Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Centralized versus Personalized Commitments and Their Influence on Cooperation in Group Interactions

The Anh Han,[1] **Luís Moniz Pereira,**[2] **Luis A. Martinez-Vaquero,**[3] **Tom Lenaerts**[4,5]
[1] School of Computing and Digital Futures Institute, Teesside University, UK
[2] NOVALINCS, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal
[3] Institute of Cognitive Sciences and Technologies, National Research Council of Italy (ISTC-CNR), Italy
[4] MLG, Département d'Informatique, Université Libre de Bruxelles, Belgium
[5] AI lab, Computer Science Department, Vrije Universiteit Brussel, Belgium
Corresponding: T.Han@tees.ac.uk

Abstract

Before engaging in a group venture agents may seek commitments from other members in the group and, based on the level of participation (i.e. the number of actually committed participants), decide whether it is worth joining the venture. Alternatively, agents can delegate this costly process to a (beneficent or non-costly) third-party, who helps seek commitments from the agents. Using methods from Evolutionary Game Theory, this paper shows that, in the context of Public Goods Game, much higher levels of cooperation can be achieved through such centralized commitment management. It provides a more efficient mechanism for dealing with commitment free-riders, those who are not willing to bear the cost of arranging commitments whilst enjoying the benefits provided by the paying commitment proposers. We show also that the participation level plays a crucial role in the decision of whether an agreement should be formed; namely, it needs to be more strict in the centralized system for the agreement to be formed; however, once it is done right, it is much more beneficial in terms of the level of cooperation as well as the attainable social welfare. In short, our analysis provides important insights for the design of multi-agent systems that rely on commitments to monitor agents' cooperative behavior.

Introduction

Before embarking on a group venture agents may seek commitments from others in the group and estimate how interested they are in contributing to the group effort, as that allows them to judge whether it is worthwhile to start the initiative or whether it is beneficial to join. Arranging such prior commitments, for instance in the form of enforceable contracts or pledges (Nesse 2001), deposit-refund schemes (Cherry and McEvoy 2013; Sasaki et al. 2015), or even emotional or reputation-based commitment devices (Frank 1988; Nesse 2001), enforces others to cooperate, as it requires them to reveal their preferences or intentions (Sterelny 2012; Han, Pereira, and Santos 2012a). Commitments have been widely studied in multi-agent and autonomous agent systems, in order to ensure high levels of cooperation among agents (Wooldridge and Jennings 1999; Castelfranchi and Falcone 2010; Winikoff 2007). They have also been utilized for ensuring good behaviors in various computerised applications such as electric vehicle charging (Stein et al. 2012) and peer-to-peer sharing networks (Rzadca et al. 2015).

Using methods from Evolutionary Game Theory (EGT) (Hofbauer and Sigmund 1998), this paper compares two different approaches to arranging commitment: personalized vs. centralized commitments. Existing game theoretical models of prior commitments have focused on the first approach, in which the agents themselves, prior to an interaction, pay the cost to arrange a reliable commitment with their peers (Han, Pereira, and Santos 2012b; Han et al. 2013a; Hasan and Raja 2013; Han et al. 2015a; Han 2016; Han, Pereira, and Lenaerts 2016). Yet, instead of having (some) agents taking the initiative of arranging the commitments, one could assume that a central authority or institution is responsible for facilitating this costly process. The presence of this institution may be to improve the level of cooperation in the population or the social welfare (e.g. public transportation arranged by government, international agreements supported by the UN, crowdsourcing systems) (Ostrom 1990; Nesse 2001; Cherry and McEvoy 2013). The institution may actually profit directly from this joint activity by requesting a fee from all committed players in order to provide the service. It is at this point unclear which behaviors benefit (or not) or how cooperation-levels change when moving from a personal to a centralized commitment model. Yet the answer to this question could be fundamental in the design of highly distributed multi-agent systems.

Our analysis is carried out in the context of the Public Goods Game (PGG), which allows us to directly compare evolutionary outcomes of our centralized approach with the personalized one described in (Han, Pereira, and Lenaerts 2016) for the PGG. The PGG can be described as follows: all players can decide whether or not to contribute an amount c to the public good (Sigmund 2010; Hauert et al. 2007) where their accumulated contribution is multiplied by a constant factor $r > 1$ before being equally distributed among all players. With r smaller than the group size (denoted by N), non-contributing free-riders gain more than contributors. Evolutionary Game Theory models (Hofbauer and Sigmund 1998;

Sigmund 2010) predict the demise of cooperation – famously known as 'the tragedy of commons' (Hardin 1968). In the personalized commitment version of the PGG, agents have, before playing the PGG, the option to propose to other members in the group to commit to contribute, paying a personal cost, ϵ, to make it credible (Han, Pereira, and Lenaerts 2016). If a sufficient number of members commit (i.e. greater than a so-called *participation level* F), the said proposers will contribute to the PGG. Otherwise, they refuse to do so. (For details see Model and Methods Section.)

