Conditions for the Evolution of Apology and Forgiveness in Populations of Autonomous Agents

Conditions for the Evolution of Apology and Forgiveness in Populations of Autonomous Agents

Tom Lenaerts and
Luis A. Martinez-Vaquero
MLG, Université Libre de Bruxelles,
Boulevard du Triomphe CP212
Brussels, Belgium and AI lab,
Vrije Universiteit Brussel,
Pleinlaan 2, Brussels, Belgium

The Anh Han
School of Computing,
Teesside University,
Middlesbrough, UK

Luís Moniz Pereira
NOVA LINCS,
Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa,
Portugal

Abstract

We report here on our previous research on the evolution of commitment behaviour in the one-off and iterated prisoner's dilemma and relate it to the issue of designing non-human autonomous online systems. We show that it was necessary to introduce an apology/forgiveness mechanism in the iterated case since without this restorative mechanism strategies evolve that take revenge when the agreement fails. As before in online interaction systems, apology and forgiveness seem to provide important mechanisms to repair trust. As such, these result provide, next to the insight into our own moral and ethical considerations, ideas into how (and also why) similar mechanisms can be designed into the repertoire of actions that can be taken by non-human autonomous agents.

Human decision making implies *trust* (Marsh 1994; Marsh and Dibben 2003; Eckel and Wilson 2004): Many face-to-face interactions are risky as participants are unsure about the intentions of others. Each participant needs to trust that their opponents will not exploit a given situation. To reduce the uncertainty about the trustworthiness of a partner we have naturally acquired capabilities that allow us to make better decisions. Typical examples are different forms of intention recognition, like reading facial expressions, tone of voice or other cues. Without such mechanisms we would be randomly guessing what choices to make.

Online as opposed to face-to-face forms of interacting block those naturally acquired assessment mechanisms, which in turn hindered the initial spread of these new forms of social interaction as it was difficult to trust other participants. Reputation scoring provided an important tool to alleviate this problem (Riegelsberger, Sasse, and McCarthy 2005; Sabater and Sierra 2005; Riegelsberger et al. 2006) and recent studies have shown that adding apology and forgiveness mechanisms in online systems provide further important improvements to reparing violations of trust in one-off encounters (Vasalou and Pitt 2005; Vasalou, Pitt, and Piolle 2006; Vasalou, Hopfensitz, and Pitt 2008). Hence technology-mediated non-autonomous interaction services appear to incorporate step by step novel mechanisms that allow us to use our natural abilities also within that context, ensuring that users trust more and more these new forms of social interaction. Note that, from a game theoretical angle, this reputation work is associated with the evolution of cooperation through indirect reciprocity (Alexander 1987; Nowak and Sigmund 1998; Ohtsuki and Iwasa 2004; Martinez-Vaquero and Cuesta 2013).

Given the importance of those mechanisms, it makes sense to ask how similar mechanisms could be used in situations where human interactions are mediated by *autonomous* non-human agents as also in those cases trust and the violation of trust may be an issue (Pinyol and Sabater-Mir 2013). With the advent of mobile devices connected through an omnipresent web, it is not difficult to imagine that while we are getting on with our daily business, our artificial counterparts are doing the same. As with humans, these artificial counterparts need to make decisions that may involve risk, requiring trust among them. (Marsh and Briggs 2009) provide a nice Ambient Intelligence example wherein the ultimate goal is to share information with other autonomous agents, which can then use it to reason with and produce socially acceptable things for the individual with whom the autonomous agents are associated (see article pages 13-15). It is highly possible that in such situations trust may be violated (intentionally or non-intentionally), requiring some form of reparation. Either this reparation can be done by the humans managing the autonomous agents, generating a spillover from the artificial to the real world, or we need to develop systems that can deal with such situations. (Marsh and Briggs 2009) argued that introducing computational variants of apology and forgiveness into such a system of autonomous agents may again provide the tools to ensure sustainable cooperative interactions.

