Evolutionary Machine Ethics Synopsis

The Anh Han[1] and **Luís Moniz Pereira**[2]

School of Computing, Media and the Arts,
Teesside University, TS1 3BX, UK.
Email: T.Han@tees.ac.uk URL: https://research.tees.ac.uk/en/persons/the-anh-han/

NOVA Laboratory for Computer Science and Informatics (NOVA LINCS)
Departamento de Informática, Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa, 2829-516 Caparica, Portugal.
Email: lmp@fct.unl.pt URL: http://userweb.fct.unl.pt/~lmp/

Abstract. Machine ethics is a sprouting interdisciplinary field of enquiry arising from the need of imbuing autonomous agents with some capacity for moral decision-making. Its overall results are not only important for equipping agents with a capacity for moral judgment, but also for helping better understand morality, through the creation and testing of computational models of ethics theories. Computer models have become well defined, eminently observable in their dynamics, and can be transformed incrementally in expeditious ways. We address, in work reported and surveyed here, the emergence and evolution of cooperation in the collective realm. We synoptically discuss how our own research with Evolutionary Game Theory (EGT) modelling and experimentation leads to important insights for machine ethics, such as the design of moral machines, multi-agent systems, and contractual algorithms, plus their potential application in human settings too.

Keywords: Machine Ethics; Intention Recognition; Commitment; Apology; Forgiveness.

1. Introduction

Some of our previous research (Pereira & Saptawijaya, 2011; Han, Saptawijaya, & Pereira, 2012; Pereira & Saptawijaya, 2015, 20016a, 2016b, 2017; Saptawijaya & Pereira, 2015, 2016a, 2018) has focused on using logic programming techniques to computational modelling of morality sans emotions. In the realm of the individual, we have addressed questions of permissibility and the dual-process of moral judgments by framing together ingredients that are essential to moral agency: abduction, integrity constraints, preferences, argumentation, counterfactuals, and updates. Computation over these ingredients has become our vehicle for modelling the dynamics of moral cognition within a single agent, without addressing the cultural dimension (Prinz, 2016), because it is still absent in machines.

In the collective realm, we have reported on computational moral emergence (Han et al., 2015a), again sans emotions, using techniques from Evolutionary Game Theory (EGT). We have shown that the introduction of cognitive abilities, like intention recognition, commitment, revenge, apology, forgiveness and guilt, reinforce the emergence of cooperation in diverse populations, comparatively to their absence, by way of EGT models. This evolutionary collective realm will be the one surveyed here, with the pointers to our specialized publications to be indicated below.

In studies of human morality, these distinct but interconnected realms – one stressing above all individual cognition, deliberation, and behaviour; the other stressing collective morals and how they have emerged with evolution – seem separate but are synchronously evinced (Pereira & Saptawijaya, 2015, 2016b). There are issues concerned with how to bridge individual cognitive abilities and their deployment in the population. Namely the ability of recognising intention in another, taking even into account how others recognize our intention; the abilities of requesting commitment, and of accepting or declining to commit; the abilities to adaptively apply complementary mechanisms; those of monitoring group participation and delegate this process to an external party; those of cooperating or defecting; plus those of apologising, be it fostered by the of taking revenge or of forgiving.

This paper summarizes our collective realm research, modelling distinct co-present strategies of cooperative and uncooperative behaviour in complex evolutionary games. For a more elaborate discussion the readers are referred to our survey in Han & Pereira 2018.

2. Evolutionary Game Theory

Game theory was first developed in the 1940's, and the first work on the subject was Theory of Games and Economic Behavior by the mathematician John von Neumann (1903-1957) and the economist Oskar Morgenstern (1902-1977), (Neumann & Morgenstern, 1944). At the time it was directed at the economy, but it was subsequently applied to the Cold War. When some such situation gets complicated, there is need to resort to sophisticated mathematical tools—and computer simulations—to deal with equations that cannot otherwise be solved.

The games theme is as complex as it is interesting and filled with diverse niches. We can envisage genes and memes ("cultural genes"), and their mutual combinations, as ongoing strategies in the game of evolution, raising issues and posing questions related to survival and winning. We can envisage too the combinatorial evolution of such strategies, and their possible mutations according to diverse conditions, which conditions can either be other game partners or the game rules (or Nature's) circumstances. The notion of game includes uncertainty, and whenever there is uncertainty there has to be some attending strategy, spelling the moves one makes with given probability. When there is co-presence of evolving strategies from several partners, along with the idea of game payoff, we are dealing with the notion of evolutionary game, which can be examined and studied in an abstract and mathematical manner.

There are zero sum games and non-zero-sum games. The zero-sum ones are those that, by their rules, some players win, some players loose. In Nature's evolution, conditions are those of non-zero sum—all can win or all can lose. Robert Wright (2001) analyses the evolution of culture and civilization with the underlying idea that, in Nature, non-zero-sum games are possible, wherefore a general gain may be obtained through cooperation, thereby leading to illuminated altruism.