Earlier work has shown that in the personalized approach, although it can lead to high levels of cooperation, full cooperation (i.e. all agents in the population always cooperate) is hardly reached, for both pair-wise and group interactions (Han, Pereira, and Santos 2012b; Han et al. 2013a; Han, Pereira, and Lenaerts 2016). The main reason is that commitment proposers are always dominated by different kinds of commitment free-riders, who do not have to pay the cost of arranging commitments while enjoying all the benefits provided by a commitment system.

This issue may be resolved if the commitment arrangement process is handled by a central authority that plays the role of proposers and enforces that all accepting agents share the setup cost. In that case it is not the (system) agents that have to carry the burden (of taking the initiative), simply removing the free-riding behavior mentioned earlier. The central authority will, based on the participation level (F), decide whether an agreement is formed and can go forward. Based on the outcome of this decision, the players in the PGG group will then decide how to play: cooperate (C) or defect (D). We will show, by closely monitoring this participation level, in so far it depends on the benefit-to-cost ratio of the PGG and the commitment parameters, that the centralized commitment approach can lead to significant improvement for cooperation.

Related Work

The problem of explaining the evolution of cooperation has been actively addressed in different fields of research, including Biology, Economics, Artificial Intelligence (AI) and Multi-agent Systems (MAS) (Nowak 2006; Sigmund 2010; Hofmann, Chakraborty, and Sycara 2011; Ranjbar-Sahraei et al. 2014; Airiau, Sen, and Villatoro 2014; Han, Pereira, and Lenaerts 2016). Among other mechanisms, such as reciprocity and costly punishment (see surveys in (Nowak 2006; Sigmund 2010)), prior- or pre- commitments have been shown to provide an important pathway for the evolution of cooperation, even in one-shot interactions, both by means of theoretical analysis (Han, Pereira, and Santos 2012b; Han et al. 2013b; Hasan and Raja 2013; Sasaki et al. 2015; Martinez-Vaquero et al. 2015; Han 2016; Han and Lenaerts 2016) and of behavioral experiments (Ostrom 1990; Cherry and McEvoy 2013). However, this literature, especially the theoretical one, has focused on the analysis of personalized commitments only. To the best of our knowledge, this paper provides the first dynamic or EGT analysis of the centralized commitment approach; interestingly, we show that this approach can lead to even higher levels of cooperation than what is achieved with the personalized one. Moreover, since prior works have shown that personalized commitments can better promote cooperation than costly punishment—a major pathway to evolution of cooperation for a wide range of game configurations, even in one-shot interactions (Han et al. 2013a; Han 2016; Han and Lenaerts 2016)—we now expect centralized commitments to even better promote cooperation compared to punishment.

Closely related to our model is the study of evolutionary mechanism design, analyzing how to efficiently interfere in an evolving dynamical system so as to influence it to achieve some desired behavior (Phelps, McBurney, and Parsons 2010; Han et al. 2015b). The interference is typically carried out (by an external decision maker) through observing system agents' behavior, thus rewarding the good and/or punishing the bad; with the goal being to minimize the interference cost while assuring a high level of good behavior. Our approach also involves an external decision maker, but differently from these works, we do not require a budget needed to reward or punish (indeed, the decision maker can even benefit from its role). In this sense, our model is related to the work on mechanism design without money, see e.g. (Procaccia and Tennenholtz 2009; Serafino and Ventre 2016), providing a potential approach to interfering in an evolving system without a budget.

Last but not least, it is noteworthy that commitments have been studied extensively in AI and MAS literature (Wooldridge and Jennings 1999; Castelfranchi and Falcone 2010; Winikoff 2007; Chopra and Singh 2009; Stein et al. 2012; Rzadca et al. 2015). Differently from our work, these studies utilize commitments for the purpose of regulating individual and collective behaviors, formalizing different aspects of commitments (such as norms and conventions) in a MAS. However, our results and approach provide novel insights into the design of such computerized and MAS systems as these require commitments to ensure high levels of cooperation or efficient collaboration within a group or team of agents; for instance, instead of letting the agents seek commitments from others by themselves, one should arrange for a centralized party to handle that, closely monitoring an appropriate participation from team members.

Model and Methods

Personalized commitment model in PGG

The modelling approach of personalized commitment utilized here follows the work in (Han, Pereira, and Lenaerts 2016). There, agents have, before playing the one-shot (i.e. non-repeated) PGG, the option to propose other members in the group to commit to contribute. To do so, the commitment proposers (COM) must pay a *personal* cost, ϵ. If a sufficient number of the members commit, i.e. greater than a given participation level F ($1 \leq F \leq N$), the agreement is formed and then the PGG is played. Otherwise, the agreement is not formed and the commitment proposers do not contribute to the PGG. Those who committed but then do not contribute have to compensate the contributing others at a personal cost, δ. There are N possible participation levels, encoded in terms of strategies, COM_F where $F \in \{1, ..., N\}$. COM_F players are those that only contribute if there are at least

F players in the whole group that agree to contribute (including the COM_F players themselves); otherwise, COM_F players do not contribute to the PGG.