This raises the question how exactly one should design agent-based systems so that high levels of cooperation can be achieved, minimising at the same time conflicts in an environment populated by autonomously deciding agents as well as the spillover of non-trust issues to the real world. As it was argued that we have acquired the strategies like revenge-taking, apologising and forgiving through natural selection (McCullough 2008), we hypothesise that evolution can also reveal the conditions that allow for these strategies to emerge, which can in turn be used to design agent-based systems. Our recent evolutionary dynamics research on the evolution of behaviours like commitments, revenge, forgiveness and apology within the context of social dilemmas examines this hypothesis and hence may allow us to

provide an answer to the earlier question (Han et al. 2013; Han, Pereira, and Lenaerts 2014; Martinez-Vaquero et al. 2015).

In the following sections we report on how we defined the notion of commitment and how it was expanded from one-off to repeated games. Then we show that in a noisy environment where errors occur apology and forgiveness provide crucial mechanisms for maintaining relationships. We conclude by discussing the relevance of these results for non-human autonomous systems, which is the focus of this AAAI symposium.

Commitment is giving up options

Creating agreements and asking others to commit to such agreements provides a basic behavioural mechanism that is present at all the levels of society, playing a key role in social interactions, ranging from personal relationships such as marriage to international and organisational ones such as alliances among companies and countries (Sosis 2000; Nesse 2001; Frank 2001). Anthropological data revealed that commitment strategies, as for instance demand-sharing (Woodburn 1982), have played an essential role in sustaining early hunter-gatherer societies.

Commitments clearly also extend to the use of online systems: When using such systems users are expected to behave according to certain rules, which may or may not be specified explicitly in an user agreement. Violations of this agreement may lead to removal from the system or has implications for the user's reputation (which could lead to loss of income in case of systems like eBay (Khopkar, Li, and Resnick 2005)). One can also imagine that non-human autonomous agents ask each other to commit to behave in a particular manner before an exchange of information is performed, imposing a penalty or compensation when either one of them violates such an agreement. Although it is not clear at this point what form of penalty would make most sense for an autonomous agent, one could imagine that a reduction in the appreciation of the human user for that agent, which would reduce the trust this human may have in the autonomous system and as a consequence the viability of this system for other users, may encourage the autonomous agent to behave according to the rules of the agreement.

In our work we defined commitment behaviour as a two-stage process wherein the commitment proposer asks her partner to commit to play cooperatively in a prisoners dilemma (PD) game. If the answer is positive the game is played and both players acquire the payoff following from the game. If the partner refuses then the game is not played and both receive no payoff. This definition follows the reasoning of (Nesse 2001) who required that committing players should give up on certain choices, which in the case of simple social dilemma's like the PD is the choice to defect. Players can of course cheat, but our prior work (see also Figure 1) shows that under a range of constraints imposed on the cost of creating the agreement (ϵ) and the compensation (δ) that needs to be paid when the agreement is violated, these cheaters (which we call fake committers) are dominated by cooperative commitment proposers (Han et al. 2013). Commitments hence force cheaters to show their hand, reducing