Sometimes, co-present strategies tend to achieve a tactical equilibrium. Take the hunter/prey relationship: neither the hunter wants to fully exterminate the prey, nor the latter can multiply indefinitely because that would exhaust the environment's resources. Some of these studies are used by Economics to understand what might be the overall result from the sum of interactions amongst the several game partners.

It is relevant to take into consideration if the game takes place only once with a given partner, or whether the same partner may be encountered on other occasions; how much recall does one have of playing with that partner; and whether the possibility of refusing a partner is allowed. Let us take a more detailed look at each of these situations in turn. We begin with the famous Prisoners Dilemma (PD), typical of the paradox of altruism. There are two prisoners, A and B, with charges on them. Either of them can denounce the other, or confess, or remain silent.

	Prisoner **B** – silence	Prisoner **B** – confession
Prisoner **A** – silence	**6 years in jail for each**	**A = 10 years in jail** **B = 2 years in jail**
Prisoner **A** – confession	**A = 2 years in jail** **B = 10 years in jail**	**8 years in jail for each**

Consider the above 2x2 payoff matrix where the lines correspond to the behaviour of A (to remain silent or confess), and the columns correspond to the behaviour of B (to remain silent of confess). At the intersection of B's «confess» column with A's «confess» row, both receive a jail sentence of 8 years. If A confesses and B does not, A will only get a 2-year sentence, whereas B gets 10 years, and vice-versa. There is an incentive for any of them to confess in order to reduce their own jail sentence. This way, it would eventually be advantageous for them not to remain silent. If one of them defects by confessing, but not the other, he will only stay in jail for 2 years whereas the other will be there for 10 years. But if both confess they will be sentenced to 8 years each. The temptation to confess is great, but so is the inherent risk, because, after all, they would mutually benefit from remaining silent, getting a 6-year sentence each in that case.

The prisoners know the rules of the game; they just do not know how the other player will act. It is advantageous for them to remain silent, but they do not know if the other one will confess. As long as one of them confesses, the silent other will be sentenced to 10 years in jail. A dilemma thus arises: it is good to remain silent, but there is the risk the other one will defect; and the one who does it faster will take the greater gain. In the worst-case scenario, both get an 8-year sentence — nobody will take the risk. This is a classic game, one where both players have the tendency to confess — and not benefit from what could be a mutual advantage, but one that they cannot assuredly profit from. Firstly, they do not have the opportunity to talk to one another; secondly, because even if they did, they would still risk being betrayed by the other. They have no joint solution in the sense that A and B could ever choose what is best for both, where there would be an assured increased advantage for the two.

All turns more complicated when one imagines A and B playing this game many times in succession, considering their experience of previous mutual behaviour in their past. In this case they can go on building mutual trust or distrust. If one betrayed the other once, the betrayed one's reaction will be vengeance, or simply intolerance, in some future opportunity. Let us now visualize a situation with multiple players and ask ourselves which will be, a long time, the best of all possible strategies — by running a computer simulation. Of course, one thing is to presuppose any one strategy can always match with any other, which is the base assumption, and then to move on to a situation where one wants to match only with certain players. Through these more realistic situations one begins to develop a game theory where social structure is included inside it.

Instead of letting a strategy evolve by choosing to copy those who win the most, one can alternatively let those who win the most to be those that reproduce the most, that is, they make more copies of themselves proportionately to the others, all the while keeping a bounded size for the whole population (since overall resources are not unlimited), through a random elimination of individuals. This other option can be adopted because those who lose more (or win less) are eliminated by virtue of their reduced number of copies, and also because only those who win more than some threshold are allowed to reproduce (reproduction is costly). The intent of this interpretation is that, throughout the game, strategies want to take over resources and occupy vital space in the population. Winning means having more energy to reproduce, while losing means not being able to persist with one's genetic/memetic continuity.

The evolutionary question that arises then is whether everyone can at length benefit more if they cooperate more. Which question hinges on how to prevent free-riders who want to gain more without having to incur in the expenses of cooperation. The evolution of any collective species clashes against this problem of balancing cooperation with opportunism. It is a strong theme in Evolutionary Psychology (Pereira 2012b), and one for which we can devise mathematical models and employ computers to perform both analytical computations, as well as long and repetitive simulations of the joint evolution of behavioural strategies in co-presence, typically done via mathematical games' implementation mixing competitive and cooperative situations, and providing mutation in strategies in order to detect focus points of long-term evolution stability.

3. Learning to recognise intentions and committing can resolve cooperation dilemmas

Few problems have motivated the amalgamation of so many seemingly unrelated research fields as has the evolution of cooperation (Nowak, 2006). Several mechanisms have been identified as catalysers of cooperative behaviour; see for example surveys by Nowak (2006) and Sigmund (2010). Yet these studies, mostly grounded on evolutionary dynamics and game theory, have neglected the important role played by intention recognition (Han and Pereira, 2013c) in behavioural evolution. In our work (Han et al., 2011, 2012a, 2012b; Han, 2013), we explicitly studied the role of intention recognition in the evolution of cooperative behaviour. The results indicate that intention recognisers prevail against the most successful strategies in the context of the Iterated Prisoner's Dilemma (IPD) (e.g. win-stay-lose-shift, and tit-for-tat like strategies), and promote a significantly higher level of cooperation, even in the presence of noise, plus the reduction of fitness associated with the cognitive costs of performing intention recognition. Our approach offers new insights into the complexity of—as well as enhanced appreciation for the elegance of—behavioural evolution when driven by elementary forms of cognition and learning ability.

