Apology and forgiveness evolve to resolve failures in cooperative agreements

Luis A. Martinez-Vaquero,[1,2] The Anh Han,[3] Luís Moniz Pereira,[4] and Tom Lenaerts[1,2]

[1]*AI lab, Computer Science Department, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium*
[2]*MLG, Département d'Informatique, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussels, Belgium*
[3]*School of Computing, Teesside University, Borough Road, Middlesbrough, UK, TS1 3BA*
[4]*NOVA Laboratory for Computer Science and Informatics, Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal*

Making agreements on how to behave has been shown to be an evolutionarily viable strategy in one-shot social dilemmas. However, in many situations agreements aim to establish long-term mutually beneficial interactions. Our analytical and numerical results reveal for the first time under which conditions revenge, apology and forgiveness can evolve and deal with mistakes within ongoing agreements in the context of the Iterated Prisoners Dilemma. We show that, when the agreement fails, participants prefer to take revenge by defecting in the subsisting encounters. Incorporating costly apology and forgiveness reveals that, even when mistakes are frequent, there exists a sincerity threshold for which mistakes will not lead to the destruction of the agreement, inducing even higher levels of cooperation. In short, even when to err is human, revenge, apology and forgiveness are evolutionarily viable strategies which play an important role in inducing cooperation in repeated dilemmas.

Corresponding author
Prof. Dr. Tom Lenaerts
Université Libre de Bruxelles
Boulevard du Triomphe CP212
1050 Brussels
email: tlenaert@ulb.ac.be
phone: +32 2 650 6004

Recently, our innate capacity to create, and commit to, prior agreements [1–3] has been proposed as an evolutionarily viable strategy inducing cooperative behavior in social dilemmas. It provides an alternative to different forms of punishment of inappropriate behavior, or of rewards to stimulate the proper one [4–8]. Commitments – defined as prior agreements with potentially posterior compensations in case the agreements fail – are wide-spread in human societies at different scales, from personal relationships such as marriage to international and organisational ones such as alliances among companies and countries [1–3, 9, 10]. Anthropological data reveals that commitment strategies, as for instance demand-sharing [11], have played an essential role in early hunter-gatherer societies. A recent body of economic experiments show that arranging prior commitments promote cooperation in diverse scenarios from one-shot to repeated games [12–14]. Analytical and numerical methods have shown that commitments are evolutionarily viable when the cost of arranging them is sufficiently small compared to the cost of cooperation both in the one-shot pairwise prisoners dilemma [15] and the one-shot public goods game [16].

However, commitment deals, like some examples mentioned earlier, are most often established to ensure favourable interactions over longer time periods, implying repeated encounters between the actors that established the agreement, as well as the appeal of repeated benefits. Experiments have shown that commitment facilitates cooperation in long-term interactions [13, 17], especially when it is voluntary. Moreover, long-term commitments are most likely more cost-efficient as the cost of setting up the agreement is paid only once for the entire duration of the agreement. Interestingly, commitment may also induce behavioral differences in repeated games: As suggested in [2], the individuals' preferred behavior in repeated interactions may shift from a conditional reciprocal to an unconditionally cooperative behavior, which will indeed be confirmed analytically and numerically in this manuscript.

Using methods from Evolutionary Game Theory [18, 19], we provide for the first time analytical and numerical insight into the viability of commitment strategies in repeated social interactions, which will be modeled through the Iterated Prisoners Dilemma (IPD) [20]. In order to study commitment strategies in the IPD a number of behavioral complexities need to be addressed. First, agreements may end before the recurring interactions are finished. As such, strategies need to take into account how to behave when the agreement is present and when it is absent, on top of proposing, accepting or rejecting such agreements in the first place. Second, as it was shown within the context of direct reciprocity [21], individuals need to deal with mistakes made by the opponent or by themselves, caused for instance by "trembling hands" or "fuzzy minds" [19, 22]: A decision needs to be made on whether to continue the agreement, or end it collecting the compensation resulting from the other's defection.

As errors might lead to misunderstandings or even breaking of commitments, individuals may have acquired sophisticated strategies to ensure that mistakes are not repeated or that profitable relationships may continue. Revenge and forgiveness may have evolved exactly to cope with those situations [23, 24]: The threat of revenge, through some punishment of withholding of a benefit, may discourage interpersonal harm. Yet often one cannot distinguish with enough certainty if the other's behavior is intentional or just accidental [25, 26]. In the latter case, forgiveness provides a restorative mechanism that ensures that beneficial relationships can still continue, notwithstanding the initial harm. An essential ingredient for forgiveness, analysed in this work, seems to be (costly) apology [23], a point emphasised in [27].

The importance of apology and forgiveness for sustaining long term relationships has been shown in different experiments [28–31]. Apology and forgiveness are of interest as they remove the interference of external institutions, which can be quite costly to all parties involved, in order to ensure cooperation. Evidence shows that there is a much higher chance that customers stay with a company (they hence forgive) that apologises for mistakes [28]. Apology leads to fewer lawsuits with lower settlements in medical error situations [32]. Apology even enters the law as an effective mechanism of resolving conflicts [33, 34]. Hence, it is important to know how apology and forgiveness can help coping with misunderstanding, on either side, in an internal way, without jeopardising the ongoing commitment. Even without explicit apology, the participants in an IPD seem to use a form of implicit apology, by cooperating in several subsequent rounds after making a mistake [35–37]. Yet, such an unclear apology might not thoroughly resolve the misunderstanding, as is the case for TFT-like strategies [35, 36, 38, 39].

In this work, once the viability of the commitment strategy within the context of the IPD is analysed, we analytically and numerically determine when explicit apology and forgiveness are evolutionarily viable, and how sincere apology needs to be for forgiveness, thus sustaining the mutually beneficial relationship, which so far we are aware have never been provided.

Results

Defining all strategies

We consider a finite population of N individuals, with $N = 100$ in our analysis. At the beginning of each generation, individuals are randomly matched to play an IPD game. In each round of this game they can either cooperate (C) or defect (D), acquiring a payoff given by the Donation game [19] —an instance of the PD— as represented by the following parametrised payoff matrix with $b > c$:

$$\begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} b-c & -c \\ b & 0 \end{pmatrix} \end{array} \tag{1}$$

With a probability ω the encounter between two individuals is repeated for another round, leading to an average number of rounds per IPD interaction $R_T = (1-\omega)^{-1}$. Individuals may make implementation mistakes,*i.e.* playing D when they intend to play C and vice versa, with a probability α. At the end of a generation, more successful individuals—those that accumulated higher total payoffs—are more likely to be imitated by less successful individuals (see Methods for more details).

At the first encounter before playing the IPD, individuals can agree to play C in every round of the game. To set up a commitment one of the players has to propose it, at a cost ϵ (players share that cost if both are proposers), while the co-player needs to decide whether to accept it. The commitment lasts as long as both players fulfill their commitment,*i.e.* they play C. If one defects then she has to pay a compensation δ to the other player and the commitment is broken.

When neither player proposes a commitment or the commitment ends, both individuals play a reactive strategy [35], which is modelled by a triplet (p_0, p_C, p_D): p_0 represents the probability of cooperating in the first round, p_C the probability of cooperating in the current round if in the previous round the co-player cooperated and similarly for p_D, which is the probability of cooperating in the current round if the co-player defected in the previous round. In this work, we will only consider pure reactive strategies in the presence of noise (α): always cooperating AllC, $(1-\alpha, 1-\alpha, 1-\alpha)$, always defecting AllD (α, α, α), TFT $(1-\alpha, 1-\alpha, \alpha)$ and *anti*-TFT (ATFT) $(1-\alpha, \alpha, 1-\alpha)$. All these strategies, except AllD, play C in the first round.

Thus, the full strategy S_i of any individual in our model is defined by three parameters $S_i = (S_c, S_{in}, S_{out})_i$, where

- $S_c \in \{P, A, NC\}$ represents whether the strategy is a *proposing* player who proposes and accepts commitments (P), an *accepting* player who does not propose a commitment but accepts those that are proposed (A), or a *non-committing* player that never accepts a commitment proposal strategy (NC). We consider at this point only simultaneous interactions, meaning that when both players propose to commit they do this at the same time, hence sharing the commitment cost.

- $S_{in} \in \{C, D\}$ indicates the behavior the player chooses when she is in a commitment: cooperating with a probability $1-\alpha$ (C) or α (D). Non-committers do not have any S_{in} strategy since they never participate in commitments, which will be represented by $S_{in} =$"$-$" in the strategy S_i.

- $S_{out} \in \{AllC, AllD, TFT, ATFT\}$ represents the reactive strategy chosen by the player when not in a commitment.

Hence the strategy of a defector or a cooperator is represented here by, respectively, $(NC, -, AllD)$ and $(A, C, AllC)$. A strategy that proposes a commitment, and honours the agreement, while defecting when the agreement is broken is represented by $(P, C, AllD)$. The strategies FAKE and FREE, discussed in [15], could be represented by $(A, D, *)$, with "$*$" representing any of the S_{out} options, and $(A, C, *)$ respectively.

Finally, we need to consider that when the commitment is broken, or when the agreement is not created, individuals can decide to simply stop the game, which results in a zero payoff for all following rounds. Proposers that do not play when the commitment is broken will not have an S_{out} strategy. These two possibilities lead to four different scenarios $\mathfrak{T}$:

- **PP**: individuals continue to play their reactive strategies both when the commitment is not set up and after an established commitment is broken,

- **NP**: they play their reactive strategies once a commitment ends, but stop interacting if a commitment is rejected (even if additional rounds could be played),

- **PN**: individuals play their reactive strategy if a proposed commitment is not accepted, but stop interacting if a commitment is broken,

- **NN**: individuals refuse to play in any of these situations.

Given these four scenarios, one can identify 20 strategies for PP, NP and PN scenarios, and 14 strategies for the NN scenario. The number of strategies will increase when an apology-forgiveness mechanism is introduced, as is shown later. It is not clear which of these scenarios leads to the most cooperative outcome. Our results, as discussed below, aim to reveal, when individuals can choose by themselves which scenario to use, which of the four scenarios is evolutionary more viable than the other. In other words, is it best to always play the game, using their S_i strategies or should one refuse the interaction when the agreement cannot be established or when it is broken (or both)? The details on how the payoffs for each strategy are calculated are provided in Methods.

The emergence of revenge after the commitment is broken

We first study commitments in the absence of an apology-forgiveness mechanism. In this situation, there are two main differences with the model in [15] next to the iterated nature of the game: the inclusion of noisy C and D actions and how players react when commitments fail or cannot be established (previously called *scenarios PP, PN, NP* and *NN*).