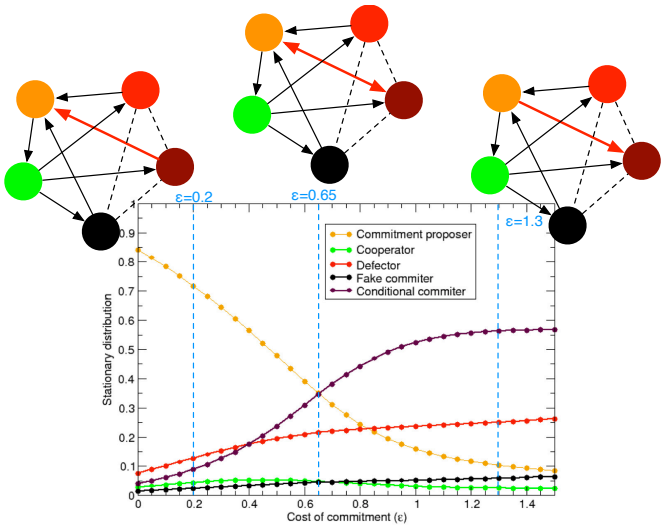

Figure 1: Frequency of the strategies in the set $\{C, D, COM, FAKE, COND\}$ as a function of the cost of commitment (ϵ) in the prisoners dilemma (PD, defined by parameters T, R, P and S). Depending on the cost of setting up the agreement (ϵ), commitment proposing strategies (COM) can be dominant, generating cooperative behaviour in the PD to levels higher than what would be expected if commitments were absent. Increasing the cost leads to an increase in players that will accept to commit but refuse to pay the commitment cost (Conditional committers or COND) relative to the commitment proposing strategies. Details for the method producing these results and their analysis can be found in (Han et al. 2013). Each graph above the plot visualises in a simplified manner the selection dynamics between the different strategies for a specific cost of setting up the agreement. As can be seen, increasing the cost only alters the relative dynamics between proposing and conditional committers. Parameter values: $T = 2,\ R = 1,\ P = 0,\ S = -1; \beta = 0.1; \delta = 4\ N = 100$. This figure was regenerated using the information in (Han et al. 2013)

the risk for the agent that wants to cooperate in the game. A commitment is the manifestation of an intention, the other side of the coin of intention recognition. These results remain valid when moving from pairwise PD to the n-player public goods game (Han, Pereira, and Lenaerts 2014). Note in Figure 1 that another type of free-riders, which we call conditional committers (they will never pay the cost but will accept an agreement when someone else proposes), emerges when the cost of setting up the agreement increases.

Notwithstanding the power of making agreements, the problem is that once the agreement is violated the interaction ends, which clearly is detrimental for long-term interactions or the long-term use of non-human autonomous agents as a social extension of users. Clearly agreements, which include the use of online ambient intelligence systems, are established to ensure persistent interactions amongst users.

Evolution of Revenge versus Tit-for-tat

To determine the effect of commitment behaviour on long-term interactions we extended the previous work to the Iterated Prisoners Dilemma (IPD) (Martinez-Vaquero et al. 2015). In this extension proposer strategies will ask their co-player to commit for as long as the game lasts (which is modelled by a probability ω that represents the probability that another round is played) as opposed to only one round. Offences, which can either be accidental or on purpose, are modelled as noise in the decision of each player, meaning that with a probability α a player will play D when she intended to play C and vice versa. The higher α the more frequent this error occurs. Clearly this approach of modelling violations is not handling explicitly the intentionality of the offence nor does it take into account any emotional effects or repeated violations. Nonetheless it allowed us to tune the frequency of violations in order to study their impact on strategy evolution. Strategies like Tit-for-tat loose their advantage over defectors when errors occur frequently ($\alpha > 0.01$) (see also (Nowak and Sigmund 1993; Sigmund 2010)) , requiring other strategies capable of dealing with such situations.

When an offence occurs before the end of the game (which is determined by the probability ω), an autonomous agent needs to decide whether, after collecting the compensation δ, to end the game (hence disregarding ω) or to continue playing, using another action or strategy from the one used before the violation occurred (in our case this is C). Additionally, commitment proposers could simply refuse to play the IPD if the partner does not commit. These choices result in four different strategies (see also Figure 2) for dealing with the start and end of agreements, which were explored in detail in (Martinez-Vaquero et al. 2015).

In summary, our results showed: First that strategies that do not end the agreement upon violation but keep playing until the game finishes dominate strategies that end also the game immediately after the offence. Moreover, even when no agreement is established it is better continue the game. Second, when analysing in detail, we observed that the most successful strategies are those that propose commitments (and are willing to pay their cost) and, following the agreement, cooperate unless a mistake occurs. When the commitment is broken then these individuals take revenge by defecting in the remaining interactions. This result is relatively important as it confirms analytically that revenge evolves to ensure that individuals behave appropriately during repeating interactions (McCullough 2008; McCullough, Kurzban, and Tabak 2011). The threat of revenge, through some punishment or withholding of a benefit, may discourage interpersonal harm. This result is furthemore intriguing as revenge by withholding the benefit from the transgressor may lead to a more favourable outcome for cooperative behaviour in the noisy IPD as opposed to the well-known TFT strategy. Even when the game continues after the agreement do we see that AllD is always better than TFT, for all noise levels.

