public good. Without delegation, the presence of free-riders is observed to disturb other group members cooperative efforts who switch to punishing strategies. We aim to contribute by answering the question: Can artificial delegates still serve as commitment devices if their principals are given the opportunity to revise their algorithms when confronted with new collective challenges?

Data collection and Methods

Following [1], we design a series of **behavioral experiments** based on the **collective-risk dilemma game** [2]: a threshold public good game, where a group of individuals must choose how much to contribute from their Private Accounts to a Public Account over a finite number of rounds. If they fail to reach a collective target, they risk losing whatever remains in their Private Accounts at the end of the game. Moreover, they are only allowed to make small contributions in each round: in this way, they are forced to play the game round-by-round, so that their initial strategies are vulnerable to adaptation once they realize how others are playing the same game. Specifically our game is defined by the parameters: group size of 4, 10 rounds, participants can contribute 0, 2 or 4 in each round, initial endowment of 40 in each individual Private Account, risk probability of 50%, and collective target is 80. To answer the aforementioned questions, we collect data on both a delegation - where participants are asked to program an agent to contribute on their behalf before the game starts - and a no-delegation - where participants make their own contributions in every round - condition. After playing a game of collective risk (Game 1), participants are then asked to play a second game (Game 2) with the same group with which they played Game 1. The second game appears to them as a **surprise restart** [3] of the first, therefore deterring the use of direct reciprocity strategies in the first game. Data was collected online: the experimental platform was designed in LIONESS Lab and the participants were recruited through Prolific.

Results and Discussion

With this work we find that, even though delegation is not more successful in avoiding the collective tragedy (illustrated in the first column of Fig. 1 and confirmed by a Fisher's exact test that returns **no statistical difference between delegation and no-delegation success rates** in either of the games played), such appears to result from a incorrectly fine-tuning the agent's algorithm, rather than from participants purposefully not wanting to reach the target. In fact, on average, **groups from the delegation treatment accumulate more on their Public Account** than groups in the no-delegation treatment (p-value$= 0.006$ in Game 1 and p-value$= 0.024$ in Game 2, Welch's t-test). This can be observed in the second column of Fig. 1, where in Game 2 a greater negative deviation of the Public Account values is evident in the case of no-delegation, suggesting that for this treatment, failed groups actually gave up early on reaching the target. Indeed, the last column of Fig. 1 further illustrates this idea, by showing the round-by-round cumulative group contributions to the Public Account, here not only separated by treatment condition but also by group outcome in the game. In the top panel of this column in Fig. 1, pertaining to Game 1 results, we see that successful groups overshoot the target in the delegation group, whereas the ones in no-delegation stop contributing once the target is reached; failed groups in the delegation treatment undershoot the target, but never stop contributing until the end, while no-delegation groups contributions, on average, flatten around round 4, seemingly giving up on the target. In the bottom panel of this last column of Fig. 1, Game 2 results are shown, which corroborate the idea that **delegation groups only miss the target by mistake**, since both the successful and the failed groups in this condition are now, on average, closer to what would be the efficient round-by-round contribution black dashed-line; which contrasts with the **failed no-delegation groups which, in Game 2, seem to give up on the collective target** even earlier (around round 2). This story is especially compelling if one takes into consideration the results presented in the third column of Fig. 1, where we see that **inequality between individual gains is greater, on average, within delegation groups** than no-delegation ones in both Game 1 and Game 2 (p-value$\ll 0.001$ for both Game 1 and Game 2, Welch's t-test). Unfortunately, collective-risk scenarios are posed to repeat themselves, and the next one always appears as a surprise, that we must tackle within the same group or community within which we endured the latter. As humans, we take our lessons and build grudges over the past, which may set us to fail again in the future. This research shows that **delegation to autonomous agents might be a key to overcome past failures, even when inequality was experienced, and reset our focus to reach the collective target ahead**.

References

[1] 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.