Similarly to (Han, Pereira, and Lenaerts 2016), besides COM_F strategies ($1 \leq F \leq N$), agents can adopt the following strategies: i) traditional unconditional contributors (C, who always commit when being proposed a commitment deal, contribute whenever the PGG is played, but do not propose commitment); ii) unconditional non-contributors (D, who do not accept commitment, defect when the PGG is played, and do not propose commitment); iii) fake committers (FAKE, who accept a commitment proposal yet do not subsequently contribute whenever the PGG is actually played) and iv) commitment free-riders (FREE, who defect unless being proposed a commitment, which they then accept and cooperate subsequently in the PGG). Detailed payoff calculation for interaction among these strategies can be found in (Han, Pereira, and Lenaerts 2016).

It has been shown that in this approach of personalizing commitment, although it can lead to high levels of cooperation, full cooperation is hardly reached (Han, Pereira, and Santos 2012b; Han et al. 2013a; Han 2016; Han, Pereira, and Lenaerts 2016). The main reason is that commitment proposers (COM_F) are always dominated by pure cooperators (C), who do not have to pay the cost of arranging commitments while being able to maintain perfect cooperation among themselves. That dominance leads to the unavoidable presence of defectors even when these defectors can be effectively dealt with by the commitment proposers, since cooperators themselves are strongly dominated by defective strategies in the absence commitments (see already Fig. 1). This issue can be avoided if the burden of proposing and arranging commitments is taken away just from the proposers by having a centralized party to handle that. Then, all those who accept to join a central proposal to commit have the duty to share its cost. There is no longer a distinction between who is a proposer or not, as we shall see next.

Centralized commitment model in PGG

Agents receive a request to commit to contribute to a one-shot PGG by a centralized authority, who wants to ensure high levels of cooperation in the population of agents (see again some examples in Introduction). They can then decide whether or not to accept to join the commitment. If there are enough participants in the commitment, i.e. the number of committed players are at least F ($1 \leq F \leq N$), the central authority decides then that the agreement is formed. Otherwise, the agents just play the regular one-shot PGG. In the former case, the committed players (agree to) equally share a cost ϵ [1] but then those who defect payment if the PGG goes through have to pay a fine δ [2].

A strategy is defined by three decisions: i) she accepts (A) or not (N) the agreement; ii) cooperates (C) or defects (D) if the agreement is formed; iii) cooperates (C) or defects (D) if it is not formed. Note that playing A assumes also implicitly that one is prepared to pay a share of the cost of setting up the agreement. There are eight such strategies in total. For a clear comparison with the personalized model, we will focus here on five, i.e. two accepting strategies, ACD and ADD, and three strategies that will not accept the authority's proposal to commit, i.e. NCD, NDD and NDC. The three remaining strategies ACC, ADC and NCC can be omitted as they are each dominated by at least one of the former strategies—because in the absence of a commitment, playing D is the dominant option (i.e., ACC, ADC and NCC are dominated by ACD, ADD and NCD, respectively). Comparing these strategies to those discussed in the personalized model, one can say that COM and ACD, FAKE and ADD and D and NDD are clearly equivalent strategies. NCD and FREE are slightly different in the sense that FREE still accepts a proposal to commit. Since in this centralized commitment model it is assumed that accepting also implies paying a part of the cost, the NCD is closest to the FREE strategy as it is also willing to cooperate but not prepared to pay the cost. The biggest difference with the personalized model is that the ACC strategy (which corresponds to the C strategy in the personalized model) is no longer viable here. It is replaced by the closest non-dominated NDC strategy.

We denote $\Pi_{A,B}(k)$ and $\Pi_{B,A}(k)$ as the payoffs of strategy A and B, respectively, in a group with k A-strategists and $(N-k)$ B-strategists. We have