Our results hence seem to indicate that when autonomous agents engage in information exchange or other social activities through agreements (with associated compensations)

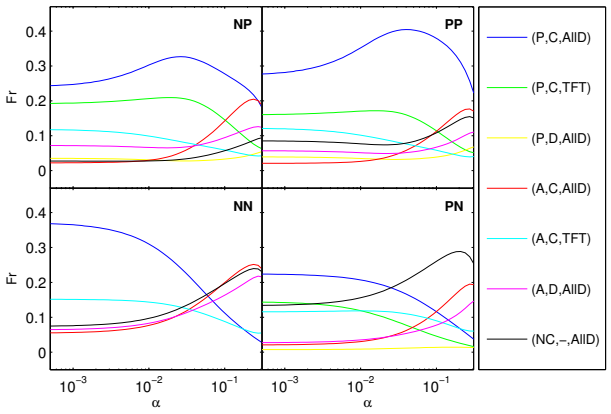

Figure 2: Frequency of most abundant strategies relative to the level of noise (α) in the iterated prisoners dilemma (IPD). In this case each strategy consist of 3 elements : 1) whether to initiate (P), accept (A) or refuse (NC) a commitment; 2) how to behave when an agreement is established (C or D); 3) how to behave when the agreement fails or even never started ($AllC$, $AllD$, TFT). Using finite-population evolutionary dynamics (see (Martinez-Vaquero et al. 2015) for method details) we examine the relative importance of the surviving strategies. The lower left scenario assumes that whenever an agreement is not established or when it fails, there is no game (NN). The top right assumes that whatever happens the game is always played until its end using the relevant actions in the agreement (PP). The two remaining plots are variations where the game ends either when there is no agreement (NP) or when it fails (PN). Black is the strategy that corresponds to defection, with serves as a reference for the other results. The PP panel shows that commitment proposing strategies which take revenge when the agreement fails, i.e. (P,C,AllD), are most viable when the game continues. This is also the case all other panels, yet in NN and PN but it breaks down for higher noise levels. See supplementary information and Supplementary Figure 1 in (Martinez-Vaquero et al. 2015) for details. Parameters: $\beta = 0.1$; $\epsilon = 0.25$; $\delta = 4$; $N = 100$.

then these agents, if simplistic, should be overly strict with offenders and actively disrupt their participation in the shared activity. Clearly, this is not what we would expect as real-world users from our artificial counterparts. It would require that users mediate continuously any problem emerging between our artificial alter-ego's and those belonging to other users. As also argued by (Marsh and Briggs 2009) a solution presents itself through the introduction of computational implementations of apology and forgiveness.

Evolution of Apology and Forgiveness

Apology and forgiveness are of interest as they remove the interference of external institutions (in the case of autonomous agents this would be human users), which can be quite costly to all parties involved, in order to ensure co-

operation. Evidence from other domains shows that there is a much higher chance that customers stay with a company (they subsequently forgive) that apologises 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 (Smith 2014).

Given this importance we need to determine how apology and forgiveness can evolve in a population of non-human autonomous agents as mechanisms to create sustainable social interactions. Understanding how they may evolve provides direct insight into our own moral and ethical considerations, providing ideas into how (and also why) similar mechanisms can be designed into the repertoire of actions that can be taken by autonomous agents.

Going back to the IPD model discussed in the previous section, agents now need to decide whether it is worthwhile to end the agreement, collect the compensation and start taking revenge when a mistake is made or whether it is better to forgive the co-player and continue the mutually beneficial agreement. Reactive anger and immediate revenge needs to be pondered against the potential longer term interest of continued playing allowed by forgiveness. Moreover, should forgiveness follow a costless or costly form of apology? In other words, what are the moral and ethical considerations of our nonhuman players to decide to forgive an opponent, where morality for them is defined in terms of costs and benefits obtained in the IPD?

We extended the strategies with the option to apologise and/or forgive, where apology was defined either as an external (pre-defined and global for all individuals, for which the results are visualised in Figure 3) or individual (susceptible to evolution) parameter in the model. Additionally, the apology was associated with a cost (γ) for the apologiser. In both cases, we were able to show 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, which means that, the amount offered when apologising has to be high enough (yet not too high), which is also corroborated by a recent experimental psychology paper (McCullough et al. 2014). As visualised in Figure 3, sincerity means that the cost of apologising (γ) is higher than a certain threshold, which in our case is closely related to the cost of cooperation ($c = 1$) typically used in social dilemmas. Introducing apology and forgiveness in the commitment model produces even higher cooperation levels than in the revenge-based outcome mentioned earlier.

