

# Why Is It So Hard to Say Sorry? Evolution of Apology with Commitments in the Iterated Prisoner's Dilemma\*

The Anh Han<sup>1,2</sup> and Luís Moniz Pereira<sup>3</sup> and Francisco C. Santos<sup>4,5</sup> and Tom Lenaerts<sup>1,2</sup>

<sup>1</sup> AI lab, Vrije Universiteit Brussel, <sup>2</sup> MLG group, Université Libre de Bruxelles, Belgium

<sup>3</sup> Centro de Inteligência Artificial (CENTRIA), Universidade Nova de Lisboa, Portugal

<sup>4</sup> GAIPS/INESC-ID & Instituto Superior Técnico, Portugal,

<sup>5</sup> ATP-group, Inst. Investigação Interdisciplinar, Lisbon, Portugal

## Abstract

When making a mistake, individuals can apologize to secure further cooperation, even if the apology is costly. Similarly, individuals arrange commitments to guarantee that an action such as a cooperative one is in the others' best interest, and thus will be carried out to avoid eventual penalties for commitment failure. Hence, both apology and commitment should go side by side in behavioral evolution. Here we provide a computational model showing that apologizing acts are rare in non-committed interactions, especially whenever cooperation is very costly, and that arranging prior commitments can considerably increase the frequency of such behavior. In addition, we show that in both cases, with or without commitments, apology works only if it is sincere, i.e. costly enough. Most interestingly, our model predicts that individuals tend to use much costlier apology in committed relationships than otherwise, because it helps better identify free-riders such as fake committers: *'commitments bring about sincerity'*. Furthermore, we show that this strategy of apology supported by commitments outperforms the famous existent strategies of the iterated Prisoner's Dilemma.

## 1 Introduction

Apology is perhaps the most powerful and ubiquitous mechanism for conflict resolution, with an abundance of experimental evidence from Economics, Psychology and Criminal Justice analysis [Ohtsubo and Watanabe, 2009; Takaku *et al.*, 2001; Petrucci, 2002; Ho, 2012; Atran *et al.*, 2007]. An apology can resolve a conflict without having to involve external parties (e.g. teachers, parents, courts), which may cost all sides of the conflict significantly more. Evidence shows that there is a much higher chance that customers stay with a company that apologizes for mistakes [Abeler *et al.*, 2010]. Apology leads to fewer lawsuits with lower settlements in medical error situations [Liang, 2002]. Apology even enters the law as an effective mechanism of resolving conflicts [Petrucci, 2002], and has been implemented in several

computerized systems such as human-computer interaction (HCI) and online markets so as to facilitate users' positive emotions and cooperation [Tzeng, 2004; Park *et al.*, 2012; Vasalou *et al.*, 2008; Utz *et al.*, 2009]. As such, one can hypothesize that apology is embedded in our behavior and is evolutionarily stable in social situations dominated by conflicts and misconceptions, which is what we will show here.

Apologies typically occur in situations where interactions are repeated [Boerlijst *et al.*, 1997; Sigmund, 2010]. In the context of such Game Theoretical research, an apology is often implicitly modeled by means of one or several cooperative acts after a wrongful or misread defective action. Such a behavior can be summarized as 'I hit you once, so you are allowed to hit me back'. It is clearly an inefficient way of apologizing for mistakes, as it might not thoroughly resolve the conflict, as is the case for Tit-For-Tat-like (TFT) strategies [Axelrod, 1984; Nowak and Sigmund, 1993; Boerlijst *et al.*, 1997; Imhof *et al.*, 2007; Han *et al.*, 2011]. It would be more natural to explicitly apologize so as to resolve the conflict before the next interaction occurs, even at a cost, thereby securing a cooperative behavior from the co-player and avoiding an escalation of the conflict.

In the current work, we study, using methods provided in Evolutionary Game Theory (EGT) [Hofbauer and Sigmund, 1998; Sigmund, 2010], whether an explicit form of apology can be a viable strategy for the *evolution of cooperation* in the iterated Prisoner's Dilemma (IPD) [Trivers, 1971; Axelrod, 1984; Sigmund, 2010]. In each round of this game, two players simultaneously decide to either cooperate (C) or defect (D), where the game is designed so that the rational option is to always play D. However, since the game is repeated cooperation may be beneficial, especially when the probability of playing again with the same partner is high [Trivers, 1971]. Several strategies that perform well in this IPD game have been discovered, as for instance the famous TFT and Win-Stay-Lose-Shift (WSLS) strategies (see Section 3.2 for detailed definitions). Yet neither of these strategies considers explicitly the possibility of apologizing.

Here we provide a model containing strategies that explicitly apologize when making a mistake between rounds, where mistakes are modeled by a noise parameter  $\alpha$ . This explicit apology is represented by a cost  $\gamma > 0$ , paid by the apologizer to the other player. As is known, TFT is not able to return to cooperation readily, as a mistake simply leads to several

\*TAH and FCS are supported by FWO Belgium and FCT-Portugal, respectively; Corresponding author: h.anh@ai.vub.ac.be.

rounds of retaliation. Yet when a player apologizes for her mistake, she simply compensates the other player, ensuring that this other player will keep on playing C. In a population consisting of only apologizers perfect cooperation can easily be maintained. Yet other behaviors that exploit such apology behavior could emerge, destroying any benefit of the apology action. We show here that when the apology occurs in a system where the players first ask for a commitment before engaging in the interaction [Han *et al.*, 2012a; 2012b; Han, 2013], this exploitation can be avoided.