In Figure 1a, where the four scenarios are analysed separately, we plot for each scenario the frequencies of the most dominant strategies relative to the frequency of defectors (*i.e.* $(NC, -, AllD)$ which do not commit when requested, hence have no S_{in}, and always defect when no agreement is established or an established one is broken) as a function of the errors (α) that individuals can make while playing. These relative frequencies reveal for which noise levels the latter can suppress the defection strategy (see Supplementary Table 1 for the outcome for all strategies). In all four scenarios the dominating strategies are, for most noise levels, of the type $(P, C, *)$, which correspond to proposers that cooperate as long as the commitment lasts and then behave reactively when the commitment is broken (see Supplementary Information for additional data on the results obtained for different conditions, including the benefit-to-cost ratio, the cost of establishing a commitment, and the penalty for breaking the commitment). One can see that in general the relative frequency of these proposers decreases as the noise level increases: The higher the noise the more likely a commitment is broken unintentionally. The scenarios PP and NP differ from the other two inasmuch as proposers can only sustain themselves in the latter for lower noise levels ($\alpha < 10^{-1}$). The $(P, C, AllD)$ strategies are better off in all four scenarios, dominating in the PP and NP scenarios even for high noise levels and low benefit-to-cost ratios, which correspond to more severe social dilemmas. As in the PN and NN scenarios proposers can only sustain themselves for lower noise levels, commitment proposing strategies in IPD seem to be more successful if they are capable to actively take revenge by withholding the benefit (through defection) against individuals breaking commitments before the end of the game.

Commitment proposing strategies survive in combination with different types of accepting strategies (Supplementary Table 1). These accepting strategies are reminiscent of the FREE, *i.e.* the accepting strategies that cooperate when an agreement is established, and FAKE, *i.e.* the accepting strategies that defect in the commitment, types analysed before in the context of the one-shot PD and public goods games [15, 16]. Interestingly, revenge also plays an important role here as, under highly erroneous conditions, these accepting strategies will also prefer to withhold the benefits from the proposer when the agreement ends. Note that the same transition from TFT to AllD, for higher noise levels, can be observed when no commitments are possible in the IPD (see Supplementary Figure 2).

Figure 1b confirms the earlier observation that proposing strategies are more successful when they can actively take revenge: if players can choose how to act outside of the commitment instead of having it imposed externally (in other words, each individual decides which one of the four scenarios to use in its strategy), the best strategies are those that defect (or play TFT in a few cases) after the commitment is broken. Hence, not playing when there is no agreement or when it is broken (NN) or only when the agreement is broken (PN) is less viable as a strategy than continuing to play in all situations (PP) or only refusing to play when no agreement can be established (NP). Interestingly, the possibility of proposing prior commitments changes the nature of the repeated game as it induces the emergence of revenge or retaliation rather than reciprocity or avoiding to interact (corresponding to the PN and NN scenarios) once commitment is broken [2, 14]. This result is in contradiction to what happens when one does not have the option of proposing commitments in the IPD, where TFT is the most important strategy for *low levels of noise* (Supplementary Figure 2). As such, commitments reduce the advantage of TFT in comparison to AllD, altering the game and resulting in the situation where AllD becomes more viable than TFT.

One could hypothesise that revenge may lead to a lower level of cooperation since proposers end up defecting when they are not in a commitment. Nevertheless, Figure 2 reveals that the presence of these retaliating commitment proposing strategies (in each of the four scenarios) increases the level of cooperation: When comparing the black line to the coloured lines in that figure, cooperation increases (Figure 2a) and defection decreases (Figure 2b), yet this decrease hides that certain scenarios suffer from an increase in games not being played (Figure 2c). For instance, although the level of defection in NN seems lower than in PP, one needs to take into account that not playing could be considered an alternative form of defection. Hence, when combining defection and not playing, the PP scenario has the highest level of cooperation (≈ 0.6) and the lowest level of defection (≈ 0.4), making it the best approach to induce cooperation in a population (see also Figure 1b).

Forgiveness requires a sincere apology to ensure cooperation

Introducing apology and forgiveness requires us to extend the strategy S_i with at least one additional parameter representing apology and forgiveness (for a more elaborate model description see Methods). Under the assumption that forgiveness occurs if and only if an apology took place, the apology parameter q_{apo} determines whether a player apologises after defecting, paying a compensation amount γ to the other player. With this definition, the strategy of a player i is now extended to $S_i = (S_c, S_{in}, S_{out}, q_{apo})_i$. We assume a strategy apologises when $q_{apo} = 1$ and does not when $q_{apo} = 0$.

We focus here on costly apology ($\gamma > 0$) and how it induces forgiveness as costless apology ($\gamma = 0$), being equivalent to forgiveness without apology, does not substantially change the conditions under which proposers are better than

pure defectors: Forgivers only do better when the benefit-to-cost ratio is high enough (see Supplementary Figures 6 and 7).

Figure 3 shows that when the compensation (γ) given upon apology is bigger than or equal to the cost of cooperating ($\gamma \gtrsim c$), proposers that cooperate during commitments, apologise when they defect by mistake and forgive when receiving an equivalent apology become the best strategists in the IPD, in all scenarios (see also Supplementary Information). They reach a maximum when $c < \gamma < \delta$ and continue dominating the population until γ becomes too high ($\gamma \approx 7$ for the PP scenario and 0.1 noise, and even higher for other scenarios, but with similar patterns; see Supplementary Information), leading to the situation where revenge,*i.e.* $(P, C, AllD, q_{apo} = 0)$, becomes once again the better choice. However, when the cost of apology is not high enough (lower than c), fake proposers and acceptors,*i.e.* $(P, D, AllD, q_{apo} = 1)$ and $(A, D, AllD, q_{apo} = 1)$, take over. These fake proposers and acceptors systematically exploit the apology-forgiveness mechanism, leading to the decrease of cooperation. Hence our results show that apology needs to be sufficiently sincere, meaning not too low, not too high ($\delta > \gamma > c$), in order for forgiveness to function properly, which intuitively makes a lot of sense. Actually, one can show that the cooperative proposer is a dominant strategy against the defecting proposer when $\gamma > c$ and against the defecting acceptor when $\gamma > c + 3\epsilon/4$ in the absence of noise and if all of them apologise (see Methods). According to Figure 3, reducing the noise affects the importance of the apologising strategy ($P, C, AllD, q_{apo} = 1$) relative to the defecting and non-apologising strategy ($P, C, AllD, q_{apo} = 0$), yet the patterns described above remain valid.

Under the assumption of high noise levels ($\alpha \approx 0.1$), which reduces the level of cooperation (see Figure 2), we can now see in Figure 4 that apology plus forgiveness seriously boost the level of cooperation and reduce defection when $\gamma > c$. Yet if the apology is not sincere enough ($\gamma < c$) one can observe the opposite behavior, even in the PP scenario case. Introducing noise in the apology and forgiveness decisions does not generate qualitative differences in these conclusions. In Supplementary Information we analyse the influence of the average number or rounds (see Supplementary Figures 9-11), showing that in repeated interactions commitments become increasingly beneficial especially when after commitment is broken one takes apology and revenge into account.

Evolution selects sincerity in apology and forgiveness

Clearly, the decision of how strongly to apologise or when to accept an apology are personal choices. As such, individuals can apologise at different costs γ and can forgive defection conditional on a personal threshold τ_γ. Therefore, if the strategy is forgiving, the parameter τ_γ is used to decide whether the player will forgive the opponent or not.

Limiting here the analysis only to strategies that defect when they are not committing we determine which threshold and apology values evolve under natural selection. Reducing the number of strategies to these ones does not reduce the generality of the results, since they are the dominant ones (that always accumulate in almost 100% fraction of the population) as we have shown before.

Figure 5 reveals which thresholds (τ_γ) are preferred and which apologies (γ) are required for the PP scenario (additional results show almost the same results are obtained for the other scenarios). First one can observe that expecting a higher apology than one actually offered ($\tau_\gamma > \gamma$) is always a bad strategy: in all the situations visualised in the figure, this situation leads to loss of cooperative commitment proposers, and hence cooperation in general. As was learned too from the results in Figure 3 and Figure 4, it is still not a good strategy to pay too high cost to apologise, as this behavior tends to disappear from the population. We see that the dominating strategies have apology values (γ) in the same region as the ones shown in Figure 3. In Figure 5 one can also observe that the higher the noise the more strategies converge to concrete values of γ and τ_γ, in other words, the more important is the apology-forgiveness mechanism due to a higher number of mistakes, as also shown in Figure 3. Yet, one can also observe in Figure 5 that apology and forgiveness are less important in more severe games (*i.e.* very low benefit-to-cost ratios).

Nevertheless, results show that even in the case of individual choices, apology and forgiveness provide an important mechanism ensuring that commitments can remain stable and both parties can continue to profit from their original agreement.

Discussion

Creating agreements and asking others to commit to such agreements provides a basic behavioral mechanism that is present at all the levels of society, playing a key role in social interactions [2, 10, 14]. Although it was shown that this behavior is evolutionary viable, little analytical and numerical insight is available on how to handle agreements and commitments in repeated interactions. The results discussed in this work fill this gap by clarifying and extending the observations made in experiments like [12, 13], while also showing that, similar to the one-shot interaction scenario, the introduction of ongoing subsisting commitments leads to higher levels of cooperation whenever the cost is sufficiently small and the compensation is high enough. Our work reveals how, when moving to repeated games, the detrimental effect of having a large arrangement cost is moderated as a subsisting commitment can play its role for several interactions. In these scenarios, the most successful individuals are those that propose commitments (and are willing to pay their cost) and, following the agreement, cooperate unless a mistake occurs. But if the commitment is broken then these individuals take revenge and defect in the remaining interactions, confirming analytically what has been argued in [23, 24]. This result is intriguing as revenge by withholding the benefit from the transgressor may lead to a

more favorable outcome for cooperative behavior in the IPD as opposed to the well-known reciprocal behavior such as TFT-like strategies.

Yet, as mistakes during any (long-term) relationship are practically inevitable, individuals need to decide whether it is worthwhile to end the agreement and collect the compensation when a mistake is made or whether it is better to forgive the co-player and continue the mutually beneficial agreement. To study this question the commitment model was extended with an apology-forgiveness mechanism, where apology was defined either as an external or individual parameter in the model. In both cases, we have shown that forgiveness is effective if it takes place after receiving an apology from the co-players. However, to play a promoting role for cooperation, apology needs to be sincere, in other words, the amount offered in the apology has to be high enough (yet not too high), which is also corroborated by a recent experimental psychology paper [40]. This extension to the commitment model produces even higher cooperation levels than in the revenge-based outcome. In the opposite case, fake committers that propose or accept to commit with the intention to take advantage of the system (defecting and apologising continuously) will dominate the population. In this situation, the introduction of the apology-forgiveness mechanism destroys the increase of the cooperation level that commitments by themselves produce. Hence there is a lower-limit on how sincere apology needs be as below this limit apology and forgiveness even reduce the level of cooperation one could expect from simply taking revenge. It has been shown in previous works that mistakes can even induce the outbreak of cheating or intolerant behavior in society [41, 42], and only a strict ethics can prevent them [42], which in our case would be understood as forgiving just when apology is sincere.

Commitments in repeated interaction settings may take the form of loyalty [17, 43], which is different from our commitments regarding posterior compensations, which do not assume a partner choice mechanism. Loyalty commitment is based on the idea that individuals tend to stay with or select partners based on the length of their prior interactions. We go beyond these works by showing that, even without partner choice, commitment can foster cooperation and long-term relationships especially when accompanied with a sincere apology and forgiveness whenever mistakes are made.