- $\Pi_{ACD,ADD}(k) = \frac{rk}{N}c - c - \frac{\epsilon}{N}\ \forall 1 \leq k \leq N$;
  $\Pi_{ADD,ACD}(k) = \frac{rk}{N}c - \frac{\epsilon}{N} - \delta\ \forall 0 \leq k \leq N-1$;
- $\Pi_{ACD,NCD}(k) = rc - c - \frac{\epsilon}{k}$ if $k \geq F$ and 0 otherwise;
  $\Pi_{NCD,ACD}(k) = rc - c$ if $k \geq F$ and 0 otherwise;
- $\Pi_{ACD,NDD}(k) = \frac{rk}{N}c - c - \frac{\epsilon}{k}$ if $k \geq F$ and 0 otherwise;
  $\Pi_{NDD,ACD}(k) = \frac{rk}{N}c$ if $k \geq F$ and 0 otherwise;
- $\Pi_{ACD,NDC}(k) = \frac{rk}{N}c - c - \frac{\epsilon}{k}$ if $k \geq F$ and $\frac{r(N-k)}{N}c$ otherwise;
  $\Pi_{NDC,ACD}(k) = \frac{rk}{N}c$ if $k \geq F$ and $\frac{r(N-k)}{N}c - c$ otherwise;
- $\Pi_{ADD,NCD}(k) = \frac{r(N-k)}{N}c - \frac{\epsilon}{k} - \delta$ if $k \geq F$ and 0 otherwise;
  $\Pi_{NCD,ADD}(k) = \frac{r(N-k)}{N}c - c$ if $k \geq F$ and 0 otherwise;
- $\Pi_{ADD,NDD}(k) = -\frac{\epsilon}{k} - \delta$ if $k \geq F$ and 0 otherwise; $\Pi_{NDD,ADD}(k) = 0$
- $\Pi_{ADD,NDC}(k) = -\frac{\epsilon}{k} - \delta$ if $k \geq F$ and $\frac{r(N-k)}{N}c$ otherwise;
  $\Pi_{NDC,ADD}(k) = 0$ if $k \geq F$ and $\frac{r(N-k)}{N}c - c$ otherwise;
- $\Pi_{NCD,NDD}(k) = 0$; $\Pi_{ND,NC}(k) = 0$;
- $\Pi_{NCD,NDC}(k) = \frac{r(N-k)}{N}c$; $\Pi_{NDC,NC}(k) = \frac{r(N-k)}{N}c - c$;
- $\Pi_{NDD,NDC}(k) = \frac{r(N-k)}{N}c$; $\Pi_{NDC,ND}(k) = \frac{r(N-k)}{N}c - c$.

[1] For a clear comparison, it is assumed here that the costs of arranging or managing a commitment (ϵ) are the same in both models. For better or worse, this may vary depending on the efficiency of the authority in place to help monitor the commitment (more discussion in Future Work). However, our later analysis shows that the centralized model permits a much higher cost than the personalized one wherein the corresponding system leads to high levels of cooperation.

[2] It is worth noting that in the centralized model the amount δ is not transferred as a compensation to the committed players as in the personalized model. This is to simplify the deployment of the centralized commitment mechanism (as compensation might be more difficult through the third party than direct personal transaction), while, our results show that this clear disadvantage of the centralized model still allows it to outperform the personalized one.

Population Setup and Evolutionary Dynamics

Analytical and numerical results obtained here use EGT methods for finite populations (Imhof, Fudenberg, and Nowak 2005; Sigmund 2010), exactly as in our previous work (Han, Pereira, and Lenaerts 2016). Below, we briefly recall its details, whose full description can be found in Methods section of (Han, Pereira, and Lenaerts 2016). Here evolutionary dynamics is shaped by social learning (Hofbauer and Sigmund 1998; Sigmund 2010), whereby the more successful agents tend to be imitated more often by others. Moreover, the pairwise comparison rule is used to model social learning, where an agent A with fitness f_A adopts the strategy of another agent B with fitness f_B with probability (Traulsen, Nowak, and Pacheco 2006): $\left(1+e^{-\beta(f_B-f_A)}\right)^{-1}$. The parameter β represents the 'intensity of selection', i.e., how strongly the agents base their decision to imitate on the fitness difference $(f_B - f_A)$. For $\beta = 0$, we obtain the limit of neutral drift, while for large β, imitation becomes increasingly deterministic.

Next, with some mutation probability, an agent switches randomly to a different strategy without imitating another agent. For a small mutation limit, the behavioral dynamics can be described by a Markov Chain, where each state corresponds to a monomorphic population (i.e. consisting of agents all with the same strategy), whereas the state transition probabilities are given by the fixation probability of a single mutant, see calculation below. The resulting Markov Chain has a stationary distribution that characterizes the average time the population spends in each of these monomorphic end states (e.g. see Fig. 1).

In finite populations, the groups engaging in PGG are given by multivariate hypergeometric sampling. We denote

$$H(k,N,m,Z) = \binom{m}{k}\binom{Z-m}{N-k}/\binom{Z}{N}.$$

Hence, in a population of x i-strategists and $(Z-x)$ j-strategists, the average payoffs to $i-$ and $j-$ strategists are:

$$P_{ij}(x) = \sum_{k=0}^{N-1} H(k,N-1,x-1,Z-1)\,\Pi_{ij}(k+1),$$
$$P_{ji}(x) = \sum_{k=0}^{N-1} H(k,N-1,x,Z-1)\,\Pi_{ji}(k).$$

The probability to change the number k of agents using strategy i by ± 1 in each time step can be written as

$$T^{\pm}(k) = \frac{Z-k}{Z}\frac{k}{Z}\left[1+e^{\mp\beta[P_{ij}(k)-P_{ji}(k)]}\right]^{-1}. \quad (1)$$

The fixation probability of a single mutant with a strategy i in a population of $(N-1)$ agents using j is given by

$$\rho_{j,i} = \left(1+\sum_{i=1}^{Z-1}\prod_{j=1}^{i}\frac{T^-(j)}{T^+(j)}\right)^{-1}. \quad (2)$$

In the limit of neutral selection (i.e. $\beta = 0$), $\rho_{B,A}$ equals the inverse of population size, $\rho_N = 1/Z$.