An apology mechanism without commitment, that is, having no consequence when opting for not apologizing, simply benefits fake apologizers. On the other hand, commitment without a conflict resolution mechanism leads to a broken relationship as soon as a mistake occurs. Hence, these two mechanisms go side by side to promote a perfect cooperation in the repeated interaction setting. In all examples discussed earlier in this introduction, one can already observe the presence of a commitment mechanism, either implicit such as when members of a family or students of the same class must follow certain rules of behaviors, or explicit as is the case for legal contracts signed in advance by individuals or companies [Nesse, 2001]. Even if one does not call upon prior commitment all the time, its existence is important to ensure or facilitate the willingness to apologize for wrongdoing.

The remainder of the paper is structured as follows. Section 2 discusses relevant literature of apology and commitment. Section 3 describes our model for apology with commitments in the IPD, and methods to analyze the model. Then, Section 4 shows our analytical and computer simulation results. The paper ends with a discussion and conclusions obtained from our results.

## 2 Related Work

Direct reciprocity has been the major explanation for the evolution of cooperation in the repeated interaction settings [Axelrod, 1984; Nowak, 2006; Sigmund, 2010]. The iterated Prisoner’s Dilemma (IPD) is the standard framework to investigate this problem. There have been a number of proposed strategies that perform well in this game, i.e. that lead to the emergence of cooperation. The most famous ones are TFT-like strategies (detailed description in the next section) [Axelrod, 1984; Nowak and Sigmund, 1993; Boerlijst *et al.*, 1997; Sigmund, 2010], where the idea of apology is implemented only implicitly. The current work explores how viable explicit apology is in the same type of game, determining the conditions that need to be in place to ensure its survival.

We are by no means the first to use an explicit form of (costly) apology. It has been considered in several Economic and Ecological models, see for example [Ohtsubo and Watanabe, 2009; Okamoto and Matsumura, 2000; Ho, 2012]. Therein an apologizing act was modeled using costly signaling. A common conclusion derived from these studies is that apology must be sincere, i.e. the apologizing signal is costly enough, which is line with observation of our model below. Differently however, the present study investigates whether apology supported by prior commitments—a combined strategic behavior that is ubiquitously observed in

agents/humans interactions—is a viable strategy for the emergence of high levels of cooperation in the repeated interaction setting. And, to the best of our knowledge, this is the first attempt to address this question.

In addition, there have been some computational models of commitment in the repeated interaction setting, see e.g. [de Vos *et al.*, 2001; Back and Flache, 2008]. The idea here is that individuals may benefit from becoming committed to long-term partners, that is, the longer they interact with someone, the more they are willing to interact with them again in the future. The success of such a commitment strategy stems from considering a resource-limited environment—each individual can engage only in a limited number of interactions—hence, committing to more frequently interacting partners is to guarantee to have interactions with more aligned or similar interests. In those models commitments are more like loyalty to ones’ partners. In our model, commitments are explicitly agreed upon in advance, similar to contracts and promises [Nesse, 2001; Han, 2013]. And furthermore, apology is not studied therein at all.

Last but not least, it is important to note a large body of literature on apology [Tzeng, 2004; Park *et al.*, 2012; Vasalou *et al.*, 2008; Utz *et al.*, 2009] and commitment [Wooldridge and Jennings, 1999; Winikoff, 2007] in AI and Computer Science, just to name a few. The main concern of these works is how these mechanisms can be formalized, implemented, and used to enhance cooperation, for example, in human-computer interactions and online market systems [Tzeng, 2004; Park *et al.*, 2012; Vasalou *et al.*, 2008; Utz *et al.*, 2009], as well as general multi-agent systems [Wooldridge and Jennings, 1999; Winikoff, 2007]. In contrast to them, the present work studies the combination of those two strategic behaviors from an evolutionary perspective, that is, whether they can be a viable strategy for the evolution of cooperation. However, as will be seen, the results from our study would provide important insight for the design and deployment of these mechanisms; for instance, what kind of apology should be provided to customers when making mistakes, and whether apology can be enhanced when complemented with commitments to ensure better cooperation, e.g. compensation from customers for wrongdoing.

## 3 Model and Methods

### 3.1 Iterated Prisoner’s Dilemma

Interactions are modeled as symmetric two-player games defined by the payoff matrix

$$\begin{array}{cc} & \begin{array}{c} C \\ D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} R, R & S, T \\ T, S & P, P \end{pmatrix} \end{array}$$

A player who chooses to cooperate (C) with someone who defects (D) receives the sucker’s payoff  $S$ , whereas the defecting player gains the temptation to defect,  $T$ . Mutual cooperation (resp., defection) yields the reward  $R$  (resp., punishment  $P$ ) for both players. Depending on the ordering of these four payoffs, different social dilemmas arise [Hofbauer and Sigmund, 1998; Sigmund, 2010]. Namely, in this work we are concerned with the Prisoner’s Dilemma (PD), where

$T > R > P > S$ . In a single round, it is always best to defect, but cooperation may be rewarding if the game is repeated. In iterated PD (IPD), it is also required that mutual cooperation is preferred over an equal probability of unilateral cooperation and defection ( $2R > T + S$ ); otherwise alternating between cooperation and defection would lead to a higher payoff than mutual cooperation.

For convenience and a clear representation of results, we later mostly use the Donation game [Sigmund, 2010]—a famous special case of the PD—where  $T = b$ ,  $R = b - c$ ,  $P = 0$ ,  $S = -c$ , satisfying that  $b > c > 0$ , where  $b$  and  $c$  stand respectively for “benefit” and “cost” (of cooperation).