A substantial body of economic experiments on commitments, apology and forgiveness, have been carried out, and the results from this work are in close accordance with the outcomes of those experiments [14, 26, 29, 31]. In [14], a PGG experiment shows that when commitment is arranged in advance, and set up afterwards, high levels of cooperation are observed. But if the commitment fails to form (i.e. some participants do not agree to commit), the players act significantly less cooperative than when they had no opportunity to join a commitment. This outcome is similar to the emergence of AllD strategy whenever commitment is not formed or when it is formed but then broken in our system. Next, several economic experiments show that apology only promotes cooperation when it is sincere,*i.e.* costly enough [26, 29, 31]. Ohtsubo's experiment [31] shows that a costlier apology is better at communicating sincerity, and as a consequence will be more often forgiven. This observation is shown to be valid across cultures [29]. In another laboratory experiment [26], the authors showed apologies work because they can help reveal the intention behind the wrongdoers preceding offence. In compliance with this observation, in our model, an apology is mostly made by those who intended to cooperate but defect by mistake.

In conclusion, our results demonstrate that even when "to err is human" [44], behaviors like revenge and forgiveness can evolve to cope with mistakes, even when they occur at high rates. On the other hand, mistakes are not necessarily intentional and even when they are it might still be worthwhile to continue a mutually beneficial agreement. Yet, as shown in this work, a sincerity threshold exists where the cost of apologising should exceed that of cooperation to induce the latter.

Methods

A. Payoffs under commitments

Payoffs introduced in the manuscript depend on the concrete strategies that players i and j decide to choose. A commitment is set up only if both players are proposers and as such, both share the cost of establishing it ($w_\epsilon^{ij} = w_\epsilon^{ji} = -\epsilon/2$), or only one of the players (i) is a proposer and the other is an acceptor (j) and then only the first one has to pay that cost ($w_\epsilon^{ij} = -\epsilon$ and $w_\epsilon^{ji} = 0$). Denote R_C^{ij} the number of rounds the players are, on average, in the commitment. Hence, R_C^{ij} is a function of the probability that the commitment is not broken in the next round, denoted by Ω_{ij}, and the probability that the IPD game continues for another round ω, which can be written as follows:

$$R_C^{ij} = \frac{1}{1 - \Omega_{ij}\,\omega}. \tag{2}$$

We denote by $\mathbf{p}_{\alpha,ij}$ the vector that represents the probability that players i and j actually play CC, CD, DC, and DD, respectively, in a round. The probability that the commitment continues once both players choose their actions depends on the apology-forgiveness mechanism and is represented by the vector $\mathbf{q}_{c,ij} = (1, q_{ij}, q_{ji}, q_{ij}q_{ji})$. Then

$$\Omega_{ij} = \mathbf{p}_{\alpha,ij} \cdot \mathbf{q}_{c,ij}. \tag{3}$$

During the commitment, the i-player obtains a payoff per round

$$w_C^{ij} = \frac{h_{ij}}{\Omega_{ij}}, \tag{4}$$

$$h_{ij} = \sum_k p^k_{\alpha,ij} q^k_{c,ij} \left(g^k + g^k_\gamma\right). \tag{5}$$

except in the last round, where she receives w^{ij}_{last}. We have represented $\mathbf{g} = (b-c, -c, b, 0)$ as the vector that contains the payoffs coming directly from the IPD payoff matrix that the first player obtain in states (CC, CD, DC, DD). The vector $\mathbf{g}_\gamma = (0, \gamma, -\gamma, 0)$ stands for the payoffs linked to the apologies needed to maintain the commitment when any player defects. The payoff received in the last round can be computed as

$$w^{ij}_{last} = \frac{h^{last}_{ij}}{1 - \Omega_{ij}\,\omega}, \tag{6}$$

$$h^{last}_{ij} = h_{ij}(1-\omega) + \overline{h}_{ij}, \tag{7}$$

where $\overline{h}_{ij}$ denotes the payoff that the i-strategist obtains if the commitment is broken:

$$\overline{h}_{ij} = \sum_k p^k_{\alpha,ij} q'^k_{c,ij} \left(g^k + g^k_\delta\right), \tag{8}$$

$$\mathbf{q'}_{c,ij} = (0, 1-q_{ij}, 1-q_{ji}, q_{ji}-q_{ij}), \tag{9}$$

$$\mathbf{g}_\delta = (0, \delta, -\delta, \delta-\gamma). \tag{10}$$

Note that the last element of $\mathbf{q'}_{c,ij}$ and $\mathbf{g}_\delta$ vectors takes into account whether one or only one of the players forgives a mutual defective behaviour in the commitment.

Vector $\mathbf{p}_{\alpha,ij}$ depends on the strategies S^i_{in} and S^j_{in}, as well as on the noise, so that functions Ω_{ij} and h_{ij}, and payoff w^{ij}_{last} depend on them as well. Four different scenarios can be described as a function of these strategies:

- Both players intend to cooperate when they commit $S^i_{in} = S^j_{in} = C$:

$$\mathbf{p}_{\alpha,ij} = \left((1-\alpha)^2, \alpha(1-\alpha), \alpha(1-\alpha), \alpha^2\right), \tag{11}$$

$$\Omega_{ij} = (1-\alpha)^2 + \alpha(1-\alpha)(q_{ij}+q_{ji}) + \alpha^2\, q_{ij}\, q_{ji}, \tag{12}$$

$$h_{ij} = (b-c)(1-\alpha)^2 + \left[(b-\gamma)\, q_{ji} + (\gamma-c)\, q_{ij}\right]\alpha(1-\alpha), \tag{13}$$

$$h_{ji} = (b-c)(1-\alpha)^2 + \left[(\gamma-c)\, q_{ji} + (b-\gamma)\, q_{ij}\right]\alpha(1-\alpha), \tag{14}$$

$$h^{last}_{ij} = (b-c)(1-\alpha) - (\delta-\gamma)\,\alpha\,(q_{ij}-q_{ji}) - \omega\, h_{ij}, \tag{15}$$

$$h^{last}_{ji} = (b-c)(1-\alpha) + (\delta-\gamma)\,\alpha\,(q_{ij}-q_{ji}) - \omega\, h_{ji}. \tag{16}$$

- Both players intend to defect in a commitment $S^i_{in} = S^j_{in} = D$:

$$\mathbf{p}_{\alpha,ij} = \left(\alpha^2, \alpha(1-\alpha), \alpha(1-\alpha), (1-\alpha)^2\right), \tag{17}$$

$$\Omega_{ij} = (1-\alpha)^2\, q_{ij}\, q_{ji} + \alpha(q_{ij}+q_{ji}) + \alpha^2, \tag{18}$$

$$h_{ij} = (b-c)\,\alpha^2 + \left[(b-\gamma)\, q_{ji} + (\gamma-c)\, q_{ij}\right]\alpha(1-\alpha), \tag{19}$$

$$h_{ji} = (b-c)\,\alpha^2 + \left[(\gamma-c)\, q_{ji} + (b-\gamma)\, q_{ij}\right]\alpha(1-\alpha), \tag{20}$$

$$h^{last}_{ij} = (b-c)\,\alpha - (\delta-\gamma)(1-\alpha)(q_{ij}-q_{ji}) - \omega\, h_{ij}, \tag{21}$$

$$h^{last}_{ji} = (b-c)\,\alpha + (\delta-\gamma)(1-\alpha)(q_{ij}-q_{ji}) - \omega\, h_{ji}. \tag{22}$$

- Player i intends to cooperate and her co-player j intends to defect $S_{in}^i = C$ and $S_{in}^j = D$:

$$\mathbf{p}_{\alpha,ij} = \left(\alpha(1-\alpha), (1-\alpha)^2, \alpha^2, \alpha(1-\alpha)\right), \tag{23}$$

$$\Omega_{ij} = (1-\alpha)^2 q_{ij} + \alpha(1-\alpha)(1+q_{ij}q_{ji}) + \alpha^2 q_{ji}, \tag{24}$$

$$h_{ij} = (b-c)\,\alpha(1-\alpha) + (b-\gamma)\,\alpha^2 q_{ji} + (\gamma-c)(1-\alpha)^2 q_{ij}, \tag{25}$$

$$h_{ji} = (b-c)\,\alpha(1-\alpha) + (\gamma-c)\,\alpha^2 q_{ji} + (b-\gamma)(1-\alpha)^2 q_{ij}, \tag{26}$$

$$h_{ij}^{last} = b\,\alpha - c(1-\alpha) + (1-2\alpha)\delta - [(1-\alpha)q_{ij} - \alpha\, q_{ji}]\,(\delta-\gamma) - \omega\, h_{ij}, \tag{27}$$

$$h_{ji}^{last} = -c\,\alpha + b(1-\alpha) - (1-2\alpha)\delta + [(1-\alpha)q_{ij} - \alpha\, q_{ji}]\,(\delta-\gamma) - \omega\, h_{ji}. \tag{28}$$

- Player i intends to defect and player j intends to cooperate $S_{in}^i = D$ and $S_{in}^j = C$. This case is equivalent to switch i and j indices in the previous case.

Since commitments last as far as nobody defects, $w_C^{ij} = b - c$ in the absence of any apology-forgiveness mechanism.

Payoffs without commitments

When individuals play their reactive strategies S_{out}, payoffs can be computed using the method described by [19]. In each round of this game there are four possible states (CC, CD, DC, DD) depending on the actions of player i and j. Taking into account that the action of a player in the current round is given by the action of the co-player in the previous one, the process can be described as a Markov chain in the state space. The stochastic matrix $\mathbf{Q}$ that represents the transition probabilities is given by

$$\mathbf{Q} = \begin{pmatrix} p_{Ci}p_{Cj} & p_{Ci}(1-p_{Cj}) & (1-p_{Ci})p_{Cj} & (1-p_{Ci})(1-p_{Cj}) \\ p_{Di}p_{Cj} & p_{Di}(1-p_{Cj}) & (1-p_{Di})p_{Cj} & (1-p_{Di})(1-p_{Cj}) \\ p_{Ci}p_{Dj} & p_{Di}(1-p_{Dj}) & (1-p_{Di})p_{Dj} & (1-p_{Ci})(1-p_{Dj}) \\ p_{Di}p_{Dj} & p_{Di}(1-p_{Dj}) & (1-p_{Di})p_{Dj} & (1-p_{Di})(1-p_{Dj}) \end{pmatrix}. \tag{29}$$

The initial probabilities for the four states are given by the vector

$$\mathbf{P_0} = (p_{0i}\,p_{0j},\, p_{0i}(1-p_{0j}),\, (1-p_{0i})p_{0j},\, (1-p_{0i})(1-p_{0j}))\,. \tag{30}$$

Then the total payoff that a i-strategist obtains playing with a j-strategist in the lack of commitments is

$$W_{out}^{ij} = \mathbf{g}\cdot\mathbf{P_0}\,(\mathbf{I} - \omega\mathbf{Q})^{-1} \tag{31}$$

where $\mathbf{I}$ is the identity matrix of size 4.