It is important to note that intention recognition techniques have been studied actively in AI for several decades (Charniak and Goldman, 1993; Sadri, 2011; Han and Pereira, 2013a, 2013b, 2013c), with various applications such as for improving human-computer interactions, assistive living, moral reasoning, and team work (Pereira and Han, 2011a; Roy et al., 2007; Han et al., 2012d; Heinze, 2003; Han and Pereira, 2013b). Intentionality has been also shown to play a crucial role in making moral judgments, e.g. as captured in the Doctrines of Double and of Triple Effect (Hauser, 2006; Mikhail, 2007). Therefore, our results, both analytically and through extensive agent-based simulations, provide important insights into designing of moral agents and machines that are capable recognising others' intentions and taking them into account in their moral decision judgement. A clear implication is that, by virtue of such designs, moral agents in a society will be able to maintain high levels of cooperative behaviours.

Now, conventional wisdom suggests that clear agreements need to be made prior to any collaborative effort in order to avoid potential frustrations for the participants. We have shown (Han et al., 2013a) that this behaviour may actually have been shaped by natural selection, as argued in (Nesse, 2011). Our research demonstrates that reaching prior explicit agreement about the consequences of not honouring a deal provides a more effective road to facilitating cooperation than simply punishing bad behaviour after the fact, even when there is a cost associated to setting up the explicit agreement. Typically, when starting a new project in collaboration with someone else, it pays to establish up-front how strongly your partner is prepared to commit to it. To ascertain the commitment level one can ask for a pledge and stipulate precisely what will happen should the deal not be honoured.

In our study, EGT is used to show that when the cost of arranging commitments (for example, that of hiring a lawyer to make a contract) is justified with respect to the benefit of the joint endeavour (for instance buying a house), and that, when the compensation is set sufficiently high, commitment proposers become prevalent, thence leading to a significant level of cooperation. Commitment proposers can get rid of fake cooperators that agree to cooperate with them yet act differently, thus also avoiding interaction with the bad guys that only aim to exploit the efforts of the cooperative ones. Interestingly, we have shown that whenever the compensation cost reaches a certain threshold (roughly equal the sum of the cost of arrangement commitment plus the benefit of cooperation), no further improvement is achieved by increasing the compensation. This outcome implies that, for regulating legal contracts, it is not required to set extreme penalties for small issues, which might otherwise lead to undesirable side-effects, such as the unwillingness to commit due to the contracts figuring extreme penalties.

More interestingly, our research (Han, et al., 2015a, Han, et al., 2015b) into the synergy of the two presented mechanisms, those of intention recognition and prior commitment, sheds new light on promoting cooperative behaviour. This work employs EGT methods in agent-based computer simulations to investigate mechanisms that underpin cooperation in differently composed societies. High levels of cooperation can be achieved if reliable agreements can be arranged. Formal commitments, such as contracts, promote cooperative social behaviour if they can be sufficiently enforced, and the costs and time to arrange them provide mutual benefit.

On the other hand, an ability to assess intention in others has been demonstrated to play a role in promoting the emergence of cooperation. Indeed, this ability to assess the intentions of others based on experience and observations facilitates cooperative behaviour without resort to formal commitments like contracts. To wit, our research found that the synergy between intention recognition and commitment strongly depends on the confidence and accuracy of the intention recognition capacity. To reach high levels of cooperation, commitments might be unavoidable whenever intentions cannot be assessed with sufficient confidence and accuracy. Otherwise, it is advantageous to wield solely intention recognition so as to avoid the costly arranging of commitments.

4. Combining commitment and costly punishment to prevent antisocial behaviour

We have compared prior commitment with costly posterior punishment, a strategy that makes no prior agreements at all and simply punishes wrongdoers afterwards. Previous studies show that, by punishing bad behaviour strongly enough, cooperation can be promoted in a population of self-interested individuals, see e.g. (Fehr & Gachter, 2002; Han, 2016). Yet these studies also show that the punishment must sometimes be quite excessive in order to obtain significant levels of cooperation. Our own study shows that arranging prior agreements can significantly reduce the impact-to-cost ratio of punishment. Higher levels of cooperation can actually be attained by dint of lower levels of punishment.

More interestingly, through the observation that prior commitment and posterior punishment complement each other, nicely dealing with different types of defective behaviours, we investigated different ways in which these two strategies can be combined. First of all, in (Han and Lenaerts 2016), we have shown that a simple probabilistic combination of the two mechanisms can promote a higher level of cooperation rather than either commitment or punishment alone. It is based on the assessment that arranging prior commitment reduces the effect-to-cost ratio required by costly punishment to perform efficiently, particularly when the cost of arrangement is sufficiently low. While costly punishment can enable one to deal with commitment free-riders, i.e. those who can escape sanctioning when interacting with the commitment strategy simply by avoiding commitment. Our analytical and simulation results show that a combined strategy leads to substantial enhancement in terms of the level of cooperation. Notably, this level is most significant when the punishment cost is sufficiently large and the impact of punishment reaches a threshold. As such, our results have shown that the combined strategy can simultaneously overcome the weaknesses of both strategies.