The repetition in the IPD is modeled as follows. After the current interaction, another interaction between the interacting players occurs with probability  $\omega \in (0, 1)$ , resulting in an average of  $(1 - \omega)^{-1}$  rounds in the game.

The IPD is played under the presence of noise, that is, an intended action,  $C$  or  $D$ , can fail, and become its opposite, with probability  $\alpha \in [0, 1]$  [Sigmund, 2010].

### 3.2 Strategies in Iterated Prisoner’s Dilemma

The iterated PD is usually known as a story of Rapoport’s tit-for-tat (TFT), which won both Axelrod’s tournaments [Axelrod, 1984; Axelrod and Hamilton, 1981]. *TFT* starts by cooperating, and does whatever the opponent did in the previous round. It will cooperate if the opponent cooperated, and will defect if the opponent defected. But if there are erroneous moves because of noise (i.e. an intended move is wrongly performed with a given execution error, referred here as “noise”), the performance of *TFT* declines, in two ways: (i) it cannot correct errors (e.g., when two *TFT*s play with one another, an erroneous defection by one player leads to a sequence of unilateral cooperation and defection) and (ii) a population of *TFT* players is undermined by random drift when the pure cooperator *AllC* mutants appear (which allows exploiters to grow). Tit-for-tat is then replaced by generous tit-for-tat (GTFT), a strategy that cooperates if the opponent cooperated in the previous round, but sometimes cooperates even if the opponent defected (with a fixed probability  $p > 0$ ) [Nowak and Sigmund, 1992]. *GTFT* can correct mistakes, but still suffers from random drift.

Subsequently, *TFT* and *GTFT* came to be outperformed by win-stay-lose-shift (WSLS) as the winning strategy chosen by evolution [Nowak and Sigmund, 1993]. *WSLS* starts by cooperating, and repeats the previous move whenever it did well, but changes otherwise. *WSLS* corrects mistakes better than *GTFT* and does not suffer random drift. However, it is severely exploited by the pure defector, i.e. the *AllD* players.

### 3.3 Models

**Apology with commitments.** We propose a new strategy, COMA (proposing commitment and apologizing when making mistakes), which before the first interaction of the IPD, proposes its co-player to commit in a long-term cooperation. If the co-player agrees to commit then, in each of the next rounds, if any of them defects and the other cooperates, the defecting one has to apologize the other for his wrongdoing, compensating the amount  $\gamma$ . COMA always honors its

wrongful violation of a commitment, that is, it always cooperates and apologizes when defecting by mistake.

If the co-player does not apologize, the relationship is broken and interactions cease between them<sup>1</sup>. When the relationship is broken (i.e. the commitment is violated), a punishment cost,  $\delta$ , for the defaulting player, provides compensation for the non-defaulting one. To arrange a commitment, COMA players have to pay a cost,  $\epsilon$ . Additionally, if the co-player does not agree to commit, both get 0.

Hence, when playing with COMA, depending on whether committing to a proposal or not, and when committed, whether apologizing for defection or not, three types of defectors can be distinguished:

- *Pure defectors (AllD)*, who never accept a commitment proposal. These players are afraid of having to pay the compensation, as imposed by the commitment deal, if they agreed.
- *Fake apologizers (FAKA)*, who accept to commit, but then defect and do not apologize. However, they accept apology from the other (for example, when because of noise, COMA defects in the first round while FAKA cooperates). These players assume that they can exploit COMA without suffering a severe penalty.
- *Fake committers (FAKC)*, who accept to commit, but then always defect and apologize for wrongdoing to prolong the relationship. These players assume the cost of apology can be offset by the benefit from exploiting their COMA co-players.

From a rational standpoint, one can see that any strategy which starts by cooperating and continues to cooperate as long as the opponent cooperates (or defects, but apologizes), should agree to commit when playing with a COMA. Reaping benefit from mutual cooperation is their motive, and a positive compensation is guaranteed when being exploited. Furthermore, in case of an execution error, such cooperative strategies should compensate to avoid the penalty imposed by the commitment deal, on the one hand, and, on the other hand, receive the benefit from further mutual cooperation. These strategic behaviors are even more advantageous in a pairwise comparison of the strategy with COMA, because commitment proposers have to pay the arrangement cost only when the commitment is agreed upon, and moreover, are the only ones of the pair who have to do so.

As such, we only consider cooperative strategies, including AllC, TFT, GTFT and WSLS, that commit when being proposed, and apologize when defecting. Thus, to evaluate the performance of COMA, we examine two different settings corresponding to two different population compositions, including the following strategies

- **(S1):** COMA, AllC, AllD, FAKA and FAKC.
- **(S2):** COMA, AllC, AllD, FAKA, COMA, together with TFT, GTFT and WSLS.

<sup>1</sup>The results obtained below remain robust if, instead of both players getting 0 when a relationship is broken, they both obtain payoffs from mutual defections (subject to noise) for all forthcoming interactions. Note that when COMA has decided to always defect, the best option of the co-player is to always defect as well.