[2] 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.

[3] Ananish Chaudhuri. Belief heterogeneity and the restart effect in a public goods game. *Games*, 9(4):96, 2018.

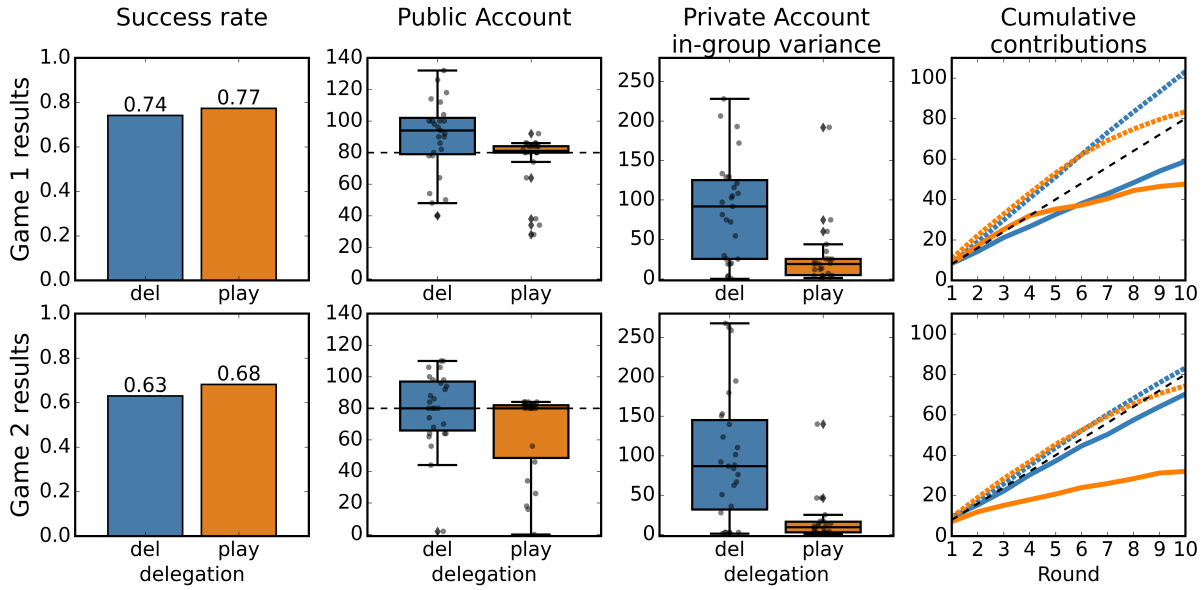

Figure 1: Experimental results pertaining to 27 delegation groups (*del*) and 22 no-delegation groups (*play*), that played the two games that compose the full experiment, denoted as Game 1 (top row panels) and Game 2 (bottom row panels). The first column shows the average success rate obtained by the groups in each treatment condition in both Game 1 (top) and Game 2 (bottom). The second column shows the distribution of Public Account values accumulated by each group at the end of the game for each treatment condition and for each game played (Game 1, top and Game 2, bottom). A black dashed line is shown at $y = 80$ denoting the threshold needed for the group to successfully avoid the collective risk. The third column presents the distribution of Private Account in-group variance, or in other words, a measure of how unequal the Private Accounts in each group were at the end of the game, for each different treatment condition and game played. Finally, the last column illustrates the round-by-round group play, by showing how the Public Account value accumulated in the course of each game, averaged over the different treatments (same colors for *del* and *play* as in previous panels) and success obtained (successful groups are marked with a dashed line, and correspond to the two upper lines, while unsuccessful groups correspond to the two bottom lines in each figure) for each game played (again, Game 1 is presented on the top panel while Game 2 is on the bottom panel). A black dashed-line is added to portray the trajectory of a group that would in every round contribute 8, guaranteeing an efficient group outcome (and fair if the efforts were equally distributed among group members) where the group would end up contributing exactly 80.