We have studied another combination approach to exploiting the complementarities of the two mechanisms, in which they are now co-present in the population (Han 2016). Interestingly, it provides a novel solution to prevent antisocial punishment, the one where defectors can punish cooperators, a major challenge in the studies of the evolution of cooperation (Raihani and Bshary, 2015; Power et al., 2011). Namely we have shown, in the context of the one-shot PD, that, if in addition to using punishment the agents in a population can also propose cooperation agreements to their co-players prior to an interaction, then social punishment and cooperation can evolve together, even in the presence of said antisocial punishment. Antisocial punishers can be significantly restrained by commitment proposing agents since only those who dishonour a commitment deal can be enforced to pay compensation. On the other hand, since arranging a commitment deal is costly, its regime can be replaced by social punishers who do not have to pay this cost, while still maintaining cooperation among them. Our results have shown that when both strategic options of commitment and punishment are present, social punishment dominates a population with antisocial punishment players, leading to a significantly higher level of cooperation compared to the cases when either of the strategic options is absent. This is a notable observation since arranging prior commitments, by itself, is already a strong mechanism that can enforce a substantial level of cooperation. By paying the extra cost of commitment for a punishment strategy that was vulnerable to antisocial behaviours and defection, there results a significant improvement in terms of cooperation. That is, the commitment mechanism catalyses the emergence of social punishment and cooperation.

5. Commitments can resolve group cooperation dilemmas. On Avoidance, Restriction, Participation Monitoring, and Delegation

Public goods, like food sharing and social health systems, may prosper when prior agreements to contribute are feasible and all participants commit to do so. Yet, free-riders may exploit such agreements (Han et al., 2013a), thus requiring committers to decide not to enact the public good whenever sufficient others are not attracted to committing. This decision removes all benefits from free-riders (non-contributors), but also from those who are wishing to establish the beneficial resource. In (Han et al., 2014) we show, in the framework of the one-shot Public Goods Game (PGG) and EGT, that implementing measures to delimit benefits to “immoral” free-riders, often leads to more favourable societal outcomes, especially in larger groups and in highly beneficial public goods situations, even if doing so incurs in new costs. PGG is the standard framework for studying emergence of cooperation within group interaction settings (Sigmund, 2010). In a PGG, players meet in groups of a fixed size, and all players can choose whether to cooperate and contribute to the public good or to defect without contributing to it. The total contribution is multiplied by a constant factor and is then equally distributed among all, regardless of whether they have contributed initially. Hence, contributors always gain less than free-riders, thus disincentivizing cooperation. In this scenario, arranging a prior commitment or agreement is an essential ingredient in motivating cooperative behaviour, as abundantly observed both in the natural world (Nesse, 2001) and lab experiments (Cherry and McEvoy, 2013).

In (Han et al., 2014), we extend the PGG to examine commitment-based strategies within group interactions. Prior to playing the PGG, commitment-proposing players ask their co-players to commit to contribute to the PGG, paying a personal proposer’s cost to establish that agreement. If all of the requested co-players accept the commitment, then the proposers assume everyone will contribute. Those who commit yet later do not contribute must compensate the proposers (Han et al., 2013a). As commitment proposers may encounter non-committers, they require strategies to deal with these individuals. Simplest is to not participate in the creation of the common good. Yet, this avoidance strategy, AVOID, also removes benefits for those wishing to establish the public good, creating a moral dilemma. Alternatively, one can establish boundaries on the common good, so that only those who have truly committed have (better) access, or so that the benefit of non-contributors becomes reduced. This is the RESTRICT strategy. Our results lead to two main conclusions: (i) Both strategies can promote the emergence of cooperation in the one-shot PGG whenever the cost of arranging commitment is justified with respect to the benefit of cooperation, thus generalizing results from pairwise interactions (Han et al., 2013a); (ii) RESTRICT, rather than AVOID, leads to more favourable societal outcomes in terms of contribution level, especially when group size and/or the benefit of the PGG increase, even if the cost of restricting is quite large.

