When apology is sincere, cooperation evolves, even when mistakes occur frequently.

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

Abstract. Evolutionary psychologists have argued that revenge, apology and forgiveness are moral sentiments that humans acquired to establish and maintain long-term mutually beneficial relationships, especially since misapprehensions, intentional or not, can always occur that could lead to worse outcomes. Their argument assumes an evolutionary advantage to such emotional thinking, for which no explicit model was available. Using the iterated prisoners dilemma as context, we provided analytical and numerical results that show that these three behaviours emerge spontaneously, ensuring lasting cooperation [16]. Concretely our work revealed that apology and forgiveness are efficient even in a very noisy environment. Yet in order for apology to work, it needs to be sufficiently costly as otherwise exploiting the system by defecting and apologising is the most profitable behaviour.

1 Introduction

Commitments by individuals in social interactions are established to ensure favourable outcomes over long time periods [19]. Essentially, interaction partners are coerced to comply to certain behavioural restrictions like cooperation instead of defection within the context of a social dilemma. To ensure compliance a credible repercussion needs to be in place. Emotions are one way to ensure compliance [5]. A nice anthropological example associated with commitments is the sharing on demand among foragers [26, 21].

In our prior work we formalised commitment behaviour in the context of one-shot pairwise and n-player social dilemmas [9, 7, 8], i.e. the prisoners dilemma and the public goods game: Prior to playing the game, an individual can try to get the other to commit to cooperate. This effort is costly (ϵ) but if accepted leads to mutual cooperation, unless the opponent cheats and defects anyway. Within our formalisation we assume that when the latter happens the defector suffers a cost (δ) that benefits the one that honoured the commitment (which we call compensation). This compensation could be the result of executing a threat as well as the legal prosecution of the defecting individual. Our analytical and numerical work showed that commitment behaviour evolves and dominates in a population when the cost ϵ is smaller than the benefit of cooperation and the compensation δ is bigger than the cost of cooperation and the cost of setting up the commitment. Interestingly, the model also showed that commitments, which are established prior to the social interaction, are more effective than costly punishment [10], which acts only posteriorly.

Although commitments have shown their use in one-shot interactions, they seem to be even more relevant for long-term relationships [14], which required the expansion of our research to repeated games like the Iterated Prisoners Dilemma (IPD) [2]. Within the context of the IPD a parameter ω specifies the probability that the interaction is repeated. A second parameter α determines the likelihood that the player takes an action opposed to her decision, i.e to play D when intending to play C and vice versa. Such errors are the source of misunderstandings and might trigger the end of the commitment. In [16], we provided detailed analytical and numerical results within that context using again Evolutionary Game Theory [22] as a tool. These results are also reported in this abstract.

As commitment decisions occur over multiple rounds with the same partner in the IPD, the individual strategies are required to take into account additional issues next to the decision when to cooperate or defect: How to deal with mistakes made by the opponent or by themselves before that the interaction is terminated? To demand the compensation immediately as in the one-shot scenario or wait? When the interaction is mutually beneficial enough to continue, which mechanism should be put in place so that trust is reestablished and the interaction can continue?

As asserted by evolutionary psychologists, humans have acquired sophisticated strategies to ensure that mistakes are not repeated and that profitable relationships may continue. McCullough [17, 18] has eloquently argued that revenge for instance may have evolved exactly to cope with those situations:

> ”The threat of revenge, through some punishment or withholding of a benefit, may discourage interpersonal harm.”

A world wherein social interactions are ruled by the fear for revenge is ruled by those that can make credible threats. Such world appears to be unkind as honest mistakes would be severely punished, destroying any reason to set up mutually beneficial relationships in the first place. It is often not straightforward to identify whether or not the other's behaviour is accidental [4]. Looking again in society, we can see that we have acquired the capacity to apologise when a mistake is made and to forgive the person that makes the mistake. Forgiveness provides a restorative mechanism that, notwithstanding the initial harm, ensures that mutually beneficial relationships can continue. An essential ingredient for forgiveness to work appears to be (costly) apology [17], a point emphasised also in [23].

As we explained in [16], apology and forgiveness remove the necessity to get costly external parties (e.g. judicial institutions) involved in order to ensure cooperation. For instance, it was shown that customers prefer to continue their interaction with a company that apologises [1]. Also in case of situations of medical error, apology has been shown to lead to fewer lawsuits with lower settlements

[1] MLG, Université Libre de Bruxelles, Brussels, Belgium and AI lab, Vrije Universiteit Brussel, Brussels, Belgium. email: tlenaert@ulb.ac.be
[2] School of Computing, Teesside University, Middlesbrough, UK
[3] NOVA-LINCS,Universidade Nova de Lisboa, Caparica, Portugal
[4] Institute of Cognitive Sciences and Technologies, CNR, Rome, Italy

[15]. Apology even enters the law as an effective mechanism of resolving conflicts [24]. It therefore makes sense to ask under which conditions revenge, apology and forgiveness evolve to ensure sustainable pro-social interactions.