Evolutionary dynamics

We have chosen a discrete imitation dynamic in a population of N individuals [45, 46]. According to this dynamics, two individuals are selected at random from the population. The probability that the first individual adopts the strategy of the second one is given by a Fermi imitation probability function $\left(1+e^{-\beta\,\triangle\Pi}\right)^{-1}$ [47, 48]. The parameter β represents the intensity of selection, *i.e.* the strength individuals base their decision to imitate the others, and $\triangle\Pi$ is the difference of payoffs between both individuals. We have chosen $\beta = 0.1$ for all the calculations showed here. Note that the payoff is a measure of the success of individuals and therefore the higher the payoff the higher the probability of being imitated by others [18, 19].

A discrete dynamics like the one we are considering here always leads to an asymptotically homogeneous population. Since only mutations (invasions) can introduce new strategies, a homogeneous population is always an absorbing state. We calculate the probabilities of the different invasions as fixation probabilities,*i.e.* the probability that a single invader will eventually be imitated by all the rest of individuals, who play the resident strategy, and this under the assumption of the small mutation limit [49]. Note that due to its complexity we do not consider the possibility of mixed equilibria, like in other previous works [50]. This fixation probability is given by [19, 51]

$$\rho_{ji} = \left(1 + \sum_{m=1}^{N-1}\prod_{k=1}^{m}\frac{T^-(k)}{T^+(k)}\right)^{-1}, \tag{32}$$

where $T^+(k)$ is the probability that an individual of the resident strategy i imitates a mutant one j and $T^-(k)$ is the probability that an individual of the mutant strategy imitates a resident one in a population of k individuals playing the resident strategy. These probabilities are obtained from the imitation probability defined previously:

$$T^{\pm}(k) = \frac{k(N-k)}{N^2}\left(1+e^{\mp\beta[\Pi_i(k)-\Pi_j(k)]}\right)^{-1}, \tag{33}$$

where $\Pi_i(k)$ and $\Pi_j(k)$ denote the average payoffs of the focal player and her opponent:

$$\Pi_i(k) = \frac{(k-1)\overline{W}_{ii} + (N-k)\overline{W}_{ij}}{N-1}, \tag{34}$$

$$\Pi_j(k) = \frac{k\overline{W}_{ji} + (N-k-1)\overline{W}_{jj}}{N-1}. \tag{35}$$

The probabilities defined by equation (32) determine a transition matrix of a Markov chain among strategies, assuming a sufficiently low mutation rate [49]. The normalized eigenvector associated with the eigenvalue 1 of that matrix provides the stationary distribution of strategies [46, 52], that represents the relative time the population spends adopting each of the strategies.

Dominant strategies

One strategy A is risk-dominant against another one B [22, 53, 54] when $\pi_{A,A} + \pi_{A,B} > \pi_{B,B} + \pi_{B,A}$, where $\pi_{i,j}$ is the payoff that an individual playing the i-strategy obtains when playing against another individual that plays the j-strategy. When the apology-forgiveness mechanism is introduced, these payoffs for the cooperating proposer (PC), defecting proposer (PD), and defecting acceptor (AD), in the absence of noise, are, respectively: $\pi_{PC,PC} = -\epsilon/2+b-c$, $\pi_{PD,PD} = -\epsilon/2$, $\pi_{AD,AD} = 0$, $\pi_{PC,PD} = -\epsilon/2+\gamma-c$, $\pi_{PD,PC} = -\epsilon/2+\gamma+b$, $\pi_{PC,AD} = -\epsilon+\gamma-c$, and $\pi_{AD,PC} = -\gamma+b$. Then the cooperative proposer is risk-dominant against the defective proposer when $\gamma > c$ and against the defective acceptor when $\gamma > c+3\epsilon/4$ in the absence of noise and if all of them apologise when making a mistake.

Acknowledgments

This work was supported by the grant FRFC nr. 2.4614.12 from the *Fondation de la Recherche Scientifique* - FNRS and the grant nr. G.0391.13N provided by *Fonds voor Wetenschappelijk Onderzoek* - FWO.

Author contributions

LMV, TAH, LMP and TL designed the research. The models were implemented by LMV. Results were analysed and improved by LMV, TAH, LMP and TL. LMV, TAH, LMP and TL wrote the paper together.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

[1] Frank, R. H. Cooperation through Emotional Commitment. In Nesse, R. M. (ed.) *Evolution and the capacity for commitment*, 55–76 (New York: Russell Sage, 2001).
[2] Nesse, R. M. *Evolution and the capacity for commitment.* Russell Sage Foundation series on trust (Russell Sage, 2001).
[3] Leeds, B. A. Alliance Reliability in Times of War: Explaining State Decisions to Violate Treaties. *Int. Organ.* **57**, 801–827 (2003).
[4] Yamagishi, T. The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* **51**, 110 (1986).
[5] Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Amer. Econ. Rev.* **90**, 980–994 (2000).
[6] Sigmund, K., Hauert, C. & Nowak, M. A. Reward and punishment. *Proc. Nat. Acad. Sci.* **98**, 10757–10762 (2001).
[7] Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
[8] Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3531–3535 (2003).
[9] Sosis, R. Religion and intra-group cooperation: preliminary results of a comparative analysis of utopian communities. *Cross. Cult. Res.* **34**, 70–87 (2000).

[10] Sterelny, K. *The evolved apprentice* (MIT Press, 2012).
[11] Woodburn, J. Egalitarian Societies. *Man* **17**, 431–451 (1982).
[12] Chen, X.-P. & Komorita, S. S. The effects of communication and commitment in a public goods social dilemma. *Organ. Behav. Hum. Decis. Process.* **60**, 367–386 (1994).
[13] Kurzban, R., McCabe, K., Smith, V. L. & Wilson, B. J. Incremental commitment and reciprocity in a real-time public goods game. *Pers. Soc. Psychol. Bull.* **27**, 1662–1673 (2001).
[14] Cherry, T. L. & McEvoy, D. M. Enforcing compliance with environmental agreements in the absence of strong institutions: An experimental analysis. *Environ. Resource Econ.* **54**, 63–77 (2013).
[15] Han, T. A., Pereira, L. M., Santos, F. C. & Lenaerts, T. Good agreements make good friends. *Sci. Rep.* (2013).
[16] Han, T. A., Moniz Pereira, L. & Lenaerts, T. Avoiding or Restricting Defectors in Public Goods Games? *J. R. Soc. Interface* 20141203 (2014).
[17] Schneider, F. & Weber, R. A. Long-term commitment and cooperation. Tech. Rep., Working Paper Series, University of Zurich, Department of Economics (2013).
[18] Hofbauer, J. & Sigmund, K. *Evolutionary Games and Population Dynamics* (Cambridge University Press, Cambridge, 1998).
[19] Sigmund, K. *The Calculus of Selfishness* (Princeton University Press, Princeton, 2010).
[20] Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
[21] Trivers, R. L. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57 (1971).
[22] Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
[23] McCullough, M. E. *Beyond Revenge, the evolution of the forgiveness instinct* (Jossey-Bass, 2008).
[24] McCullough, M. E., Kurzban, R. & Tabak, B. A. Evolved mechanisms for revenge and forgiveness. In Shaver, P. R. & Mikulincer, M. (eds.) *Human aggression and violence: Causes, manifestations, and consequences. Herzilya series on personality and social psychology*, 221–239 (American Psychological Association, Washington, DC, US, 2011).
[25] Han, T. A., Pereira, L. M. & Santos, F. C. Intention recognition promotes the emergence of cooperation. *Adapt. Behav.* **19**, 264–279 (2011).
[26] Fischbacher, U. & Utikal, V. On the acceptance of apologies. *Game. Econ. Behav.*. **82**, 592–608 (2013).
[27] Smith, N. *I was wrong: The meanings of apologies*, vol. 8 (Cambridge University Press New York, 2008).
[28] Abeler, J., Calaki, J., Andree, K. & Basek, C. The power of apology. *Econ. Lett.* **107**, 233 – 235 (2010).
[29] Takaku, S., Weiner, B. & Ohbuchi, K. A cross-cultural examination of the effects of apology and perspective taking on forgiveness. *J. Lang. Soc. Psychol.* **20**, 144–166 (2001).
[30] Okamoto, K. & Matsumura, S. The evolution of punishment and apology: an iterated prisoner's dilemma model. *Evol. Ecol.* **14**, 703–720 (2000).
[31] Ohtsubo, Y. & Watanabe, E. Do sincere apologies need to be costly? test of a costly signaling model of apology. *Evol. and Hum. Behav.* **30**, 114–123 (2009).
[32] Liang, B. A system of medical error disclosure. *Qual. Saf. Health Care* **11**, 64–68 (2002).
[33] Petrucci, C. Apology in the criminal justice setting: Evidence for including apology as an additional component in the legal system. *Behav. Sci. Law* **20**, 337–362 (2002).
[34] Smith, N. *Justice Through Apologies: Remorse, Reform, and Punishment* (Cambridge University Press, 2014).
[35] Axelrod, R. *The Evolution of Cooperation* (Basic Books, New York, 1984).
[36] Boerlijst, M. C., Nowak, M. A. & Sigmund, K. The logic of contrition. *J. Theor. Biol.* **185**, 281 – 293 (1997).
[37] Fudenberg, D., Rand, D. G. & Dreber, A. Slow to anger and fast to forgive: Cooperation in an uncertain world *Am. Econ. Rev.* **102**, 720 – 749 (2012).
[38] Nowak, M. A. & Sigmund, K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* **364**, 56–58 (1993).
[39] Imhof, L. A., Fudenberg, D. & Nowak, M. A. Tit-for-tat or win-stay, lose-shift? *J. Theor. Biol.* **247**, 574–580 (2007).
[40] McCullough, M. E., Pedersen, E. J., Tabak, B. A. & Carter, E. C. Conciliatory gestures promote forgiveness and reduce anger in humans. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 11211–11216 (2014).
[41] Martinez-Vaquero, L. A. & Cuesta, J. A. Evolutionary stability and resistance to cheating in an indirect reciprocity model based on reputation. *Phys. Rev. E* **87**, 052810 (2013).
[42] Martinez-Vaquero, L. A. & Cuesta, J. A. Spreading of intolerance under economic stress: Results from a reputation-based model. *Phys. Rev. E* **90**, 022805 (2014).
[43] Back, I. & Flache, A. The Adaptive Rationality of Interpersonal Commitment. *Ration. Soc.* **20**, 65–83 (2008).
[44] Pope, A. *An Essay on Criticism, part II* (W. Lewis, Russel Street, Covent Garden, 1711).
[45] Nowak, M. A., Sasaki, A., Taylor, C. & Fudenberg, D. Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650 (2004).
[46] Imhof, L. A., Fudenberg, D. & Nowak, M. A. Evolutionary cycles of cooperation and defection. *Proc. Natl. Acad. Sci. USA* **102**, 10797–10800 (2005).
[47] Blume, L. Now noise matters. *Game. Econ. Behav.*. **44**, 251–271 (2003).
[48] Traulsen, A., Nowak, M. A. & Pacheco, J. M. Stochastic dynamics of invasion and fixation. *Phys. Rev. E* **74**, 011909 (2006).
[49] Wu, B., Gokhale, C. S., Wang, L. & Traulsen, A. How small are small mutation rates?. *J. Math. Biol.* **64**, 803–827 (2012).
[50] Martinez-Vaquero, L. A., Cuesta, J. A. & Sánchez, A. Generosity pays in the presence of direct reciprocity: A comprehensive study of 2×2 repeated games. *PLoS ONE* **7**, e35135 (2012).
[51] Karlin, S. & Taylor, H. M. *A First Course in Stochastic Processes* (Academic Press, New York, 1975), second edn.
[52] Fudenberg, D. & Imhof, L. A. Imitation processes with small mutations. *J. Econ. Theory* **131**, 251–262 (2006).
[53] Kandori, M., Mailath, G. J. & Roy, R. Learning, mutation and long-run equilibria in games. *Econometrica* **61**, 29–56 (1993).
[54] Gokhale, C. S. & Traulsen, A. Evolutionary games in the multiverse. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5500–5504 (2010).