**Apology without commitments.** To reflect upon the role of commitments in supporting the establishment of apology as a powerful mechanism for conflict resolution, we examine a *pure apology* strategy, *AP*, who does not arrange commitments of any form (whether explicit or implicit). *AP* can be defined as *COMA* with  $\epsilon = \delta = 0$ . *AP* does not have to pay the initial cost to arrange commitments, but its co-player is not subject to penalty when defecting and not apologizing. The first comparison of *AP* with *COMA* is that *AP* would perform better when playing against apologizing strategies (i.e. *AIIIC* and *FAKC*) because it can avoid the commitment arrangement cost. However, it would perform worse against fake apologizers (*FAKA*) who now do not have to suffer any consequence (see already Fig. 1). Furthermore, because there is no penalty for not apologizing, a strategy *CFAKA* who always cooperates if there is no commitment in place, but does not apologize for mistake, can also exploit *AP*. Hence, we will examine the following setting where the population consists of these strategies:

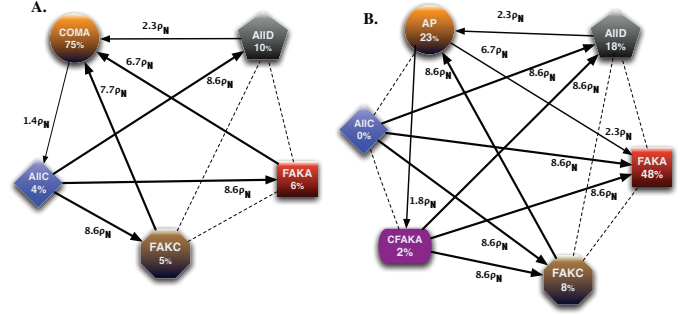
- (**S1'**): *AP*, *AIIIC*, *AIID*, *FAKA*, *CFAKA*, and *FAKC*.

### 3.4 Evolution in finite populations

Our analysis is based on Evolutionary Game Theory methods for finite populations [Nowak *et al.*, 2004; Imhof *et al.*, 2005]. In such a setting, individuals' payoff represents their *fitness* or social *success*, and evolutionary dynamics is shaped by social learning [Hofbauer and Sigmund, 1998; Sigmund, 2010], whereby the most successful individuals will tend to be imitated more often by the others. In the current work, social learning is modeled using the so-called pairwise comparison rule [Traulsen *et al.*, 2006], assuming that an individual *A* with fitness  $f_A$  adopts the strategy of another individual *B* with fitness  $f_B$  with probability given by the Fermi function,  $(1 + e^{-\beta(f_B - f_A)})^{-1}$ . The parameter  $\beta$  represents the 'imitation strength' or 'intensity of selection', i.e., how strongly the individuals base their decision to imitate on fitness comparison. For  $\beta = 0$ , we obtain the limit of neutral drift – the imitation decision is random. For large  $\beta$ , imitation becomes increasingly deterministic.

In the absence of mutations or exploration, the end states of evolution are inevitably monomorphic: once such a state is reached, it cannot be escaped through imitation. We thus further assume that, with a certain mutation probability, an individual switches randomly to a different strategy without imitating another individual. In the limit of small mutation rates, the behavioral dynamics can be conveniently described by a Markov Chain, where each state represents a monomorphic population, whereas the transition probabilities are given by the fixation probability of a single mutant [Fudenberg and Imhof, 2005; Imhof *et al.*, 2005; ?]. The resulting Markov Chain has a stationary distribution, which characterizes the average time the population spends in each of these monomorphic end states.

Let  $N$  be the size of the population. Suppose there are at most two strategies in the population, say,  $k$  individuals using strategy A ( $0 \leq k \leq N$ ) and  $(N - k)$  individuals using strategies B. Thus, the (average) payoff of the individual that



**Figure 1: Transition probabilities and stationary distributions:** (A) of strategies in (S1) and (B) of strategies in (S1'). The black arrows are only shown for the transition directions that are rather more likely than neutral. Parameters:  $\epsilon = 1$ ;  $\delta = 2.5$  (panel A). In both panels:  $N = 100$ ;  $\omega = 0.9$ ;  $b = 2$ ;  $c = 1$ ;  $\alpha = 0.05$ ;  $\beta = 0.1$ ;  $\gamma = 2$ ; Note that  $\rho_N = 1/N$  denotes the neutral fixation.

uses A and B can be written as follows, respectively,

$$\begin{aligned} \Pi_A(k) &= \frac{(k-1)\pi_{A,A} + (N-k)\pi_{A,B}}{N-1}, \\ \Pi_B(k) &= \frac{k\pi_{B,A} + (N-k-1)\pi_{B,B}}{N-1}, \end{aligned} \quad (1)$$

where  $\pi_{X,Y}$  stands for the payoff an individual using strategy *X* obtained in an interaction with another individual using strategy *Y*.

Now, the probability to change the number  $k$  of individuals using strategy A by  $\pm 1$  in each time step can be written as

$$T^\pm(k) = \frac{N-k}{N} \frac{k}{N} \left[ 1 + e^{\mp\beta[\Pi_A(k) - \Pi_B(k)]} \right]^{-1}. \quad (2)$$

The fixation probability of a single mutant with a strategy A in a population of  $(N-1)$  individuals using B is given by [Traulsen *et al.*, 2006; Fudenberg and Imhof, 2005]

$$\rho_{B,A} = \left( 1 + \sum_{i=1}^{N-1} \prod_{j=1}^i \frac{T^-(j)}{T^+(j)} \right)^{-1}. \quad (3)$$

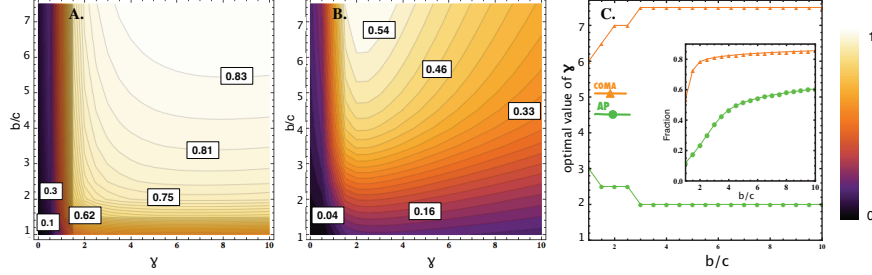
In the limit of neutral selection (i.e.  $\beta = 0$ ),  $\rho_{B,A}$  equals the inverse of population size,  $1/N$ .