Figure 3 also shows that when the apology is not sincere enough (lower than the cost of cooperation), fake committers , those that propose (or accept) to commit with the intention of taking advantage of the system by defecting and apologising continuously, will dominate the population. In this situation, the introduction of the apology-forgiveness mechanism lowers the cooperation level below that which revenge-based commitments can generate by themselves. Hence our results identify two moral/ethical thresholds on how sincere apology needs be in IPD to sustain cooperation: i) When the apology cost is lower than this limit the level of cooperation is reduced relative to the level one could expect

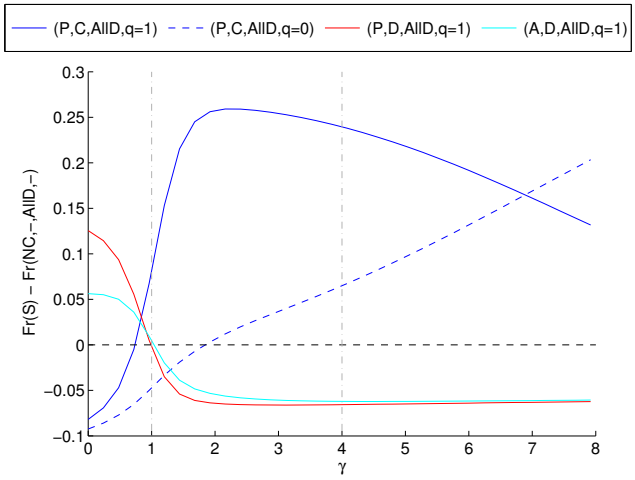

Figure 3: Frequency of strategies relative to the cost of apologising (γ) for high noise ($\alpha = 0.1$) in the iterated prisoners dilemma. Adding apologies and forgiveness extends each strategy with a fourth element: the probability to apologise or not, i.e. q. The plot shows the frequency of the most abundant strategies in the PP scenario relative to the number of true defectors, i.e. (NC, -, AllD). When the cost of apologising goes beyond 1, one can observe a steep increase in the commitment proposing strategies that apologise ($q = 1$) when they make mistakes. Below 1, strategies that strategically exploit the apologise mechanisms dominate. When the cost of apology becomes too high revenge, i.e. (P,C, AllD, q=0), is a better option to sustain cooperation. This figure was adapted from (Martinez-Vaquero et al. 2015). Parameters: $\beta = 0.1$; $\epsilon = 0.25$; $\delta = 4$; $N = 100$.

from simply taking revenge. ii) When the apology cost becomes too high it does not pay to be kind, since simply taking revenge when a commitment is broken leads to a more favourable outcome for individuals. An ongoing analysis shows that these results remain consistent when moving to n-player public goods games.

Discussion

The results discussed in the previous sections reveal under which conditions strategies that incorporate commitments, apology and forgiveness evolve and become dominant in a population for one-off and repeated interactions. As argued earlier, understanding the conditions that determine their evolution may provide insight into how systems comprising non-human autonomous agents need to be designed to ensure sustained functionality.

Considering one-off encounters, apology and forgiveness are not immediately relevant. As opposed to the work on reputation-based indirect reciprocity systems discussed in the introduction, agents have no reputation and hence violations will not affect their probability to team up with another agent, thus there is no reason to restore the reputation. It would nonetheless be interesting to evaluate how commitments could benefit from reputation information. One can

imagine that when reputations are highly positive it may not be necessary to pay the cost of setting up an agreement with another agent and simply cooperate as there is no reason to believe that the partner will not act as expected. Only when reputations are intermediate or fairly low will a commitment with corresponding compensation be necessary, but this will depend on the inherent value of the compensation for the offender. Our earlier work revealed that the compensation does not need to be as excessive as in punishment-based approaches (Han et al. 2013). Lowering the reputation may by itself be the compensation, which would make the commitment model very similar to a system based on indirect reciprocity. Note that one could consider to introduce a compensation mechanism as the one we envision in here for buyers and sellers in online commercial systems like eBay, which would allow either buyer or seller to demand compensation when harmed. The compensation could be inversely proportional to the reputation. In a similar sense the compensation could be inversely proportionate to the recognition of positive intentions (Han et al. 2015), meaning that if a positive intention is recognised then less compensation is demanded to the co-player for commitment.