Figure Captions

FIG. 1: Success of commitments and revenge after commitments break. Stationary distribution of the most dominant strategies (proposers that cooperate within the commitment) relative to the stationary distribution of the pure defectors as a function of noise for PP, NP, PN and NN scenarios separately (**a**) and together (**b**). Different lines correspond to different S_{out} . We assumed $\omega = 0.9$, $b/c = 2$, $\epsilon = 0.25$, and $\delta = 4$.

FIG. 2: Commitments increase the level of cooperation. Levels of cooperation (**a**), defection (**b**) and non-playing (**c**) for the dominant strategies (proposers that cooperate within the commitment), as a function of the noise for the different scenarios. The black lines correspond to the situation where commitments cannot be made, serving as a baseline for the other approaches. We assumed $\omega = 0.9$, $b/c = 2$, $\epsilon = 0.25$, and $\delta = 4$.

FIG. 3: Forgiveness is evolutionary viable if apology is sincere. Stationary distribution of the main strategies with respect to the stationary distribution of the pure defectors as a function of the apology cost for the PP scenario and $\alpha = 0.01$ (left) and $\alpha = 0.1$ (right). Vertical dashed lines mark the values of c and δ. We assumed $\omega = 0.9$, $b/c = 2$ (with $c = 1$), $\epsilon = 0.25$, and $\delta = 4$.

FIG. 4: Sincere apology increases the level of cooperation. Levels of cooperation (**a**), defection (**b**) and non-playing (**c**) for the main strategies (proposers that cooperate within the commitment) as a function of the apology cost for the different scenarios. Vertical dashed lines mark the values of c and δ. We assumed $\omega = 0.9$, $b/c = 2$, $\alpha = 0.1$, $\epsilon = 0.25$, and $\delta = 4$.

FIG. 5: Thresholds for conditional forgiveness. Stationary distribution of cooperative proposers as a function of the cost of their apologies γ and the threshold τ_γ they require to forgive a co-player for the PP scenario. We assumed $\omega = 0.9$, $\epsilon = 0.25$ and $\delta = 4$.

Supplementary Information.
Apology and forgiveness evolve to resolve failures in cooperative agreements.

Luis A. Martinez-Vaquero[α,β], The Anh Han[γ],
Luís Moniz Pereira[λ] and Tom Lenaerts[α,β,*]

[α] AI lab, Computer Science Department, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
[β] MLG, Département d'Informatique, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussels, Belgium
[γ] School of Computing, Teesside University, Borough Road, Middlesbrough, UK TS1 3BA
[λ] NOVA Laboratory for Computer Science and Informatics, Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

*corresponding authors: Tom.Lenaerts@ulb.ac.be

1 Commitments without apology-forgiveness mechanism

1.1 Stationary distribution of all strategies

In order to complement the Figure 1A of the Results section, we show in Supplementary Table 1 the frequencies corresponding to every strategy for the four scenarios separately for the usual values of parameters and noise 0.1 and 0.001. As one can see, strategies (P,C,AllD), (P,C,TFT) and (NC,-,AllD) are usually the main ones.

1.2 Influence of the different parameters of the model

In Results we showed the stationary distributions of the main strategies as a function of the noise for the four different scenarios. In Supplementary Figures 1–5 we complement that information showing the stationary distributions of strategies as a function of noise, b/c, ϵ, and δ for different parameters; Supplementary Figure 2 is focused on the case where commitments are not allowed. We see that an increase in the benefit-to-cost ratio leads to a decrease of the presence of pure defectors and an increase or maintenance of the level of proposers as well as the cooperation level. In PP and NP scenarios, proposers that play TFT outside the commitments benefit from this increase even more than those that defect in that situation, at least for low noise. Obviously the cheaper it becomes to set up a commitment the more successful the proposers are and vice versa for acceptors. Very low δ benefits those strategies that accept a commitment yet defect when playing the game, whereas if this value is increased, proposers that cooperate in commitments gain importance.

2 Commitments in the presence of apology-forgiveness mechanism

2.1 Costless apology

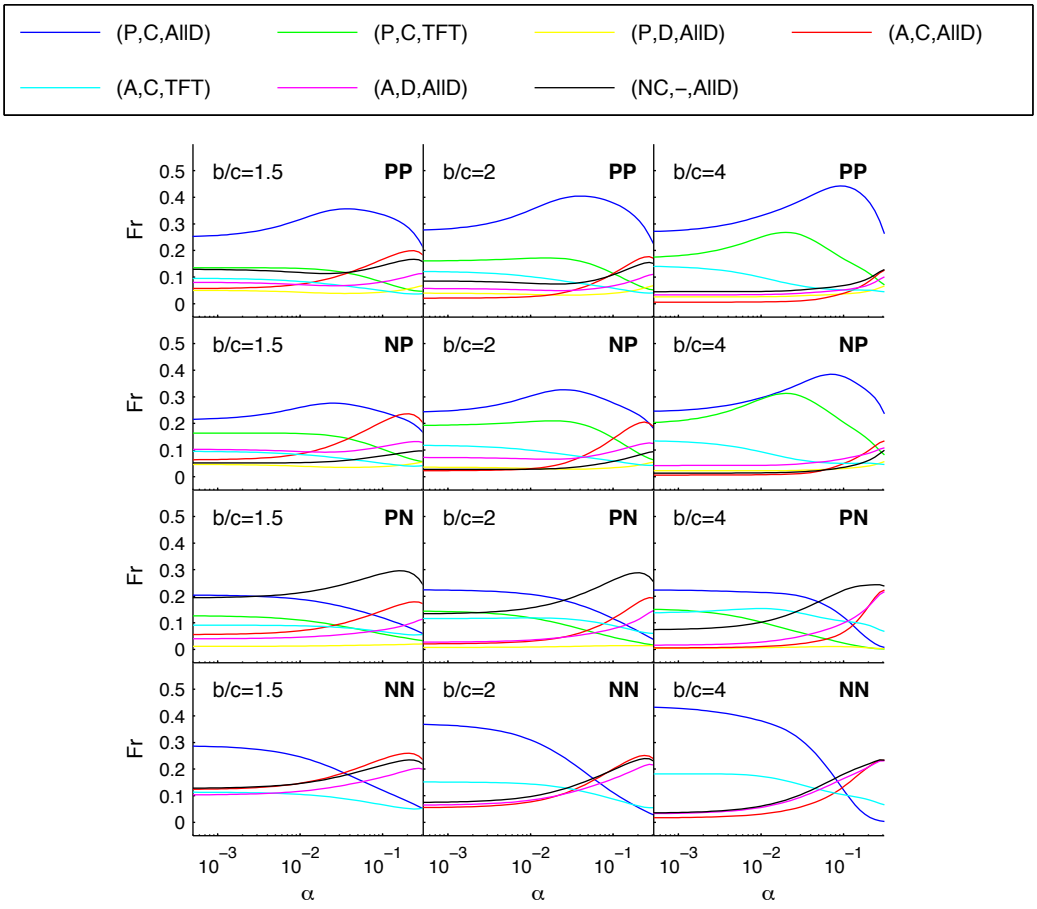
We started by analysing the effect of costless apology, which is equivalent to consider that players always apologise at no cost ($\gamma = 0$ and $q_{apo} = 1$). If we incorporate strategies that forgive in commitments with a given probability $q_{for} = q$, we see in Supplementary Figures 6 and 7 that the costless apology does not change substantially the conditions under which proposers are better than pure defectors. Forgivers do better when the benefit is high enough ($b \gtrsim 3$ for the PP scenario). In that situation, the higher the probability of forgiving the better at least for $q \lesssim 0.7$. We see the opposite behaviour for lower values of the benefit: forgiving is worse and worse when its probability is increased.

2.2 Apology-forgiveness

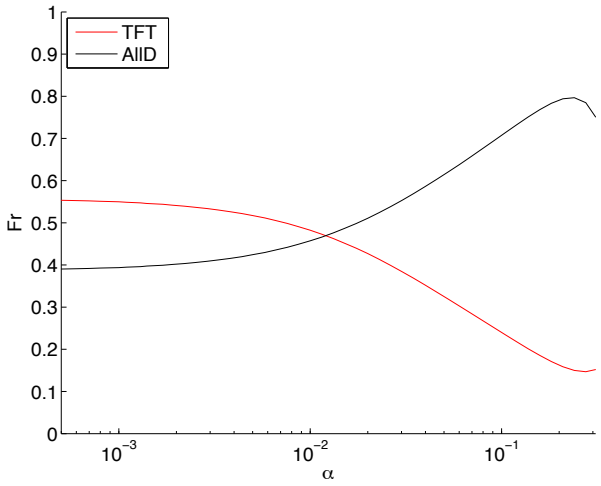
In order to complement the Figure 3 of the Results, we show in Supplementary Figure 8 the stationary distributions of the main strategies under the presence of apology-mechanism for the four scenarios. We confirm what we discussed in Results and Conclusions: the apology-forgiveness mechanism only works under sincere apologies for all the scenarios.