In another approach to commitment-based strategic behaviour in the context of the PGG (Han et al., 2017a), we consider a different set of strategies, envisaging that a restriction measure may not always be possible as it is both costly and takes additional effort to implement. Namely, we consider that before engaging in a group venture agents often secure prior commitments from other members of the group and based on the level of participation (i.e. how many group members commit) they can then decide whether or not it is worthwhile to join the group effort (Nesse 2011; Cherry and McEvoy, 2013). This approach is inspired in that many group ventures can be launched only when the majority or a minimum of the participants do commit to contribute to a common good

We have shown that arranging prior commitments while imposing a minimal participation when interacting in groups can help ensure agents' cooperative behaviour. Namely, our results have shown that if the cost of arranging the commitment is sufficiently small compared to the cost of cooperation, commitment-arranging behaviours is frequent, leading thereby to a high level of cooperation in the population. Moreover, an optimal participation level emerges depending both on the dilemma at stake and on the cost of arranging the commitment. The harsher the common good dilemma is, and the costlier it becomes to arrange the commitment, then the more participants should explicitly commit to the agreement to ensure the success of the joint venture

In yet another approach to commitment-based strategic behaviour in the context of the PGG (Han et al., 2017b), we consider that agents can delegate the commitment arrangement and participation monitoring processes in the above-described approaches, to a (beneficiary or non-costly) central authority or institution. The institution may itself benefit from improving the level of cooperation in the population or the social welfare (e.g. public transportation arranged by government, international agreements supported by the UN, crowd-sourcing systems) (Nesse 2011; Cherry and McEvoy, 2013). It may also profit directly from this joint activity by requesting a fee from all committed players in order to provide the service. We have shown that this centralised approach to arranging commitments out-performs the described (personalised) commitment strategy. By having a centralised party to help arrange commitments from the group members instead of leaving it to them to have the initiative, it removes the commitment free-riding issue that prevented the personalized approach to achieve full cooperation (Han et al., 2013a, 2017a). We have shown that the participation level plays a crucial role in the decision of whether an agreement should be formed; namely, it needs to be stricter in the centralized system for the agreement to be formed; however, once it is done right, it is much more beneficial in terms of the level of cooperation as well as the attainable level of social welfare.

6. Why is it so hard to say sorry? Commitments bring about sincerity

When making a mistake, individuals are willing to apologise to secure further cooperation, even if the apology is costly. Similarly, individuals arrange commitments to guarantee that an action such as a cooperative one is in the others' best interest, and thus will be carried out to avoid eventual penalties for commitment failure. Hence, both apology and commitment should proceed side by side in behavioural evolution. In Han et al. (2013b), we studied the relevance of a combination of these two strategies in the context of the IPD. We show that apologising acts are rare in non-committed interactions, especially whenever cooperation is very costly, and that arranging prior commitments can considerably increase the frequency of apologising behaviour. In addition, we have shown that, with or without commitments, apology resolves conflicts only if it is sincere, i.e. costly enough. Most interestingly, our model predicts that individuals tend to use a much costlier apology in committed relationships than otherwise, because it helps better identify free-riders, such as fake committers.

Apology is perhaps the most powerful and ubiquitous mechanism for conflict resolution (Abeler et al., 2010; Ohtsubo & Watanabe, 2009), especially among individuals involved in long-term repeated interactions (such as a marriage). An apology can resolve a conflict without having to additionally involve external parties (e.g. teachers, parents, courts), which may cost all sides of the conflict significantly more. Evidence supporting the usefulness of apology abounds, ranging from medical error situations to seller-customer relationships (Abeler et al., 2010). Apology has been implemented in several computerized systems, such as human-computer interaction and online markets, to facilitate users' positive emotions and cooperation (Tzeng, 2004; Utz et al., 2009).

In (Han et al., 2013b), we describe a model containing strategies that explicitly apologise when making an error between rounds. An apologising act consists in compensating the co-player with an appropriate amount (the higher the sincerer), in order to ensure that this other player cooperates in the next actual round. As such, a population consisting of only apologisers can maintain perfect cooperation. However, other behaviours that exploit this apologetic behaviour could emerge, such as those that accept apology compensation from others but do not apologise when making mistakes (fake apologisers), destroying any benefit of the apology behaviour. Employing EGT (Sigmund, 2010), we show that when the apology occurs in a system where the players first ask for a commitment before engaging in the interaction (Han et al., 2012b, 2012c; Han, 2013), this exploitation can be avoided. Our results lead to these conclusions: (i) Apology alone is insufficient to achieve high levels of cooperation; (ii) Apology supported by prior commitment leads to significantly higher levels of cooperation; (iii) Apology needs to be sincere to function properly, whether in committed relationships or commitment-free ones (which is in accordance with existing experimental studies, e.g. Ohtsubo and Watanabe (2009)); (iv) A much costlier apology tends to be used in committed relationships than in commitment-free ones, as it can help better identify free-riders such as fake apologisers: "*commitments bring about sincerity*".

Our study provides important insights for the design and deployment of apology and commitment mechanisms; for instance, what kind of apology should be provided to customers when mistakes are made, and whether apology can be enhanced if complemented with commitments to ensure cooperation, e.g. compensation for customers who suffer wrongdoing.

7. Apology and forgiveness evolve to resolve failures in cooperative agreements

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 (Martinez-Vaquero et al., 2015, 2017) reveal for the first time under which conditions revenge, apology and forgiveness can evolve, and deal with mistakes within on-going agreements in the context of the IPD. We showed that, when 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, playing an important role in inducing cooperation in repeated dilemmas.