Considering a set  $\{1, \dots, q\}$  of different strategies, these fixation probabilities determine a transition matrix  $M = \{T_{ij}\}_{i,j=1}^q$ , with  $T_{ij, j \neq i} = \rho_{ji}/(q-1)$  and  $T_{ii} = 1 - \sum_{j=1, j \neq i}^q T_{ij}$ , of a Markov Chain. The normalized eigenvector associated with the eigenvalue 1 of the transposed of  $M$  provides the stationary distribution described above [Fudenberg and Imhof, 2005; Imhof *et al.*, 2005], describing the relative time the population spends adopting each of the strategies.

**Analytical condition for risk-dominance.** An important criteria for pairwise comparison of strategies in finite population dynamics is *risk-dominance*, that is, whether it is more probable that an A mutant fixating in a homogeneous population of individuals adopting B than a B mutant fixating in

$$\begin{pmatrix} (b-c)(1-\alpha) & b\alpha-c(1-\alpha) & b\alpha-c(1-\alpha) & b\alpha-c(1-\alpha) & (b-c)(1-\alpha) \\ b-b\alpha-c\alpha & (b-c)\alpha & (b-c)\alpha & (b-c)\alpha & \mathbf{0} \\ b-b\alpha-c\alpha & (b-c)\alpha & (b-c)\alpha & (b-c)\alpha & m_2 \\ b-b\alpha-c\alpha & (b-c)\alpha & (b-c)\alpha & (b-c)\alpha & m_4 \\ (b-c)(1-\alpha)-\bar{\epsilon} & \mathbf{0} & m_1 & m_3 & (b-c)(1-\alpha)-\frac{\epsilon}{2} \end{pmatrix}$$

**Figure 2: Payoff matrix for the five strategies, AllC, AllD, FAKA, FAKC and COMA**, where for the sake of a clean representation we denote  $\bar{\epsilon} = \epsilon(1-\omega)$ ;  $m_1 = \frac{-c+\delta+\alpha(b+c-2\delta+b\omega-\gamma\omega)}{1+(1-\alpha)^2\omega} - \bar{\epsilon}$ ;  $m_2 = \frac{b-b\alpha-\delta-\alpha(c-2\delta+c\omega-\gamma\omega)}{1+(1-\alpha)^2\omega}$ ;  $m_3 = (1-2\alpha)\gamma + b\alpha + c\alpha - c - \bar{\epsilon}$ ; and  $m_4 = b - b\alpha - c\alpha - (1-2\alpha)\gamma$ . Note that all the terms of order  $O(\alpha^2)$  in the numerators of  $m_1$  and  $m_2$  are ignored.



**Figure 3: (A) Frequency of COMA as a function of  $b/c$  and  $\gamma$  in (S1); (B) Frequency of AP as a function of  $b/c$  and  $\gamma$  in (S1').** For each game configuration, COMA has a significantly greater frequency than AP. AP performs poorly for difficult IPDs (i.e. for small benefit-to-cost ratio  $b/c$ ). For all  $b/c$ , the optimal apology cost  $\gamma$  for AP tends to converge to a threshold, beyond which AP fraction decreases (panel B); For COMA, such an optimal value of apology cost is much higher, and monotonically increases with  $b/c$  (panel A), see panel C. Parameters:  $N = 100$ ;  $\omega = 0.9$ ;  $\delta = 2.5$ ;  $\epsilon = 1$ ;  $\gamma = 2$ ;  $\beta = 0.1$ .

a homogeneous population of individuals adopting A. When the first is more likely than the latter (i.e.  $\rho_{B,A} > \rho_{A,B}$ ), A is said to be *risk-dominant* against B [Kandori *et al.*, 1993; Nowak, 2006], which holds for any intensity of selection and in the limit of large  $N$  when

$$\pi_{A,A} + \pi_{A,B} > \pi_{B,A} + \pi_{B,B}. \quad (4)$$

## 4 Results

### 4.1 Analytical conditions for viability of COMA

To begin with, we derive the (average) payoff matrix for all pairwise interactions of strategies in (S1), in the presence of noise (Fig. 2, see Appendix A for details).