	$\epsilon = 10^{-3}$				$\epsilon = 10^{-1}$			
	NN	**PN**	**NP**	**PP**	**NN**	**PN**	**NP**	**PP**
(P,C,AllC)	**0.36**	0.04	0.03	0.02	**0.11**	<0.01	0.01	<0.01
(P,C,AllD)		**0.22**	**0.25**	**0.28**		**0.11**	**0.27**	**0.38**
(P,C,ATFT)		0.06	0.06	0.04		0.03	0.02	0.01
(P,C,TFT)		**0.14**	**0.19**	**0.16**		**0.04**	**0.14**	**0.11**
(P,D,AllC)	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
(P,D,AllD)		<0.01	0.04	0.04		0.01	0.03	0.04
(P,D,ATFT)		<0.01	<0.01	<0.01		<0.01	<0.01	<0.01
(P,D,TFT)		<0.01	0.02	0.02		<0.01	0.01	<0.01
(A,C,AllC)	0.01	0.01	0.01	0.01	0.01	0,02	<0.01	<0.01
(A,C,AllD)	0.06	0.02	0.02	0.02	0.20	0.12	0.15	0.12
(A,C,ATFT)	0.01	<0.01	<0.01	<0.01	0.19	0.02	0.01	<0.01
(A,C,TFT)	0.15	0.12	0.12	0.12	0.09	0.09	0.06	0.06
(A,D,AllC)	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
(A,D,AllD)	0.07	0.03	0.07	0.06	0.17	0.08	0.1	0.07
(A,D,ATFT)	<0.01	<0.01	<0.01	<0.01	0.01	0.01	<0.01	<0.01
(A,D,TFT)	0.11	0.08	0.04	0.05	0.07	0.06	0.02	0.02
(NC,-,AllC)	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.01	<0.01
(NC,-,AllD)	**0.08**	**0.13**	**0.03**	**0.08**	**0.20**	**0.26**	**0.06**	**0.11**
(NC,-,ATFT)	<0.01	<0.01	<0.01	<0.01	0.02	0.02	0.02	<0.01
(NC,-,TFT)	0.12	0.09	0.09	0.07	0.08	0.07	0.05	0.03

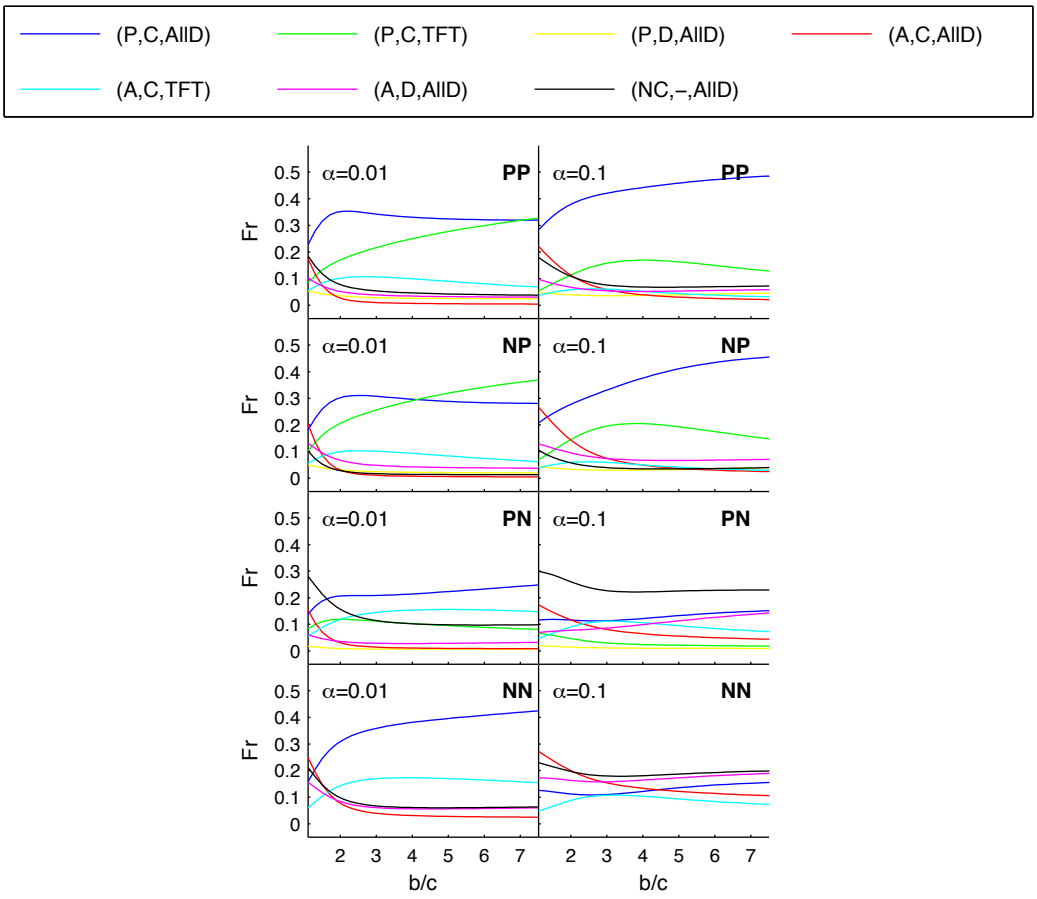
Supplementary Table 1: Stationary distribution of all the strategies for all the scenarios separately. The strategies plots in Figure 1a of the Results section are marked in bold. We assumed $\omega = 0.9$, $b/c = 2$, $\epsilon = 0.25$, and $\delta = 4$.



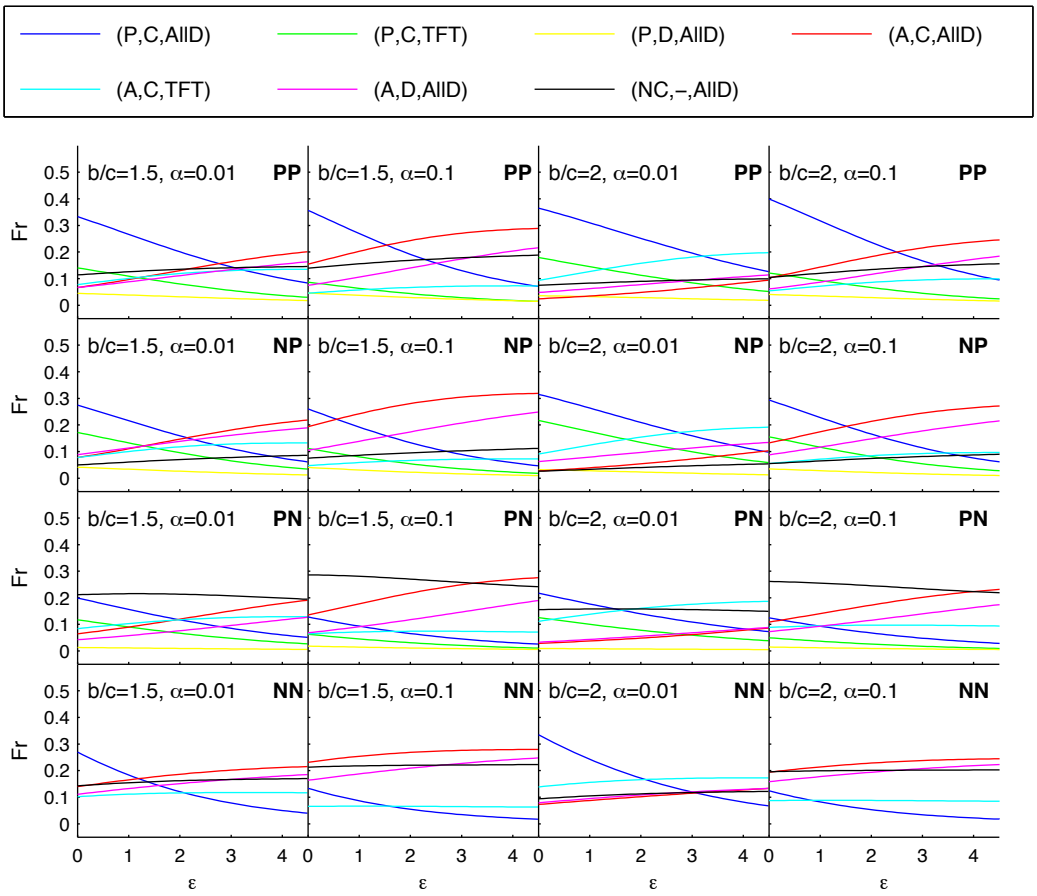
Supplementary Figure 1: Stationary distribution of the main strategies as a function of the noise. We consider different benefit-to-cost ratio and scenarios, and assume $\omega = 0.9$, $\epsilon = 0.25$, and $\delta = 4$.



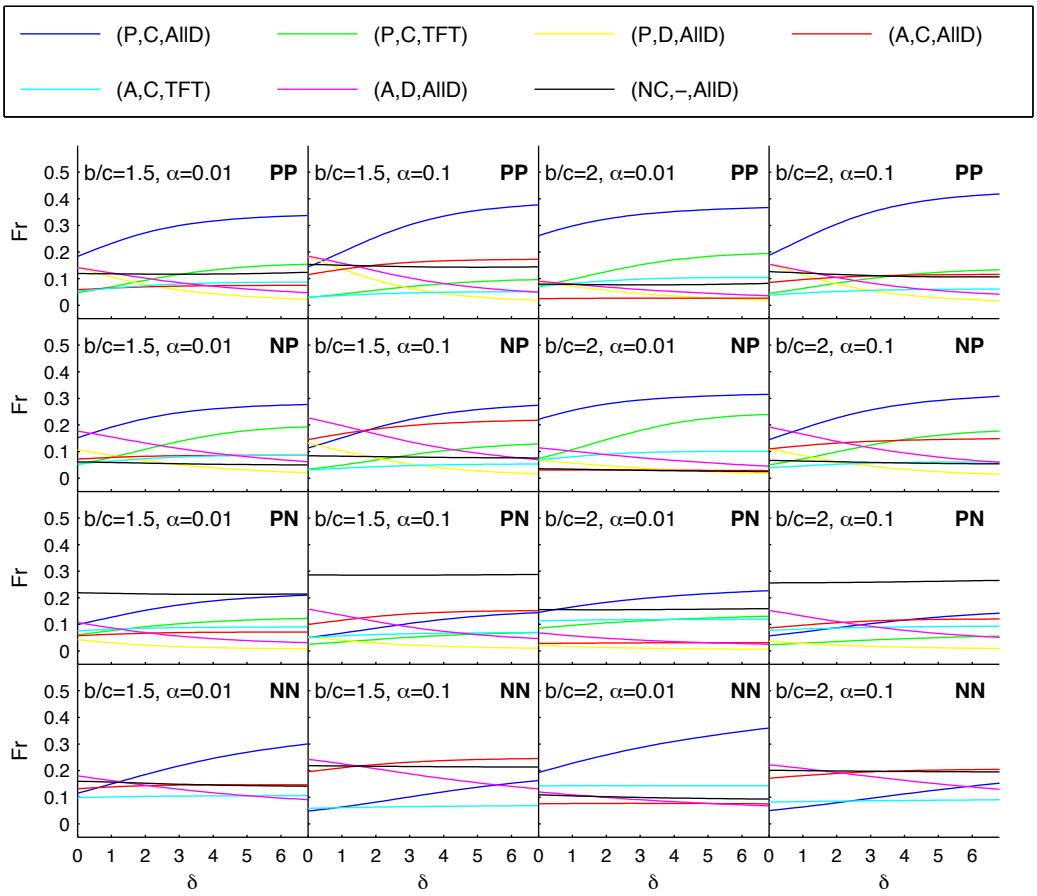
Supplementary Figure 2: Stationary distribution of TFT and AllD as a function of the noise when commitments are not allowed. We assumed $b/c = 2$ and $\omega = 0.9$.