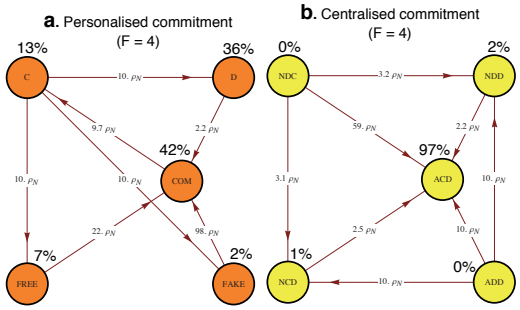

Figure 1: ***Stationary distribution and fixation probabilities for the (a) personalized commitment and (b) centralized commitment models***. For clarity only the transitions that are stronger than neutral are shown. In the former case (panel a), note the cyclic pattern from C to defective strategies to COM and back. COM is most abundant in the population but defection is still present at high frequency. In the latter case (panel b), ACD is risk-dominant against all other four strategies, thereby reaching almost full cooperation in the population (97%). Parameters: $N = 5$, $Z = 100$, $r = 3$, $\delta = 4$, $\epsilon = 0.5$, $\beta = 0.25$, $F = 4$; $\rho_N = 1/Z$ denotes the neutral fixation probability.

Risk-dominance condition An important analytical criteria to determine the viability of a given strategy is whether it is risk-dominant with respect to other strategies (Nowak 2006; Gokhale and Traulsen 2010). Namely, one considers which selection direction is more probable: an i mutant fixating in a homogeneous population of agents playing j or a j mutant fixating in a homogeneous population of agents playing i. When the first is more likely than the latter, i is said to be *risk-dominant* against j, which holds for any β and in the limit of large Z when

$$\sum_{k=1}^{N}\Pi_{ij}(k) \geq \sum_{k=0}^{N-1}\Pi_{ji}(k). \quad (3)$$

Results

We start by providing analytical conditions for when ACD can be a viable strategy, by being risk-dominant when playing against other strategies. This analysis shows when the centralized proposer has sufficient authority to induce cooperative behavior. We then provide numerical simulation results to support the analytical observations and beyond.

When will ACD strategists dominate the others?

To begin with, using Eq. 3, ACD is risk-dominant against ADD if

$$\sum_{k=1}^{N}\left(\frac{rk}{N}c - c - \frac{\epsilon}{N}\right) \geq \sum_{k=0}^{N-1}\left(\frac{rk}{N}c - \frac{\epsilon}{N} - \delta\right), \quad (4)$$

which is simplified to

$$\delta \geq \frac{(N-r)c}{N}. \quad (5)$$

Now, ACD is risk-dominant against NCD if

$$\sum_{k=F}^{N}\left(rc - c - \frac{\epsilon}{k}\right) \geq \sum_{k=F}^{N-1}(rc-c), \quad (6)$$

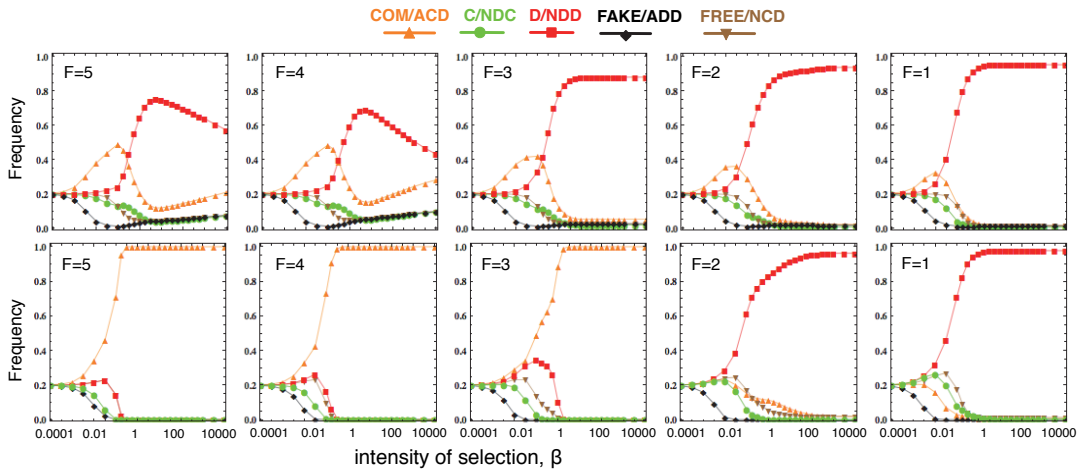

Figure 2: **Frequency of each strategy as a function of β, for different F, in a population of five strategies for: (top row) personalized commitment (bottom row) centralized commitment models**. For the personalized commitment model (top row), commitment proposers are most abundant when the intensity of selection β is small, even for small F, but when β is large, pure defectors dominate the population, especially when F is small. However, even for the best scenarios, the level of defection is still high (greater than 35%). To the contrary, in the centralized commitment model (bottom row), when F is sufficiently large (≥ 3), ACD is most abundant, reaching 100% when β is sufficiently large. However, when F is small ($F = 1$ or 2), defection is most abundant even for small β. Parameters: $N = 5$, $Z = 100$, $r = 3$, $\delta = 4$, $\epsilon = 0.5$.

