Apology and forgiveness evolve to resolve failures in cooperative agreements[1]

Luis Martinez-Vaquero [ab] The Anh Han [c] Luis Moniz Pereira [d]

Tom Lenaerts [ab]

[a] *AI lab, Vrije Universiteit Brussel, Brussels, Belgium*
[b] *MLG, Université Libre de Bruxelles, Brussels, Belgium*
[c] *School of Computing, Teesside University, Middlesbrough, UK*
[d] *NOVA-LINCS,Universidade Nova de Lisboa, Caparica, Portugal*

Abstract

When interactions are repeated mistakes, whether intentionally or not, tend to occur. Researchers have argued that revenge, apology and forgiveness are mechanisms that humans have acquired to ensure that intentional mistakes are avoided and that mutually beneficial relationships can continue. We have shown in the context of the iterated prisoners dilemma wherein agents can decide to make cooperative agreements that these three behaviours emerge spontaneously. Concretely our work reveals that apology and forgiveness are very efficient even in a very noisy environment and ensure long lasting relationships. Yet in order for apology to work, it needs to be sufficiently costly otherwise taking revenge by defecting is the most profitable behaviour. This research has direct implications for online socio-technological systems who's success depends on the trust users (and agents) have in the other users (or agents) participating in the system.

Commitment deals – defined as prior agreements with potentially posterior compensations in case the agreements fail – are most often established to ensure favourable interactions over longer time periods. Experiments have shown that commitment facilitates cooperation in long-term interactions [2], especially when it is voluntary. Moreover, long-term commitments as opposed to one-shot ones [3] 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, as suggested in [4], commitment in long-term relationships may induce behavioural differences as they may remove the need for reciprocal behaviour like tit-for-tat (TFT).

In [1] we provide for the first time analytical and numerical results for the viability of the commitment strategy within the context of an iterated prisoners dilemma (IPD) [5], expanding our prior work on commitment in the one-shot prisoners dilemma and the public goods game [3, 6]. To study commitment within the IPD context the strategy space needed to be expanded: First, as it is possible for agreements to end before the interaction is finished, 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, individuals need to decide whether to continue the agreement when a mistake is made, or end it collecting the compensation. In a cooperative agreement this occurs when a player defects even though she agreed to cooperate with her partner. All these choices define the strategy space of each individual player.

Our work reveals first of all how the detrimental effect of having a large arrangement cost, which was observed for one-shot games [3], is limited as the setup cost is only paid when starting the interaction. Individuals that propose commitments (and are willing to pay their cost) and, following the agreement, cooperate unless a mistake occurs are now the most successful players. But if the agreement is violated through a defection before the IPD is finished then these individuals take revenge by defecting in the

[1]The full paper has been published in the journal *Scientific Reports*, 5, Article number: 10639, June 2015 [1]."

remaining rounds. This observation confirms analytically what has been argued in [7]. Moreover, although defection leads here to the withholding of a the benefit from both players, this revenge-taking behaviour leads to a more favourable outcome for cooperation as opposed to the well-know TFT.

Yet, as mistakes may not be intentional and stopping a mutually beneficial interaction (i.e. cooperation in the IPD) may be strategically a bad idea, individuals may decide not to end the agreement. It might be better to apologise and forgive, continuing the agreement without taking revenge. To study this question the commitment model was extended with an apology-forgiveness mechanism, where apology was defined either as a systemic 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 work [7]. 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 is detrimental for the level of cooperation level that is produced by commitment and compensation model. 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 further as to what one would expect from simply taking revenge.

As we argued in [8] these results may have direct implications in Artificial Intelligence research: In the context of hybrid socio-technical systems which use reputation scoring to ensure trust [9], apology and forgiveness have been shown to provide additional gains in the transactions. As violations of trust may also occur within systems of interacting non-human autonomous agents [10], our behavioural results may also provide essential mechanisms to ensure that such systems will survive the critical evaluation of their users.

References

[1] 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.

[2] 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.

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

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

[5] Robert Axelrod and William Donald Hamilton. The evolution of cooperation. *Science*, 211:1390–1396, 1981.

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

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

[8] Tom Lenaerts, Luis Martinez-Vaquero, The Anh Han, and Luís Moniz Pereira. Conditions for the evolution of apology and forgiveness in populations of autonomous agents. In *2016 AAAI Spring Symposium, Technical Reports*, pages 242–248. AAAI Press, 2016.

[9] Cynthia L Corritore, Beverly Kracher, and Susan Wiedenbeck. On-line trust: concepts, evolving themes, a model. *International journal of human-computer studies*, 58(6):737–758, 2003.

[10] Stephen Marsh and Pamela Briggs. Examining trust, forgiveness and regret as computational concepts. In *Computing with social trust*, pages 9–43. Springer-Verlag, 2009.