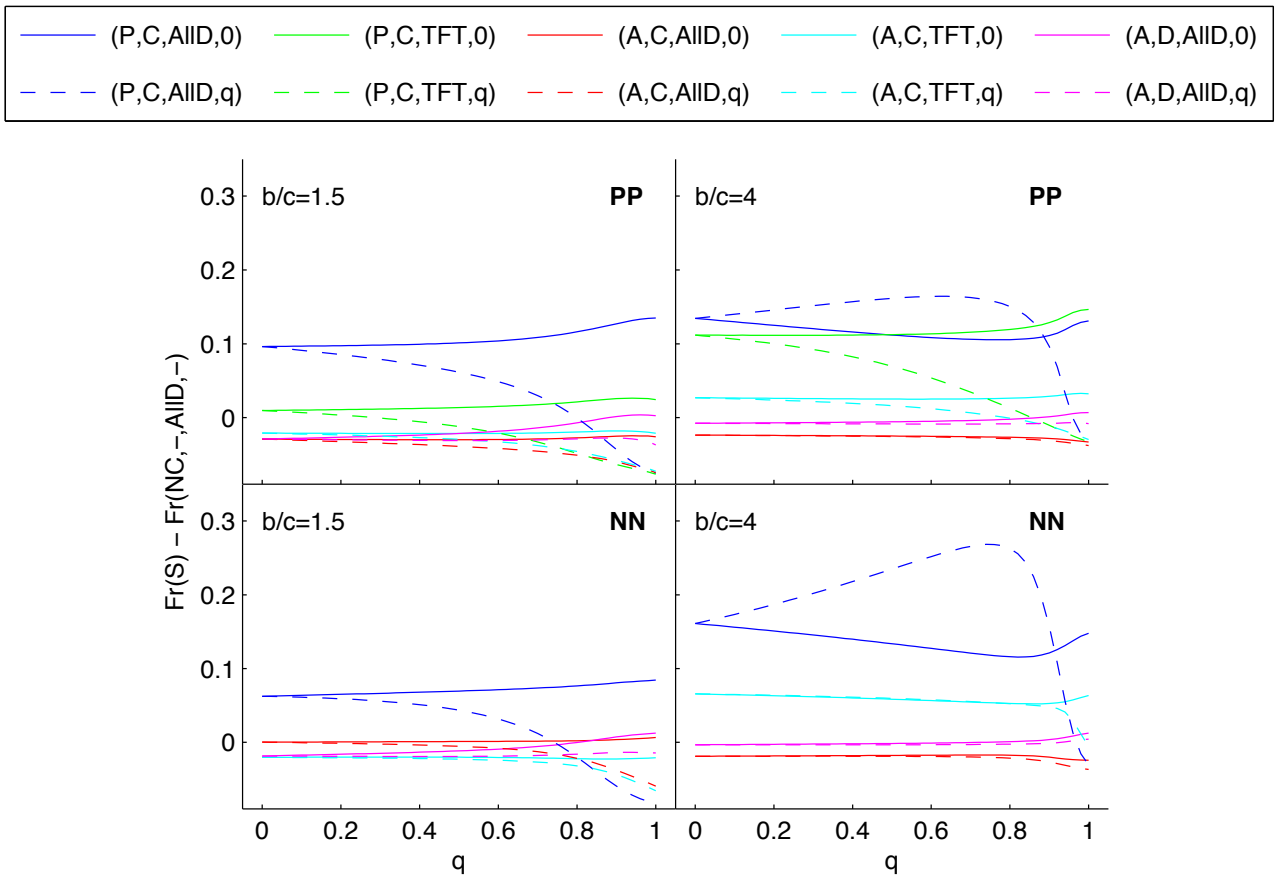
Supplementary Figure 3: Stationary distribution of the main strategies as a function of the benefit-to-cost ratio. We consider different noise and scenarios, and assume $\omega = 0.9$, $\epsilon = 0.25$, and $\delta = 4$.



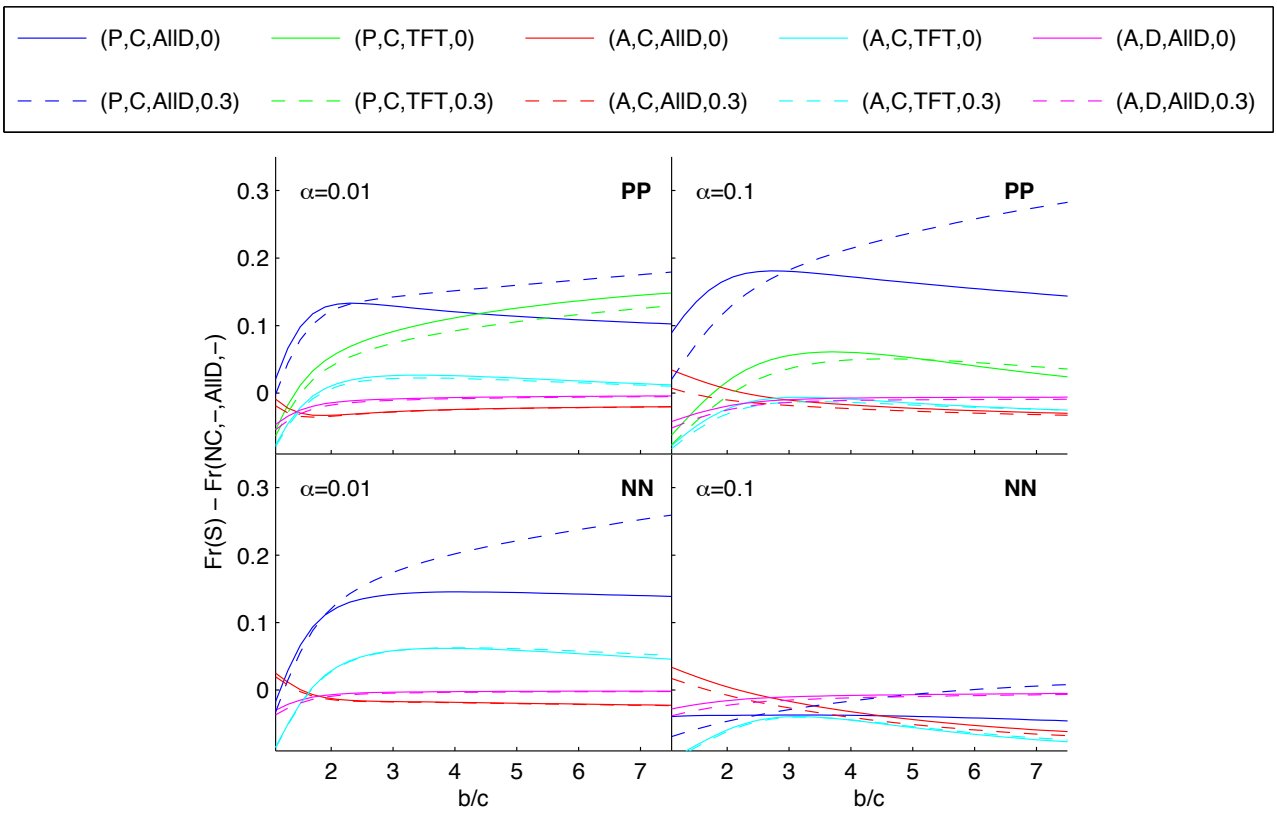
Supplementary Figure 4: Stationary distribution of the main strategies as a function of ϵ. We consider different noise, benefit-to-cost ratio, and scenarios, and assume $\omega = 0.9$ and $\delta = 4$.



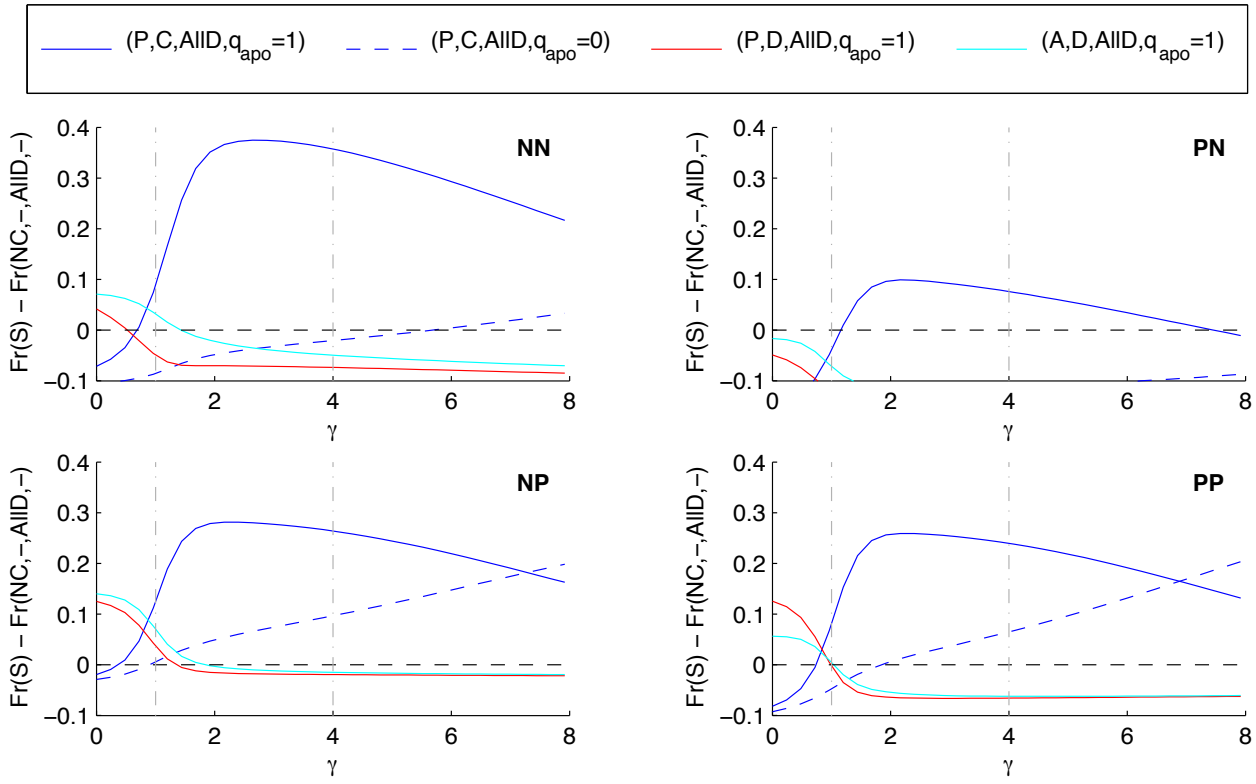
Supplementary Figure 5: Stationary distribution of the main strategies as a function of δ. We consider different noise, benefit-to-cost ratio, and scenarios, and assume $\omega = 0.9$ and $\epsilon = 0.25$.



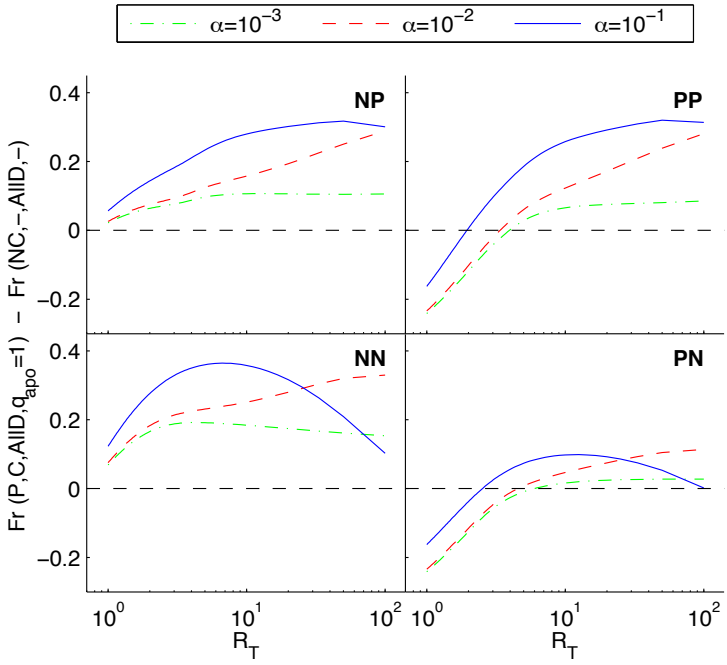
Supplementary Figure 6: Stationary distribution of the main strategies as a function of the probability of forgiveness q. We consider different benefit-to-cost ratio in the presence of costless apology for PP and NN scenarios.