which is simplified to

$$(r-1)c \geq \epsilon \sum_{k=F}^{N} \frac{1}{k}. \tag{7}$$

Furthermore, ACD is risk-dominant against NDD if

$$\sum_{k=F}^{N} \left(\frac{rk}{N} c - c - \frac{\epsilon}{k} \right) \geq \sum_{k=F}^{N-1} \left(\frac{rkc}{N} \right), \tag{8}$$

which is simplified to

$$(r+F-N-1)c \geq \epsilon \sum_{k=F}^{N} \frac{1}{k}. \tag{9}$$

Since $F \leq N$, this last condition entails the condition in Eq. (7). Finally, ACD is risk-dominant against NDC if

$$\sum_{k=F}^{N} \left(\frac{rk}{N} c - c - \frac{\epsilon}{k} \right) + \sum_{k=1}^{F-1} \left(\frac{r(N-k)}{N} c \right) \geq \sum_{k=F}^{N-1} \left(\frac{rkc}{N} \right) + \sum_{k=0}^{F-1} \left(\frac{r(N-k)}{N} c - c \right), \tag{10}$$

which is simplified to

$$(2F-N-1)c \geq \epsilon \sum_{k=F}^{N} \frac{1}{k}. \tag{11}$$

In summary, for ACD to be risk-dominant against all the other four strategies, the following inequalities need to be satisfied

$$\delta \geq \frac{(N-r)c}{N}, \tag{12}$$

$$\begin{aligned} \epsilon &\leq \frac{\min\{2F-N-1, r+F-N-1\}}{H_F} \times c \\ &= \frac{\min\{F, r\}+F-N-1}{H_F} \times c, \end{aligned} \tag{13}$$

where $H_F = \sum_{k=F}^{N} \frac{1}{k}$. The second condition implies that for the risk-dominance of ACD, it is necessary that more than half of the potential participants accept the agreement ($F \geq \frac{N+1}{2}$). Moreover, as $r > 1$ the right hand side of this inequality is always positive when $F = N$ (hence it can be satisfied with sufficiently small ϵ). On the other hand, for the first condition (Eq. 12) to hold it is sufficient that the compensation when the agreement fails is bigger than the cost of cooperation ($\delta \geq c$), as its right hand side is always smaller than c.

Thus, for a suitable arrangement of the commitment, i.e. sufficiently large F, small ϵ and large δ, these results show that ACD is risk-dominant against all other strategies. Defectors who commit to contribute (i.e. ADD) have to pay a fine when faking. Defectors who do not commit (i.e. NDD and NDC) are less likely than ACD to belong to a group with sufficient participation level so that an agreement is formed.

A centralized authority avoids cycles of cooperation and defection

As can be observed in Fig.1, the results contrast with the personalized commitment system, wherein the commitment proposing strategy (COM) is always risk-dominated by C, leading to cycles between cooperative and defective strategies and thereby an unavoidable presence of defective strategies. The Markov diagrams in the figure show that (for illustration, we use $F = 4$) the personalized commitment model (panel a), leads to a cyclic pattern from C to defective strategies (i.e. D, FREE and FAKE) to COM and back. COM is the most abundant in the population but defection is still

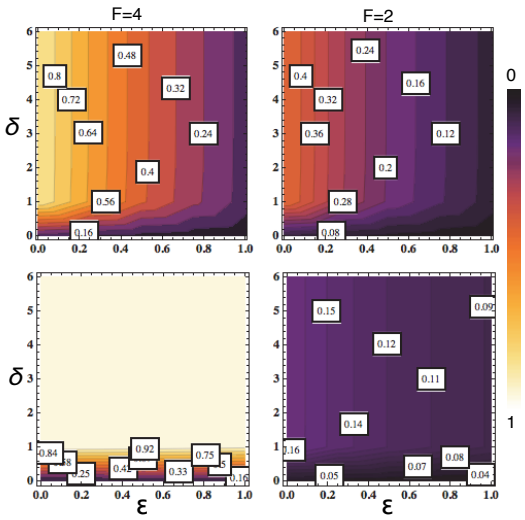

Figure 3: Frequency of COM (top row) and ACD (bottom row) as a function of ϵ and δ, in a population of five strategies for the personalized and centralized commitment models, respectively ($F = 2$ for left and $F = 4$ for right columns). Parameters: $N = 5$, $Z = 100$, $r = 3$, $\beta = 0.25$.

present at high frequency (the total frequency of D, FREE and FAKE is 45%). To the contrary, in the centralized commitment model (panel b), there are strong transitions from all other four strategies towards ACD, leading to almost full cooperation in the population (97%). It is noteworthy that cooperation in the personalized commitment model does not improve when moving from $F = 4$ to $F = 5$ (indeed, $F = 4$ is its best option). Yet a similar increase for the centralized model leads to 100% cooperation in the population.