Using methods from EGT, we provide analytical and numerical insight into the viability of commitment strategies in repeated social interactions, modelled by means of the IPD (Axelrod & Hamilton, 1981). In order to study commitment strategies in the IPD, a number of behavioural complexities need to be addressed. First, agreements may end before the recurring interactions are finished. As such, strategies need to consider 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 shown within the context of direct reciprocity (Trivers, 1971), individuals need to deal with mistakes made by an opponent or by themselves, caused for instance by "trembling hands" or "fuzzy minds" (Sigmund, 2010; Nowak, 2006). A decision needs to be made on whether to continue the agreement, or end it collecting the compensation owed from the other's defection.

As errors might lead to misunderstandings or even the 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 (McCullough, 2008; McCullough et al., 2011). The threat of revenge, through some punishment or withholding of a benefit, may discourage interpersonal harm. Yet, often one cannot distinguish with enough certainty if the other's behaviour is intentional or just accidental (Han et al., 2011; Fischbacher & Utikal, 2013). 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 our work, seems to be (costly) apology (McCullough, 2008), a point emphasised in Smith (2008).

Creating agreements and asking others to commit to them provides a basic behavioural mechanism present at all the levels of society, playing a key role in social interactions (Nesse, 2001; Sterelny, 2012; Cherry & McEvoy, 2013). Our work reveals how, when moving to repeated games, the detrimental effect of having a large arrangement cost is moderated, for a subsisting commitment can play its role during several interactions. In these scenarios, the most successful individuals are those who 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 McCullough (2008), and in McCullough et al. (2011). This result is intriguing, since revenge by withholding benefit from the transgressor may lead to a more favourable outcome for cooperative behaviour in the IPD, as opposed to the well-known reciprocal behaviour such as TFT-like strategies. Forgivers only do better when the benefit-to-cost ratio is high enough.

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 proffered in apology has to be high enough (yet not too high), which is also corroborated by recent experimental psychology (McCullough et al., 2014). 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 a commitment with the intention of taking 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. Thus, there is a lower-limit on how sincere apology needs to 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 induce the outbreak of cheating or intolerant behaviour in society (Martinez-Vaquero & Cuesta, 2013, 2014), and only a strict ethics can prevent them (Martinez-Vaquero & Cuesta, 2014), which in our case would be understood as forgiving only when apology is sincere.

Commitments in repeated interaction settings may take the form of loyalty (Schneider & Weber, 2013; Back & Flache, 2008), which is different from our commitments regarding posterior compensations, for we 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 by sincere apology and forgiveness in case mistakes are made.

8. Conclusion

We have summarized results of our own work, reported and surveyed here, on several fundamental facets concerning the emergence and evolution of cooperation in the collective realm, and have provided references to permit following it up in detail.

We have argued how a multiplicity results from our research employing Evolutionary Game Theory (EGT) modelling and experimentation does profitably lead to important insights into machine ethics, such as the design of moral machines, of multi-agent systems, and of contractual algorithms, plus their potential application in human settings too.

Acknowledgements

Thanks are due to co-authors of joint published work cited. Alphabetically: Ari Saptawijaya, Francisco C. Santos, Luis Martinez-Vaquero, and Tom Lenaerts. Pereira acknowledges support from grant FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013. Pereira and Han are supported by Future of Life Institute grant RFP2-154.

References

Abeler, Johannes, Juljana Calaki, Kai Andree, and Christoph Basek. “The power of apology.” *Economics Letters* 107, no. 2 (2010): 233-235.

Axelrod, Robert. *The evolution of cooperation*. Vol. 5145. Basic Books (AZ), 1984.

Hamilton, William D., and Robert Axelrod. “The evolution of cooperation.” *Science* 211, no. 27 (1981): 1390-1396.

Charniak, Eugene, and Robert P. Goldman. “A Bayesian model of plan recognition.” *Artificial Intelligence* 64, no. 1 (1993): 53-79.

Cherry, Todd L., and David M. McEvoy. "Enforcing compliance with environmental agreements in the absence of strong institutions: An experimental analysis." *Environmental and Resource Economics*

Fehr, Ernst, and Simon Gächter. "Altruistic punishment in humans." *Nature* 415, no. 6868 (2002): 137-140.

Fischbacher, Urs, and Verena Utikal. "On the acceptance of apologies." *Games and Economic Behavior* 82 (2013): 592-608.

Han, The Anh. "Intention recognition, commitments and their roles in the evolution of cooperation: from artificial intelligence techniques to evolutionary game theory models." *SAPERE series* 9 (2013).

Han, The Anh. Emergence of social punishment and cooperation through prior commitments. In proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016), pp. 2494-2500, Phoenix, Arizona, USA.

Han, The Anh, and Luis Moniz Pereira. 2013a. "Context-dependent incremental decision making scrutinizing the intentions of others via bayesian network model construction." *Intelligent Decision Technologies* 7, no. 4 (2013): 293-317.

Han, The Anh, and Luis Moniz Pereira. 2013b. "Intention-based decision making via intention recognition and its applications." In *Human Behavior Recognition Technologies: Intelligent Applications for Monitoring and Security*, pp. 174-211. IGI Global, 2013.