Using Eq. (4) one can show that COMA is risk-dominant against AllD, FAKC and FAKA, respectively, whenever the following conditions are satisfied:

$$\begin{aligned} \epsilon &< \frac{2(b-c)(1-2\alpha)}{1-\omega}, \\ \gamma &> \frac{3\epsilon(1-\omega)}{4(1-2\alpha)} + c, \\ \delta &> a_2\gamma + a_1\epsilon + a_0, \end{aligned} \quad (5)$$

where  $a_2 = \frac{\alpha\omega}{1-2\alpha}$ ;  $a_1 = \frac{3(1-\omega)(1+\omega-2\alpha\omega)}{4(1-2\alpha)}$ ; and  $a_0 = 2c + \frac{(2c+3b\alpha-b-10c\alpha)\omega}{2(1-2\alpha)}$ . It is easily seen that  $a_1, a_2 > 0$ , assuming that the noise level is not too large, namely  $\alpha < 0.5$ . The first condition means that for COMA to be risk-dominant against the non-committing pure defectors AllD, the arrangement cost needs to be justified with respect to the potential reward of cooperation ( $R = b - c$ ), and that the larger the average number of rounds,  $(1-\omega)^{-1}$ , the easier the cost is justified. The second condition means that the associated compensation for a mistake needs to take into account the cost of

cooperation as well as the cost of commitment arrangement, in order for COMA to win against the fake committers FAKC. The last condition states that, for COMA to be risk-dominant against the fake apologizers FAKA, the compensation cost associated with a commitment deal needs to positively correlate with the cost of apology (which encourages co-players to apologize to keep on the relationship), and also take into account the costs of arranging commitment and of cooperation.

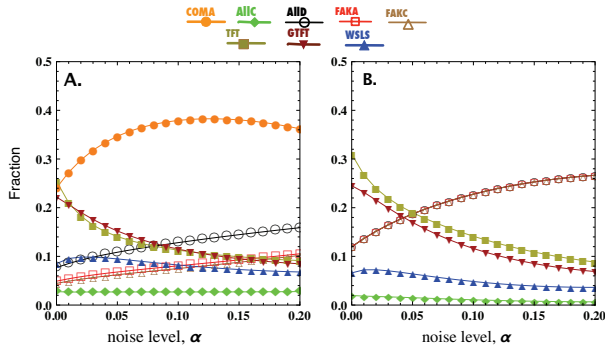
For AP, as a special case of COMA where  $\delta = \epsilon = 0$ , the first condition becomes apparent. The second condition condition becomes  $\gamma > c$ . However, it becomes extremely more difficult for the third one,  $a_2\gamma + a_0 < 0$ , to hold. The necessary condition is  $a_0 < 0$ , which requires  $b > 6c^2$ . This is an extremely easy condition for cooperation to emerge in IPD. TFT and WSLs can easily establish high level of cooperation in such cases [Imhof *et al.*, 2007].

### 4.2 Apology with or without commitments

The above observations can be seen in Fig. 1, which shows the transition probabilities and stationary distribution of strategies in (S1) and (S1') settings. With the same game configuration, COMA dominates the population, while AP has a significantly lower frequency. Note the direction of transition from FAKA to COMA is reversed for AP. Our additional analysis shows that including CFKA in the population with COMA, i.e. (S1), does not change the relative performance of COMA, as the strategy is dominated by AllC which is already in the population.

For varying benefit-to-cost ratio ( $b/c$ ), COMA has a significantly higher frequency than AP (Fig. 3). AP performs

<sup>2</sup>It is because  $a_0 > \frac{(4c(1-2\alpha)+(2c+3b\alpha-b-10c\alpha)\omega)}{2(1-2\alpha)} = \frac{(6c-b)(1-3\alpha)\omega}{2(1-2\alpha)}$ . Hence, for small enough  $\alpha$ , namely  $\alpha < 1/3$ , from  $a_0 < 0$  we obtain  $b > 6c$ .



**Figure 4: Frequencies of each strategy in (S2) (panel A) and (S2), but without COMA (panel B), as a function of noise level  $\alpha$ .** Without COMA, pure defectors prevail, while with COMA, the latter becomes the prevalent strategy. Parameters:  $N = 100$ ;  $\omega = 0.9$ ;  $b = 2$ ;  $c = 1$ ;  $\delta = 2.5$ ;  $\epsilon = 1$ ;  $\gamma = 2$ ;  $\beta = 0.1$ .

poorly for difficult IPD (i.e. for small  $b/c$ ). For all cases, the optimal apology cost  $\gamma$  for AP tends to converge to a threshold, beyond which AP fraction decreases (see Fig. 3B); For COMA, such an optimal value of apology cost is much higher, and monotonically increases with  $b/c$  (Fig. 3A), see Fig. 3C. That is, in a committed interaction, a much higher frequency of apologizing acts are used. Interestingly, much higher costly apology should willingly be used in the committed interactions. That is, ‘commitments bring about sincerity’.

For both cases, with or without commitments, apology works poorly if it is not costly enough (Figs. 3A and 3B). It means that for apology to function, it must be sincere. This observation is in accordance with previous experimental evidence [Ohtsubo and Watanabe, 2009; Takaku *et al.*, 2001].

### 4.3 Apology supported by commitments prevails

We examine the performance of COMA when in a population now extended with strategies TFT, GTFT and WLSL, i.e. (S2) setting. Fig. 4A shows that COMA outperforms all other strategies under different levels of noise. In addition, to clarify the role of COMA, it is removed from the population and we run the simulation with the same parameters’ values (Fig. 4B). Defectors now take over the population. This is partly because COMA can deal with noise better than TFT, GTFT and WLSL, and furthermore, it can deal with all kinds of defectors much better than them.

It is noteworthy that our additional analysis shows that these achieved remarkable performances of COMA are robust for various average number of rounds of the IPD and for different levels of intensity of selection, as well as a wide range of the commitment parameters ( $\epsilon$  and  $\delta$ ).

## 5 Concluding Remarks

We have shown, analytically as well as by numerical simulations, that apology supported by commitments can promote the emergence of cooperation, in the sense that the population spends most of the time in the homogenous state in which individuals adopt such a strategy. Note that a population of COMA can maintain a perfect level of cooperation, even in

the presence of noise, as well as a population of unconditional cooperators AllC can.