Centralized commitment induces cooperation for diverse selection strengths, as long as participation is high This risk-dominance property in the centralized commitment model leads to even greater advantages as compared to the personalized model when the payoff advantage obtained from the game is important, i.e. when the intensity of selection β increases (see the Eq. 1). In particular, Fig. 2 compares the two models for different F and varying intensity of selection β. For the personalized commitment model (top row), commitment proposers COM are most abundant when β is small, even for small F, but when β is large, pure defectors (D) dominate the population, especially when F is small. However, even for the best scenarios for COM (namely, $F = 4$ and intermediate β), the level of defection is still high (greater than 35%). To the contrary, in the centralized commitment model (bottom row), when F is sufficiently large (≥ 3), ACD is most abundant, reaching 100% when β is sufficiently large. However, when F is small ($F \leq 2$), defection is most abundant even for small β. This is in accordance with the analytical conditions, which indicate that it must hold that $F \geq \frac{N+1}{2} = 3$ for ACD to be an evolutionarily viable strategy (risk-dominant against others). This implies that in the latter model, it needs to be more strict in terms of the level of participation required from other players for the agreement to be formed; however, once it is done right, it is much more beneficial in terms of the level of cooperation and social welfare.

This interesting observation regarding COM and ACD is robust for varying the parameters of a commitment system, ϵ and δ, see Fig. 3 (for illustration we show for $F = 2$ and 4). We compare the frequency COM and ACD in the two commitment models, as they are ones that mainly generate cooperation or social welfare in both models. For the personalized commitment model (top row), COM is abundant when ϵ is small enough and when δ is sufficiently high. For $F \geq 3$, the frequency of COM can be larger than 50% and can reach approx. 80% ($F = 4$), but full cooperation is never reached. Differently, in the centralized commitment model (bottom row), when F is sufficiently large (≥ 3), whenever δ is sufficiently large ($\delta \geq c = 1$), ACD is most abundant, reaching even 100% for $F = 4$ or 5. Moreover, ACD dominates the population (i.e. $\geq 50\%$ frequency) for a much larger range of ϵ than COM. However, when F is small ($F = 1$ or 2), ACD performs worse than COM, having very low frequency even for small ϵ. Again, the results show that commitment needs to be more strict in terms of the participation level required from other players for the agreement to be formed in the centralized model, but once it is assured, it guarantees much higher levels of cooperation.

Conclusions and Future Work

The present paper describes a novel, centralized approach to arranging prior commitments in group interactions, showing that it outperforms the dominant model of personalized commitments in the literature. By having a (beneficent or non-costly) central authority or institution to help arrange commitments from the group members instead of leaving them to the initiative, it removes the commitment free-riding issue that prevented the personalized approach to reach full cooperation (Han et al. 2013a; Han, Pereira, and Lenaerts 2016). Unlike the personalized approach, a commitment-accepting strategy (ACD) can be risk-dominant against all other strategies in the population, leading to significantly higher levels of cooperation for a wide range of parameters (note that a population of ACD players maintain 100% cooperation). We showed, both analytically and by numerical simulation, that by requiring a sufficiently high level of participation from group members (more than half of the group size), significant levels of cooperation can be ensured.

As a future direction, we consider that a central institution may come to incorporate other duties and roles that make it more complex and onerous; for instance, those of marketing the PGG until enough players are found, or of detecting and preventing negative pool associations against the PGG, or those of the police, courts, lawyers, prisons, and enacting new laws. In these scenarios, other administration costs need to be considered. Moreover, as an enforcement institution can benefit from its role, there may be competition for it, leading to possibility of natural selection among institutions. As such, multi-level evolutionary dynamics will need to be analysed, which we plan to explore in future work.

In short, our analysis suggests novel insights for the design of MAS and computerized systems in order to ensure high levels of cooperation among agents in the systems. Namely, that can be achieved by arranging for a central party

to help agents arrange commitments instead of leaving them to the task, and requiring that an agreement is formed and enforced only when the agents' participation level is sufficiently high.

References

Airiau, S.; Sen, S.; and Villatoro, D. 2014. Emergence of conventions through social learning. *Autonomous Agents and Multi-Agent Systems* 28(5):779–804.

Castelfranchi, C., and Falcone, R. 2010. *Trust Theory: A Socio-Cognitive and Computational Model (Wiley Series in Agent Technology)*. Wiley.

Cherry, T. L., and McEvoy, D. M. 2013. Enforcing compliance with environmental agreements in the absence of strong institutions: An experimental analysis. *Environmental and Resource Economics* 54(1):63–77.

Chopra, A. K., and Singh, M. P. 2009. Multiagent commitment alignment. In *AAMAS'2009*, 937–944.

Frank, R. H. 1988. *Passions Within Reason: The Strategic Role of the Emotions*. Norton and Company.