Second, when interactions are repetitive, apology and forgiveness turn out to be important. Our results show that without apology and forgiveness, the commitment model leads to revenge, which, as was mentioned already earlier, would be counterproductive for the user appreciation of a system with non-human autonomous agents (see end of section on the evolution of revenge): An Ambient Intelligence system containing autonomous agents that represent humans in the virtual world would loose its relevance if users need to mediate continuously between their agents and those of others in such a system. Worse, conflicts from within the system could spill-over to the real world, which could be detrimental to relationships established among ourselves (see also (Marsh and Briggs 2009)). Therefore mechanisms need to be in place that ensure that conflicts are likely to be resolved, without human interference.

Our results on apology and forgiveness indicate that in order for these actions to work in a reputation-less autonomous system, offenders need to make costly apologies, where the cost should be higher than what one would receive if one behaved as expected. Yet the cost should not be too extreme (i.e. much higher than the compensation that needs to be paid when the agreement fails) as this would imply that revenge taking would be more efficient for any user in terms of costs and benefits. As there are essentially no humans directly involved in a system of autonomous agents one needs to determine what form an apology cost should take. Any measurable penalty that counters the utility of the agent could perform this role. One could imagine that the agent, since it is an extension of the human user, will aim to not harm the appreciation it receives from his or her user. Clearly the end-user is interested in the results (information or other benefits) produced by the agent. Appreciation is reduced when the relevant information is not obtained or whether to attain the information certain violations were made against the agents of relatives or friends. On the other hand, a user could appreciate that the agent resolved conflicts, without harming its prime objective. Thus reputations could now be assigned to an agent by either the agents that did business with it or by the end-user. The autonomous agent will need to balance between them, and their might be conflicts that need to be resolved when the interests of both parties are not aligned. Thus also in this case the reputations of other agents and maybe also the reputation of their users could prove useful as an extension of the commitment model including apology and forgiveness. This issues are currently under investigation.

Clearly the work on the evolution of commitment behaviour with or without apology and forgiveness can be further extended to make it better fit the needs in non-human autonomous systems. Moreover, the results we have presented are only applicable to the question of trust as far as trust can be quantified by the level of cooperation in a system, and this expressed in terms of costs and benefits. Consequently our focus is for now on cognitive trust as opposed to emotional trust (Corritore, Kracher, and Wiedenbeck 2003; Vasalou, Hopfensitz, and Pitt 2008).

Forgiveness, as explained by (Vasalou, Hopfensitz, and Pitt 2008), transforms negative motivations towards the offender into positive ones. Yet these positive ones are influenced by a number of factors: 1) the severity of the offence, 2) frequency/severity of previous offences, 3) the offender's intention, 4) the offender's apology and efforts to repair and 5) previous interactions with the offender during which he/or she has demonstrated benevolence. In the context of our work so far, only item 4 and maybe partially item 2 have been considered. Additional extensions to our model will be introduced to explore also the other aspects. For instance, in our current setup the mistakes made by the agents are external, in the sense that they are defined by a system-wide parameter. This choice allowed us to explore how the frequency of errors influences the evolving behaviours, not considering intentionality, severity etc. Yet clearly there is a relationship between the frequency of a mistake, on one hand, and its intentionality and severity on the other hand. Further developments should take such issues into account.

To conclude, morality for artificial non-humans has often stressed the cognitive and reasoning abilities by themselves. What our work has also shown is that the study of the population aspect of morality is inevitable and provides insights and cues, plus aggregate testing, of cognitive abilities envisaged as strategies, requiring that the individual and collective realms of morality studies must perforce be entwined.

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).

References

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

Alexander, R. D. 1987. *The Biology of Moral Systems*. New York: Aldine de Gruyter.

Corritore, C. L.; Kracher, B.; and Wiedenbeck, S. 2003. Online trust: concepts, evolving themes, a model. *International Journal of Human-Computer Studies* 58(6):737–758.