Han, The Anh, and Luis Moniz Pereira. 2013c. "State-of-the-art of intention recognition and its use in decision making." *AI Communications* 26, no. 2 (2013): 237-246.

Han, The Anh, and Luis Moniz Pereira. 2018. "Evolutionary Machine Ethics." in: O. Bendel (ed.), *Handbuch Maschinenethik*, Springer, February 2018.

Han The Anh, Pereira, Luis Moniz, and Francisco C. Santos. "Intention recognition promotes the emergence of cooperation." *Adaptive Behavior* 19, no. 4 (2011): 264-279.

Han, The Anh, Pereira, Luis Moniz, and Francisco C. Santos. 2012a. "Corpus-based intention recognition in cooperation dilemmas." *Artificial Life* 18, no. 4 (2012): 365-383.

Han, The Anh, Pereira, Luis Moniz, and Francisco C. Santos. 2012b. "Intention recognition, commitment and the evolution of cooperation." In *Evolutionary Computation (CEC), 2012 IEEE Congress on*, pp. 1-8. IEEE, 2012.

Han, The Anh, Pereira, Luis Moniz, and Francisco C. Santos. 2012c. "The emergence of commitments and cooperation." In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pp. 559-566. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

Han, The Anh, Pereira, Luis Moniz, and Tom Lenaerts. "Avoiding or restricting defectors in public goods games?." *Journal of the Royal Society Interface* 12, no. 103 (2015): 20141203.

Han, The Anh, Pereira, Luis Moniz , Francisco C. Santos, and Tom Lenaerts. 2013a."Good agreements make good friends." *Scientific reports* 3 (2013).

Han, The Anh, Pereira, Luis Moniz , Francisco C. Santos, and Tom Lenaerts. 2013b. "Why is it so hard to say sorry? evolution of apology with commitments in the iterated Prisoner's Dilemma." In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pp. 177-183. AAAI Press, 2013.

Han, The Anh, Pereira, Luis Moniz , Francisco C. Santos, and Tom Lenaerts. 2015a. "Emergence of cooperation via intention recognition, commitment and apology–a research summary." *AI Communications* 28, no. 4 (2015): 709-715.

Han, The Anh, Francisco C. Santos, Tom Lenaerts, and Luis Moniz Pereira. 2015b."Synergy between intention recognition and commitments in cooperation dilemmas." *Scientific reports* 5 (2015).

Han, The Anh, and Tom Lenaerts. “A synergy of costly punishment and commitment in cooperation dilemmas.” *Adaptive Behavior* 24, no. 4 (2016): 237-248.

Han, The Anh, Pereira, Luis Moniz , and Tom Lenaerts. 2017a. “Evolution of commitment and level of participation in public goods games.” *Autonomous Agents and Multi-Agent Systems* 31, no. 3 (2017): 561-583.

Han, The Anh, Luís Moniz Pereira, Luis A. Martinez-Vaquero, and Tom Lenaerts. 2017b.“Centralized vs. Personalized Commitments and their influence on Cooperation in Group Interactions.” In proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI 2017), pp. 2999-3005, San Francisco, USA.

Han, The Anh, Ari Saptawijaya, and Luis Moniz Pereira. “Moral reasoning under uncertainty.” In *International Conference on Logic for Programming Artificial Intelligence and Reasoning*, pp. 212-227. Springer, Berlin, Heidelberg, 2012.

Hauser, Marc. *Moral minds: How nature designed our universal sense of right and wrong*. Ecco/HarperCollins Publishers, 2006.

Martinez-Vaquero, Luis A., and José A. Cuesta. “Evolutionary stability and resistance to cheating in an indirect reciprocity model based on reputation.” *Physical Review E* 87, no. 5 (2013): 052810.

Martinez-Vaquero, Luis A., and José A. Cuesta. “Spreading of intolerance under economic stress: Results from a reputation-based model.” *Physical Review E* 90, no. 2 (2014): 022805.

Martinez-Vaquero, Luis A., Han, The Anh, Pereira, Luis Moniz, and Tom Lenaerts. “Apology and forgiveness evolve to resolve failures in cooperative agreements.” *Scientific reports* 5 (2015).

Martinez-Vaquero, Luis A., Han, The Anh, Pereira, Luis Moniz, and Tom Lenaerts. “When agreement-accepting free-riders are a necessary evil for the evolution of cooperation.” *Scientific Reports* 7 (2017).

McCullough, Michael. *Beyond revenge: The evolution of the forgiveness instinct*. John Wiley & Sons, 2008.

McCullough, Michael E., Robert Kurzban, and Benjamin A. Tabak. “Evolved mechanisms for revenge and forgiveness.” *Understanding and reducing aggression, violence, and their consequences* (2010): 221-239.

McCullough, Michael E., Eric J. Pedersen, Benjamin A. Tabak, and Evan C. Carter. “Conciliatory gestures promote forgiveness and reduce anger in humans.” *Proceedings of the National Academy of Sciences* 111, no. 30 (2014): 11211-11216.

McDermott, Drew V. *Mind and mechanism*. MIT Press, 2001.