Gokhale, C. S., and Traulsen, A. 2010. Evolutionary games in the multiverse. *Proc. Natl. Acad. Sci. U.S.A.* 107(12):5500–5504.

Han, T. A., and Lenaerts, T. 2016. A synergy of costly punishment and commitment in cooperation dilemmas. *Adaptive Behavior* 24(4):237–248.

Han, T. A.; Pereira, L. M.; Santos, F. C.; and Lenaerts, T. 2013a. Good agreements make good friends. *Scientific reports* 3(2695).

Han, T. A.; Pereira, L. M.; Santos, F. C.; and Lenaerts, T. 2013b. Why Is It So Hard to Say Sorry: The Evolution of Apology with Commitments in the Iterated Prisoner's Dilemma. In *IJCAI'2013*, 177–183. AAAI Press.

Han, T. A.; Santos, F. C.; Lenaerts, T.; and Pereira, L. M. 2015a. Synergy between intention recognition and commitments in cooperation dilemmas. *Scientific reports* 5(9312).

Han, T. A.; Tran-Thanh, L.; Jennings, N. R.; et al. 2015b. The cost of interference in evolving multiagent systems. In *AAMAS'2015*, 1719–1720.

Han, T. A.; Pereira, L. M.; and Lenaerts, T. 2016. Evolution of commitment and level of participation in public goods games. *Autonomous Agents and Multi-Agent Systems* 1–23.

Han, T. A.; Pereira, L. M.; and Santos, F. C. 2012a. Corpus-based intention recognition in cooperation dilemmas. *Artificial Life* 18(4):365–383.

Han, T. A.; Pereira, L. M.; and Santos, F. C. 2012b. The emergence of commitments and cooperation. In *AAMAS'2012*, 559–566.

Han, T. A. 2016. Emergence of social punishment and cooperation through prior commitments. In *AAAI'2016*, 2494–2500.

Hardin, G. 1968. The tragedy of the commons. *Science* 162:1243–1248.

Hasan, M. R., and Raja, A. 2013. Emergence of cooperation using commitments and complex network dynamics. In *IEEE/WIC/ACM Intl Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, 345–352.

Hauert, C.; Traulsen, A.; Brandt, H.; Nowak, M. A.; and Sigmund, K. 2007. Via freedom to coercion: The emergence of costly punishment. *Science* 316:1905–1907.

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

Hofmann, L.-M.; Chakraborty, N.; and Sycara, K. 2011. The evolution of cooperation in self-interested agent societies: a critical study. AAMAS '11, 685–692.

Imhof, L. A.; Fudenberg, D.; and Nowak, M. A. 2005. Evolutionary cycles of cooperation and defection. *Proc. Natl. Acad. Sci. U.S.A.* 102:10797–10800.

Martinez-Vaquero, L. A.; Han, T. A.; Pereira, L. M.; and Lenaerts, T. 2015. Apology and forgiveness evolve to resolve failures in cooperative agreements. *Scientific reports* 5(10639).

Nesse, R. M. 2001. *Evolution and the capacity for commitment*. Foundation series on trust. Russell Sage.

Nowak, M. A. 2006. Five rules for the evolution of cooperation. *Science* 314(5805):1560.

Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.

Phelps, S.; McBurney, P.; and Parsons, S. 2010. Evolutionary mechanism design: a review. *Autonomous Agents and Multi-Agent Systems* 21(2):237–264.

Procaccia, A. D., and Tennenholtz, M. 2009. Approximate mechanism design without money. In *Proc. 10th conference on Electronic commerce*, 177–186.

Ranjbar-Sahraei, B.; Bou Ammar, H.; Bloembergen, D.; Tuyls, K.; and Weiss, G. 2014. Evolution of cooperation in arbitrary complex networks. In *AAMAS'2014*, 677–684.

Rzadca, K.; Datta, A.; Kreitz, G.; and Buchegger, S. 2015. Game-theoretic mechanisms to increase data availability in decentralized storage systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 10(3):14.

Sasaki, T.; Okada, I.; Uchida, S.; and Chen, X. 2015. Commitment to cooperation and peer punishment: Its evolution. *Games* 6(4):574–587.

Serafino, P., and Ventre, C. 2016. Heterogeneous facility location without money. *Theoretical Computer Science* 636:27–46.

Sigmund, K. 2010. *The Calculus of Selfishness*. Princeton University Press.

Stein, S.; Gerding, E.; Robu, V.; and Jennings, N. R. 2012. A model-based online mechanism with pre-commitment and its application to electric vehicle charging. In *AAMAS'2012*, 669–676.

Sterelny, K. 2012. *The evolved apprentice*. MIT Press.

Traulsen, A.; Nowak, M. A.; and Pacheco, J. M. 2006. Stochastic dynamics of invasion and fixation. *Phys. Rev. E* 74:11909.

Winikoff, M. 2007. Implementing commitment-based interactions. AAMAS '07, 868–875.

Wooldridge, M., and Jennings, N. R. 1999. The cooperative problem-solving process. In *Journal of Logic and Computation*, 403–417.