To reflect on the role of commitment for the success of apology, we have shown that apology without commitments performs poorly, especially when only a small benefit can be obtained from cooperation (i.e. difficult IPDs). It is so because they can be easily exploited by the fake apologizers.

Most interestingly, our model predicts that individuals tend to use much costlier apology in committed interactions than otherwise, in order to better avoid fake committers and fake apologizers. And in line with prior experimental evidence, we have shown that, to function properly, apology needs to be sincere, whether it is to resolve conflict in a committed relationship or in commitment-free ones.

## A Deriving payoff matrix in presence of noise

We describe how to derive the analytical payoff matrix for strategies in the main text, see Fig. 2, using a similar method as that in [Sigmund, 2010, Chapter 3]. The strategies are at most one-step memory, i.e. taking into account at most the moves in the last game round. There are four possible states, corresponding to the four possible game situations ( $R, S, T, P$ ) in the last encounter. We enumerate these states by  $state_i$ , with  $1 \leq i \leq 4$ .

We consider stochastic strategies  $(f, q_1, q_2, q_3, q_4) \in [0, 1]^5$  where  $f$  is the propensity to play  $C$  in the initial round, and  $q_i$  are the propensities to play  $C$  after having been at  $state_i$ ,  $1 \leq i \leq 4$ . Let us assume that player 1 using  $(f_1, p_1, p_2, p_3, p_4)$  encounters a co-player 2 using  $(f_2, q_1, q_2, q_3, q_4)$ . We have a Markov chain in the state space  $\{state_1, \dots, state_4\}$ . The transition probabilities are given by the stochastic matrix  $Q$  below. Note that one player’s  $S$  is the other player’s  $T$

$$Q = \begin{pmatrix} p_1 q_1 & p_1(1-q_1) & (1-p_1)q_1 & (1-p_1)(1-q_1) \\ p_2 q_3 & p_2(1-q_3) & (1-p_2)q_3 & (1-p_2)(1-q_3) \\ p_3 q_2 & p_3(1-q_2) & (1-p_3)q_2 & (1-p_3)(1-q_2) \\ p_4 q_4 & p_4(1-q_4) & (1-p_4)q_4 & (1-p_4)(1-q_4) \end{pmatrix}.$$

The initial probabilities for the four states are given by the vector:  $\mathbf{f} = \{f_1 f_2, f_1(1-f_2), f_2(1-f_1), (1-f_1)(1-f_2)\}$ . In the next round, these probabilities are given by  $\mathbf{f}Q$ , and in round  $n$  by  $\mathbf{f}Q^n$ . We denote by  $\mathbf{g}$  the vector  $\{X, Y, Z, W\}$ , where  $X, Y, Z, W$  are the payoffs player 1 obtains when the game state is  $R, S, T, P$ , respectively. The payoff for player 1 in round  $n$  is given by

$$A(n) = \mathbf{g} \cdot \mathbf{f}Q^n. \quad (6)$$

For  $\omega < 1$  the average payoff per round is  $(1-\omega) \sum w^n A(n)$  [Sigmund, 2010], i.e.,

$$(1-\omega) \mathbf{g} \cdot \mathbf{f}(Id - \omega Q)^{-1} \quad (7)$$

where  $Id$  is the identity matrix of size 4.

For instance, AllC is given by  $\{1-\alpha, 1-\alpha, 1-\alpha, 1-\alpha, 1-\alpha\}$ , and FAKA is given by  $\{\alpha, \alpha, \alpha, \alpha, \alpha\}$ . In the absence of commitment and apology, we have  $g = \{R, S, T, P\}$ . The payoffs can be found in [Sigmund, 2010]. When playing with COMA, i.e. in the presence of apology and commitment,  $g$  is given by different formulas. For instance, when COMA plays with FAKA, for COMA,  $g = \{R, S + \delta, T - \gamma, P\}$ , and for FAKA,  $g = \{R, S + \gamma, T - \delta, P\}$ . Similarly for other pairs of strategies.