Supplementary Figure 7: Stationary distribution of the main strategies as a function of the benefit-to-cost ratio. We consider different noise in the presence of costless apology for PP and NN scenarios.



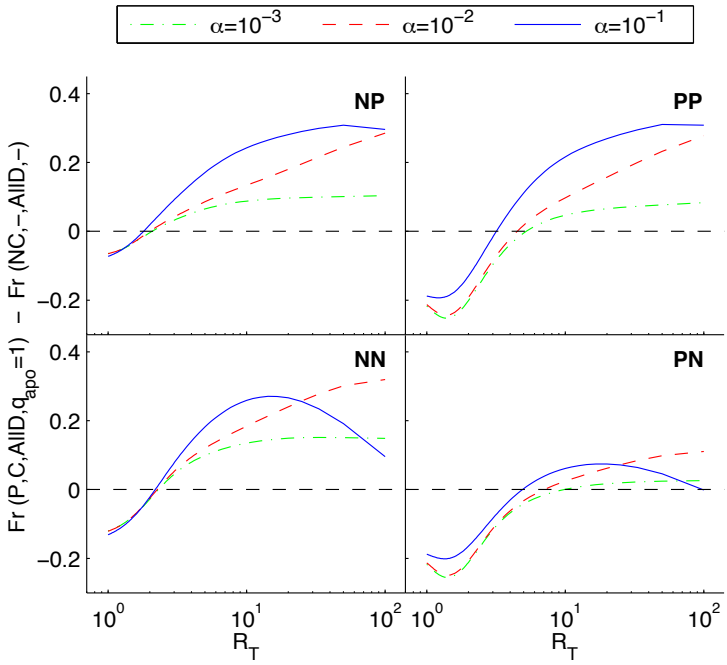
Supplementary Figure 8: Success of forgiveness for different apologies in different scenarios. Stationary distribution of the main strategies with respect to the stationary distribution of the pure defectors as a function of the apology cost for the different scenarios. Vertical dashed lines mark the values of c and δ. We assumed $\omega = 0.9$, $b/c = 2$, $\alpha = 0.1$, $\epsilon = 0.25$, and $\delta = 4$.



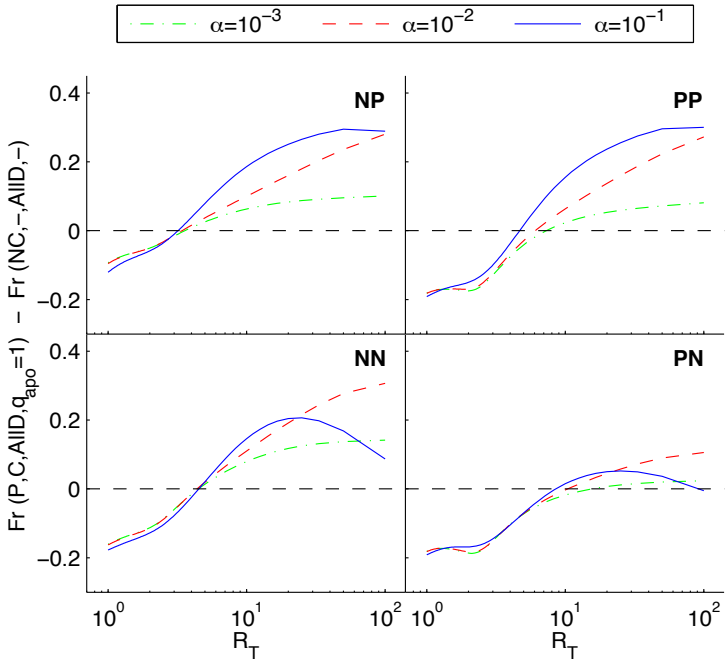
Supplementary Figure 9: Influence of the number of rounds in the success of forgiveness. Stationary distribution of the $(P, C, AllD, q_{apo} = 1)$ strategies with respect to the stationary distribution of the pure defectors as a function of the total number of rounds for the four scenarios. We assumed $b/c = 2$, $\epsilon = 0.25$, $\delta = 4$, and $\gamma = 2$.

In Supplementary Figures 9-11 we examined how the frequency of $(P, C, AllD, q_{apo} = 1)$ changes as the average number of rounds in the IPD. We can see in the NP and PP scenarios, that, when individuals can play after the commitment is broken, this frequency generally increases with R_T. When the players cannot play after the commitment is broken, *i.e.* in NN and PN scenarios, a similar observation is only seen for sufficiently low levels of noise. When noise is large, the frequency of these apologising commitment proposers drops when R_T reaches certain threshold but it remains higher than when $R_T = 1$, *i.e.* for the one-shot PD, for a wider range of R_T. That said, arranging commitments in long-term interactions is more beneficial than in the one-shot one, especially when one takes apology during and revenge after the commitment is broken into account. One of the reasons that this occurs is because the cost of arranging commitment ϵ is paid only once at the beginning of the IPD, thereby reducing its detrimental (per round) impact on a proposer for increasing R_T. This cost, if becoming too large, is highly detrimental for cooperation in the one-shot PD. This interesting observation becomes even clearer when we look at the results for varying ϵ.

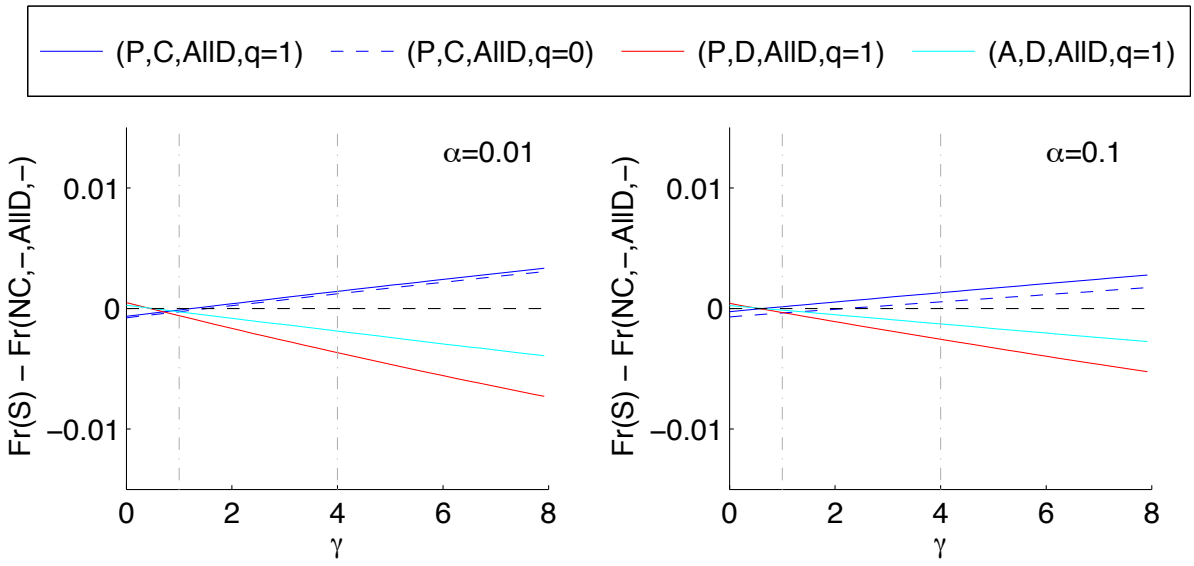
We also checked the influence of β repeating the same analysis as in Figure 3 of the Results but for $\beta = 0.001$ and $\beta = 1$ in Supplementary Figures 12 and 13, respectively. One can observe that β has no important impact on the conclusions we obtain for our study of the apology-forgiveness mechanism.



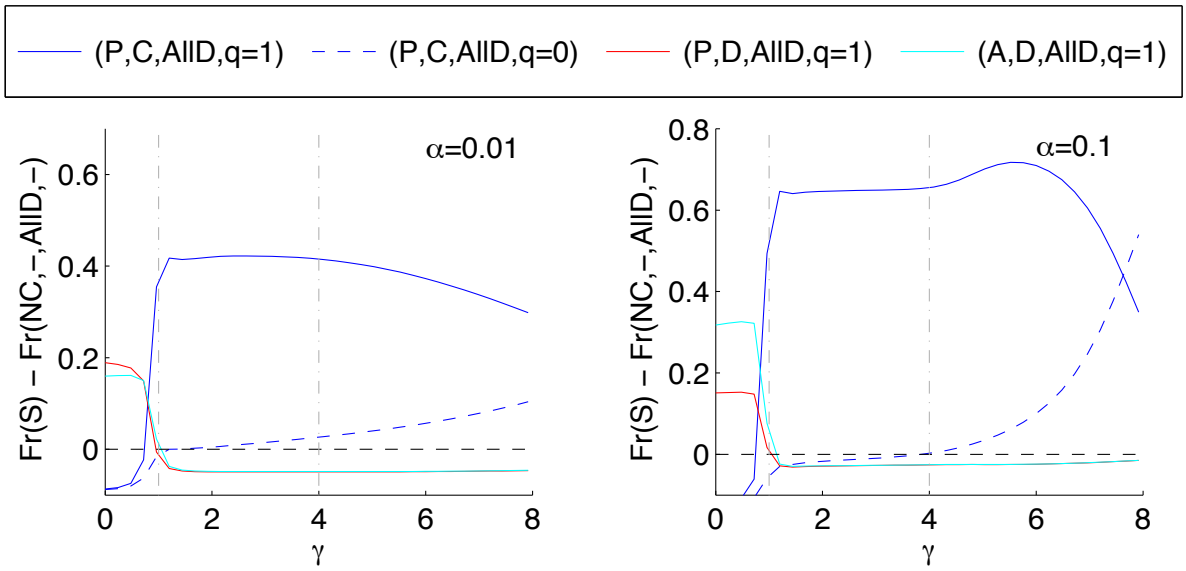
Supplementary Figure 10: Influence of the number of rounds in the success of forgiveness for $\epsilon = 1$. Same as Supplementary Figure 9 but for $\epsilon = 1$.



Supplementary Figure 11: Influence of the number of rounds in the success of forgiveness for $\epsilon = 2$. Same as Supplementary Figure 9 but for $\epsilon = 2$.



Supplementary Figure 12: Same as Figure 3 in the manuscript but for $\beta = 10^{-3}$.



Supplementary Figure 13: Same as Figure 3 in the manuscript but for $\beta = 1$.