2 Analysis

We here present some of the technical aspects of the model. For the details see [16]. Our extension of the repeated games with commitments, revenge, apology and forgiveness required the definition of new strategies that require to address three questions:

1. Whether the player proposes (P), accepts (A) or ignores (NC) commitments, with the latter meaning that they play by refusing commitment proposals.
2. Whether she cooperates (C) or defects (D) once the commitment is established.
3. How to behave once the commitment is terminated. As there may be remaining rounds of the IPD the player needs to decide how to act. In [16], 4 options were provided, always cooperating ($AllC$), always defecting ($AllD$), playing tit-for-tat (TFT) and the inverse of tit-for-tat ($ATFT$).
4. Whether to apologise ($q = 1$) when a defection occurs and continue to behave as in 2 (within a commitment).

We consider a well-mixed population formed by N individuals. In each generation, individuals are randomly matched. They first decide if they are proposing a commitment. If one or both individuals do it and the other accepts it, the proposers pay an amount ϵ ($\epsilon/2$ if both are proposing it) and the commitment is set up. In a second stage, individuals play an Iterated Prisonner's Dilemma with payoff matrix

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

A new round is repeated with probabilty ω. If the commitment was set up and one of the individuals defects and do not apologise, the defector pays an amounf δ to the other player and the commitment is broken. The payoffs that individuals obtain are then the sum of those obtained during the commitment, outside of the commitment, and the amounts exchanged in setting up the commitment, apologising and/or compensating (see all the details in [16]). Payoffs are a measure of the success of individuals and therefore the higher the payoff the higher the probability of being imitated by others [11, 22]. New generations are evolving following a discrete imitation dynamic[20, 12], where a randomly chosen individual is copying the strategy of another one according with the Fermi imitation probability function [3, 25]. Then we calculate the probability that each strategy invades the others through fixation probabilities [13, 22]. These probabilities determine a transition matrix of a Markov chain among strategies. The stationary distribution of strategies [12, 6] represents the relative time that the population spends adopting each of the strategies.

3 Results and discussion

When apology is not possible, the most successful individuals are those that propose commitments (and are willing to pay the cost ϵ) and, following the agreement, cooperate until an error occurs. Once this mistake occurs the commitment is broken and then these individuals take revenge by defecting in all the remaining rounds. This result is relatively important as it confirms analytically that it appears that individual players prefer to take revenge when apology and forgiveness are not possible [17, 18]: Withholding benefits from the wrongdoer appears to induce cooperation even better than to the well-known TFT-like strategies in the IPD extended with commitments.

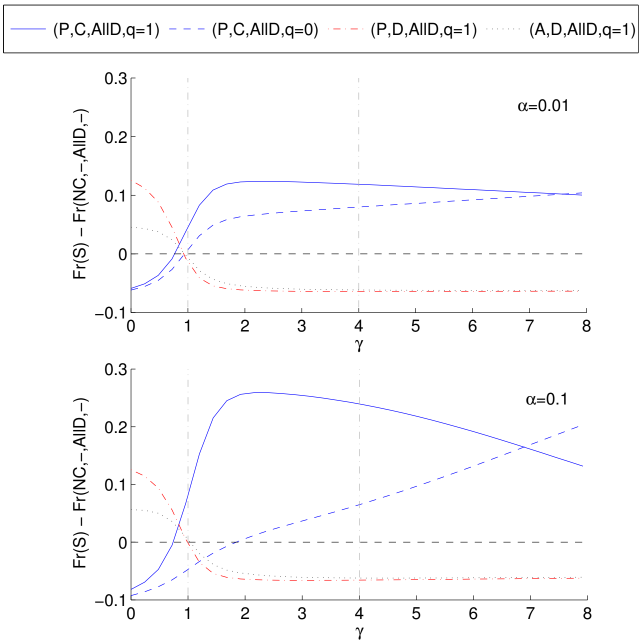

Figure 1. Stationary distribution of the main strategies with respect to the stationary distribution of the pure defectors as a function of the apology cost γ for $\alpha = 0.01$ and $\alpha = 0.1$. Vertical dashed lines mark the values of c and δ. We assumed $\omega = 0.9$, $b/c = 2$, $\epsilon = 0.25$, and $\delta = 4$. Figure reproduced from [16].

Once the model allowed for individuals to apology and forgive, revenge-taking no longer was the dominating strategy, even in situations when there was a 10% chance of making mistakes ($\alpha = 0.1$). In the simulations, apology was defined as a costly behaviour: when a player apologises she pays a fine γ to the player that was harmed, which was always accepted by the co-player. As we show in Figure 1, our analytical results revealed that apology an forgiveness, and hence long-term cooperation, evolves when the apology cost is sufficiently high, i.e. slightly higher than the cooperation cost but much less than the compensation that can be acquired by ending the commitment. When the apology cost is too high, revenge dominates apologising. When the apology cost is too low (lower than the cost of cooperation), apology becomes exploited by cheaters who use it to get more out of defecting, thus dishonest apologisers evolve. In a follow up experiment we also examined the evolutionary dynamics of an agent-based model wherein players have a personal apology value and a forgiveness threshold. This model confirmed our results.