## References

- [Abeler *et al.*, 2010] J. Abeler, J. Calaki, K. Andree, and C. Basek. The power of apology. *Economics Letters*, 107(2):233 – 235, 2010.
- [Atran *et al.*, 2007] S. Atran, R. Axelrod, R. Davis, et al. Sacred barriers to conflict resolution. *Science*, 317:1039–1040, 2007.
- [Axelrod and Hamilton, 1981] R. Axelrod and W.D. Hamilton. The evolution of cooperation. *Science*, 211:1390–1396, 1981.
- [Axelrod, 1984] R. Axelrod. *The Evolution of Cooperation*. Basic Books, ISBN 0-465-02122-2, 1984.
- [Back and Flache, 2008] Istvan Back and Andreas Flache. The Adaptive Rationality of Interpersonal Commitment. *Rationality and Society*, 20(1):65–83, 2008.
- [Boerlijst *et al.*, 1997] Maarten C. Boerlijst, Martin A. Nowak, and Karl Sigmund. The logic of contrition. *Journal of Theoretical Biology*, 185(3):281 – 293, 1997.
- [de Vos *et al.*, 2001] de Vos, R Smaniotto, and D Elsas. Reciprocal altruism under conditions of partner selection. *Rationality and Society*, 13(2):139–183, 2001.
- [Fudenberg and Imhof, 2005] D. Fudenberg and L. A. Imhof. Imitation processes with small mutations. *Journal of Economic Theory*, 131:251–262, 2005.
- [Han *et al.*, 2011] T. A. Han, L. M. Pereira, and F. C. Santos. Intention recognition promotes the emergence of cooperation. *Adaptive Behavior*, 19(3):264–279, 2011.
- [Han *et al.*, 2012a] T. A. Han, L. M. Pereira, and F. C. Santos. The emergence of commitments and cooperation. In *Procs. of AAMAS-2012*, pages 559–566, 2012.
- [Han *et al.*, 2012b] T. A. Han, L. M. Pereira, and F. C. Santos. Intention Recognition, Commitment, and The Evolution of Cooperation. In *Proceedings of IEEE Congress on Evolutionary Computation*, pages 1–8. IEEE Press, June 2012.
- [Han, 2013] T. A. Han. *Intention Recognition, Commitments and Their Roles in the Evolution of Cooperation: From Artificial Intelligence Techniques to Evolutionary Game Theory Models*, volume 9. Springer SAPERE series, May 2013.
- [Ho, 2012] B. Ho. Apologies as signals: with evidence from a trust game. *Management Science*, 58(1):141–158, 2012.
- [Hofbauer and Sigmund, 1998] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.
- [Imhof *et al.*, 2005] L. A. Imhof, D. Fudenberg, and Martin A. Nowak. Evolutionary cycles of cooperation and defection. *Proc. Natl. Acad. Sci. U.S.A.*, 102:10797–10800, 2005.
- [Imhof *et al.*, 2007] L. A. Imhof, D. Fudenberg, and M. A. Nowak. Tit-for-tat or win-stay, lose-shift? *Journal of Theoretical Biology*, 247(3):574 – 580, 2007.
- [Kandori *et al.*, 1993] M. Kandori, G.J. Mailath, and R. Rob. Learning, mutation, and long run equilibria in games. *Econometrica*, 61:29–56, 1993.
- [Liang, 2002] B.A. Liang. A system of medical error disclosure. *Quality and Safety in Health Care*, 11(1):64–68, 2002.
- [Nesse, 2001] R. M. Nesse. Natural selection and the capacity for subjective commitment. In Randolph M. Nesse, editor, *Evolution and the capacity for commitment*, pages 1–44. New York: Russell Sage, 2001.
- [Nowak and Sigmund, 1992] M. A. Nowak and K. Sigmund. Tit for tat in heterogeneous populations. *Nature*, 355:250–253, 1992.
- [Nowak and Sigmund, 1993] M. A. Nowak and K. Sigmund. A strategy of win-stay, lose-shift that outperforms tit-for-tat in prisoner’s dilemma. *Nature*, 364:56–58, 1993.
- [Nowak *et al.*, 2004] M. A. Nowak, A. Sasaki, C. Taylor, and D. Fudenberg. Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 428:646–650, 2004.
- [Nowak, 2006] M. A. Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560, 2006.
- [Ohtsubo and Watanabe, 2009] Y. Ohtsubo and E. Watanabe. Do sincere apologies need to be costly? test of a costly signaling model of apology. *Evolution and Human Behavior*, 30(2):114–123, 2009.
- [Okamoto and Matsumura, 2000] K. Okamoto and S. Matsumura. The evolution of punishment and apology: an iterated prisoner’s dilemma model. *Evolutionary Ecology*, 14(8):703–720, 2000.
- [Park *et al.*, 2012] S.J. Park, C.M. MacDonald, and M. Khoo. Do you care if a computer says sorry? In *Proceedings of the Designing Interactive Systems Conference*, pages 731–740, 2012.
- [Petrucci, 2002] C.J. Petrucci. Apology in the criminal justice setting: Evidence for including apology as an additional component in the legal system. *Behavioral sciences & the law*, 20(4):337–362, 2002.
- [Sigmund, 2010] K. Sigmund. *The Calculus of Selfishness*. Princeton University Press, 2010.
- [Takaku *et al.*, 2001] S. Takaku, B. Weiner, and K.I. Ohbuchi. A cross-cultural examination of the effects of apology and perspective taking on forgiveness. *Journal of Language and Social Psychology*, 20(1-2):144–166, 2001.
- [Traulsen *et al.*, 2006] A. Traulsen, M. A. Nowak, and J. M. Pacheco. Stochastic dynamics of invasion and fixation. *Phys. Rev. E*, 74:11909, 2006.
- [Trivers, 1971] R. L. Trivers. The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46:35–57, 1971.
- [Tzeng, 2004] Jeng-Yi Tzeng. Toward a more civilized design: studying the effects of computers that apologize. *International Journal of Human-Computer Studies*, 61(3):319 – 345, 2004.
- [Utz *et al.*, 2009] S. Utz, U. Matzat, and C. Snijders. On-line reputation systems: The effects of feedback comments and reactions on building and rebuilding trust in on-line auctions. *Intl. Journal of Electronic Commerce*, 13(3):95–118, 2009.
- [Vasalou *et al.*, 2008] A. Vasalou, A. Hopfensitz, and J.V. Pitt. In praise of forgiveness: Ways for repairing trust breakdowns in one-off online interactions. *International Journal of Human-Computer Studies*, 66(6):466–480, 2008.
- [Winikoff, 2007] M. Winikoff. Implementing commitment-based interactions. In *Procs. of AAMAS-2007*, pages 868–875, 2007.
- [Wooldridge and Jennings, 1999] M. Wooldridge and N. R. Jennings. The cooperative problem-solving process. In *Journal of Logic and Computation*, pages 403–417, 1999.