Mikhail, J., 2007. Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences*, *11*(4), pp.143-152.

Nesse, Randolph M. “Natural selection and the capacity for subjective commitment.” *Evolution and the capacity for commitment* (2001): 1-44.

Nowak, M.A., 2006. Five rules for the evolution of cooperation. *science*, *314*(5805), pp.1560-1563.

Ohtsubo, Yohsuke, and Esuka Watanabe. “Do sincere apologies need to be costly? Test of a costly signaling model of apology.” *Evolution and Human Behavior* 30, no. 2 (2009): 114-123.

Okamoto, Kyoko, and Shuichi Matsumura. “The evolution of punishment and apology: an iterated prisoner’s dilemma model.” *Evolutionary Ecology* 14, no. 8 (2000): 703-720.

Pereira, Luís Moniz. 2012a. “Turing is among us.” *Journal of Logic and Computation* 22, no. 6 (2012): 1257-1277.

Pereira, Luís Moniz. 2012b. “Evolutionary tolerance.” In *Philosophy and Cognitive Science*, pp. 263-287. Springer, Berlin, Heidelberg, 2012.

Pereira, Luís Moniz and Han The Anh. 2011a. "Intention recognition with evolution prospection and causal bayes networks." In *Computational Intelligence for Engineering Systems*, pp. 1-33. Springer Netherlands, 2011.

Pereira, Luís Moniz, and Ari Saptawijaya. "Modelling morality with prospective logic." *International Journal of Reasoning-based Intelligent Systems* 1, no. 3-4 (2009): 209-221.

Pereira, Luís Moniz, and Ari Saptawijaya. "Bridging two realms of machine ethics." In *Programming Machine Ethics*, pp. 159-165. Springer International Publishing, 2016.

Pereira, Luís Moniz, and Ari Saptawijaya. "Abduction and beyond in logic programming with application to morality." (2015). In: Magnani, L. (Ed.), IfColog Journal of Logics and their Applications, Special issue on Abduction, 3(1):37-71. London: College Publications.

Pereira, Luís Moniz, and Ari Saptawijaya. 2016a. *Programming machine ethics*. Vol. 26. Springer, 2016. ISBN: 978-3-319-29353-0, Berlin: Springer-Verlag.

Pereira, Luís Moniz, and Ari Saptawijaya. 2016b. "Counterfactuals, logic programming and agent morality." In *Applications of Formal Philosophy*, pp. 25-53. Springer, Cham, 2017., ISBN: 978-3319585055, pp. 25-54, Berlin: Springer.

Powers, Simon T., Daniel J. Taylor, and Joanna J. Bryson. "Punishment can promote defection in group-structured populations." *Journal of theoretical biology* 311 (2012): 107-116.

Prinz, Jesse. "Emotions, morality, and identity." *Morality and emotion* (2016): 13-34. ISBN: 978-1-138-12130-0, pp. 83-98. London: Routledge.

Raihani, Nichola J., and Redouan Bshary. "The reputation of punishers." *Trends in ecology & evolution* 30, no. 2 (2015): 98-103.

Sadri, Fariba. "Logic-based approaches to intention recognition." In *Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives*, pp. 346-375. IGI Global, 2011.

Saptawijaya, Ari and Luis Moniz Pereira. "Logic programming applied to machine ethics." In *Portuguese Conference on Artificial Intelligence*, pp. 414-422. Springer, Cham, 2015. LNCS vol. 9273, ISBN 978-3-319-23485-4. Berlin: Springer-Verlag.

Saptawijaya, Ari, and Luis Moniz Pereira. 2015a. "The potential of logic programming as a computational tool to model morality." In *A Construction Manual for Robots' Ethical Systems*, pp. 169-210. Springer International Publishing, 2015.

Saptawijaya, Ari, and Luis Moniz Pereira. 2015b. From Logic Programming to Machine Ethics, in: O. Bendel (Ed.), (2018) *Handbuch Maschinenethik*, Berlin: Springer.

Schneider, Frédéric, and Roberto A. Weber. "Long-term commitment and cooperation." (2013), Tech. Rep., Working Paper Series, University of Zurich, Department of Economics.

Sigmund, Karl. *The calculus of selfishness*. Princeton University Press, 2010.

Smith, Nick. *I was wrong: The meanings of apologies*. Cambridge University Press, 2008.

Sterelny, Kim. *The evolved apprentice*. MIT press, 2012.

Trivers, Robert L. "The evolution of reciprocal altruism." *The Quarterly review of biology* 46, no. 1 (1971): 35-57.

Tzeng, Jeng-Yi. "Toward a more civilized design: studying the effects of computers that apologize." *International Journal of Human-Computer Studies* 61, no. 3 (2004): 319-345.

Utz, Sonja, Uwe Matzat, and Chris Snijders. "On-line reputation systems: The effects of feedback comments and reactions on building and rebuilding trust in on-line auctions." *International Journal of Electronic Commerce* 13, no. 3 (2009): 95-118.

Neumann, J. von. “von, Morgenstern O.” *Theory of games and economic behavior* 1 (1944). Princeton: Princeton University Press.