The research we discussed here showed for the first time in an analytical and numerical manner how emotional responses like revenge, apology and forgeiveness are relevant for the evolution of cooperation. Extension of this work to the repeated public goods game confirm the conclusions (work in progress) we were able to draw, urging us to further explore the role of emotions in social interactions.

Acknowledgments

LMP acknowledges support from FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013. LMV and TL acknowledge the support of from the F.R.S.- F.N.R.S. (grant FRFC nr. 2.4614.12) and the F.W.O (grant G.0391.13N). TAH acknowledges support from URF Funding (Teesside University).

REFERENCES

[1] J. Abeler, J. Calaki, K. Andree, and C. Basek, 'The power of apology', *Economics Letters*, **107**(2), 233 – 235, (2010).

[2] Robert Axelrod and William Donald Hamilton, 'The evolution of cooperation', *Science*, **211**, 1390–1396, (1981).

[3] Lawrence Blume, 'Now noise matters', *Games and Economic Behavior*, **44**, 251–271, (2003).

[4] Urs Fischbacher and Verena Utikal, 'On the acceptance of apologies', *Games and Economic Behavior*, **82**, 592–608, (2013).

[5] Robert H Frank, *Passions within reason: the strategic role of the emotions.*, WW Norton & Co, 1988.

[6] Drew Fudenberg and Lorens A. Imhof, 'Imitation processes with small mutations', *J. Econ. Theory*, **131**, 251–262, (2006).

[7] The Anh Han, Luís Moniz Pereira, and Tom Lenaerts, 'Avoiding or Restricting Defectors in Public Goods Games?', *Journal of the Royal Society Interface*, 20141203, (2014).

[8] The Anh Han, Luís Moniz Pereira, and Tom Lenaerts, 'Evolution of commitment and level of participation in public goods games', *Autonomous Agents and Multi-Agent Systems*, 1–23, (2016).

[9] The Anh Han, Luís Moniz Pereira, Francisco C. Santos, and Tom Lenaerts, 'Good agreements make good friends', *Scientific Reports*, (2013).

[10] Christoph Hauert, Arne Traulsen, H. Brandt, Martin A. Nowak, and Karl Sigmund, 'Via freedom to coercion: the emergence of costly punishment', *Science*, **316**, 1905–1907, (2007).

[11] Josef Hofbauer and Karl Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, 1998.

[12] L. A. Imhof, D. Fudenberg, and Martin A. Nowak, 'Evolutionary cycles of cooperation and defection', *Proc. Natl. Acad. Sci. USA*, **102**, 10797–10800, (2005).

[13] Samuel Karlin and Howard M. Taylor, *A First Course in Stochastic Processes*, Academic Press, New York, second edn., 1975.

[14] Robert Kurzban, Kevin McCabe, Vernon L Smith, and Bart J Wilson, 'Incremental commitment and reciprocity in a real-time public goods game', *Personality and Social Psychology Bulletin*, **27**(12), 1662–1673, (2001).

[15] B.A. Liang, 'A system of medical error disclosure', *Quality and Safety in Health Care*, **11**(1), 64–68, (2002).

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

[17] Michael E McCullough, *Beyond Revenge, the evolution of the forgiveness instinct*, Jossey-Bass, 2008.

[18] Michael E McCullough, Robert Kurzban, and Benjamin A Tabak, 'Evolved mechanisms for revenge and forgiveness', in *Human aggression and violence: Causes, manifestations, and consequences. Herzilya series on personality and social psychology*, eds., Philip R Shaver and Mario Mikulincer, 221–239, American Psychological Association, Washington, DC, US, (2011).

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

[20] Martin A. Nowak, Akira Sasaki, Christine Taylor, and Drew Fudenberg, 'Emergence of cooperation and evolutionary stability in finite populations', *Nature*, **428**, 646–650, (2004).

[21] Nicolas Peterson, 'Demand sharing: reciprocity and the pressure for generosity among foragers', *American anthropologist*, **95**(4), 860–874, (1993).

[22] Karl Sigmund, *The Calculus of Selfishness*, Princeton University Press, Princeton, 2010.

[23] Nick Smith, *I was wrong: The meanings of apologies*, volume 8, Cambridge University Press New York, 2008.

[24] Nick Smith, *Justice Through Apologies: Remorse, Reform, and Punishment*, Cambridge University Press, 2014.

[25] Arne Traulsen, Martin A. Nowak, and Jorge M. Pacheco, 'Stochastic dynamics of invasion and fixation', *Phys. Rev. E*, **74**, 011909, (2006).

[26] James Woodburn, 'Egalitarian Societies', *Man*, **17**(3), 431–451, (September 1982).