Eckel, C. C., and Wilson, R. K. 2004. Is trust a risky decision? *Journal of Economic Behavior & Organization* 55(4):447–465.

Frank, R. H. 2001. Cooperation through Emotional Commitment. In Nesse, R. M., ed., *Evolution and the capacity for commitment*. New York: Russell Sage. 55–76.

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

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

Han, T. A.; Pereira, L.; and Lenaerts, T. 2014. Avoiding or Restricting Defectors in Public Goods Games? *Journal of the Royal Society Interface* 12(103): 20141203.

Khopkar, T.; Li, X.; and Resnick, P. 2005. Self-selection, slipping, salvaging, slacking, and stoning: the impacts of negative feedback at ebay. In *Proceedings of the 6th ACM conference on Electronic commerce*, 223–231. ACM.

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

Marsh, S., and Briggs, P. 2009. Examining trust, forgiveness and regret as computational concepts. In *Computing with social trust*, Human-Computer Interaction Series. Springer-Verlag. 9–43.

Marsh, S., and Dibben, M. R. 2003. The role of trust in information science and technology. *Annual Review of Information Science and Technology* 37(1):465–498.

Marsh, S. 1994. Trust in distributed artificial intelligence. In *Artificial Social Systems*. Springer Berlin Heidelberg. 94–112.

Martinez-Vaquero, L. A., and Cuesta, J. A. 2013. Evolutionary stability and resistance to cheating in an indirect reciprocity model based on reputation. *Phys. Rev. E* 87:052810.

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

McCullough, M. E.; Pedersen, E. J.; Tabak, B. A.; and Carter, E. C. 2014. Conciliatory gestures promote forgiveness and reduce anger in humans. *Proceedings of the National Academy of Sciences* 111(30):11211–11216.

McCullough, M. E.; Kurzban, R.; and Tabak, B. A. 2011. Evolved mechanisms for revenge and forgiveness. In Shaver, P. R., and Mikulincer, M., eds., *Human aggression and violence: Causes, manifestations, and consequences. Herzilya series on personality and social psychology*. Washington, DC, US: American Psychological Association. 221–239.

McCullough, M. E. 2008. *Beyond Revenge, the evolution of the forgiveness instinct*. Jossey-Bass.

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

Nowak, M., and Sigmund, K. 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* 364(6432):56–58.

Nowak, M. A., and Sigmund, K. 1998. Evolution of indirect reciprocity by image scoring. *Nature* 393:573–577.

Ohtsuki, H., and Iwasa, Y. 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. *J. Theor. Biol.* 231:107–120.

Pinyol, I., and Sabater-Mir, J. 2013. Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review* 40(1):1–25.

Riegelsberger, J.; Vasalou, A.; Bonhard, P.; and Adams, A. 2006. Reinventing trust, collaboration and compliance in social systems. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, 1687–1690. ACM.

Riegelsberger, J.; Sasse, M. A.; and McCarthy, J. D. 2005. The mechanics of trust: A framework for research and design. *International Journal of Human-Computer Studies* 62(3):381–422.

Sabater, J., and Sierra, C. 2005. Review on computational trust and reputation models. *Artificial intelligence review* 24(1):33–60.

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

Smith, N. 2014. *Justice Through Apologies: Remorse, Reform, and Punishment*. Cambridge University Press.

Sosis, R. 2000. Religion and intra-group cooperation: preliminary results of a comparative analysis of utopian communities. *Cross-Cultural Research* 34:70–87.

Vasalou, A., and Pitt, J. 2005. Reinventing forgiveness: A formal investigation of moral facilitation. In *Trust Management*, volume LNCS 3477 of *Proceedings of the Third International Conference, iTrust*. Springer. 146–160.

Vasalou, A.; Hopfensitz, A.; and Pitt, J. 2008. In praise of forgiveness: Ways for repairing trust breakdowns in one-off online interactions. *International Journal of Human-Computer Studies* 66(6):466–480.

Vasalou, A.; Pitt, J.; and Piolle, G. 2006. From theory to practice: forgiveness as a mechanism to repair conflicts in cmc. In *Trust Management*. Springer. 397–411.

Woodburn, J. 1982. Egalitarian Societies. *Man* 17(3):431–451.