Software sans Emotions but with Ethical Discernment

Luís Moniz Pereira

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.

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 have addressed, in work reported here, the emergence of cooperation both in the individual and collective realms, sans emotions but with ethical discernment.

1. Introduction

Some of our previous research (Pereira & Saptawijaya, 2011; Han, Saptawijaya, & Pereira, 2012; Pereira & Saptawijaya, 2015a, 2015b; Saptawijaya & Pereira, 2015a; Saptawijaya & Pereira, 2015b) 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, 2015), because this 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, and forgiveness, reinforce the emergence of cooperation in diverse populations, comparatively to their absence, by way of EGT models.

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, 2015b). There are issues concerned with how to bridge the two realms also addressed in this volume (see, for example, Gaspar, 2015). Our account affords plenty of room for an evolutionary phylogenetic emergence of morality, as illustrated below, thereby supplementing the limitations of focusing just on ontogeny. The bridging issues concern individual cognitive abilities and their deployment in the population. Namely the one of recognising the intention of another, even taking into account how others recognize our intention; the abilities of requesting commitment, and of accepting or declining to commit; those of cooperating or defecting; plus those of apologising, be it fostered by guilt, and of taking revenge or forgiving.

This chapter relies mainly on our collective realm research, and considers the modelling of distinct co-present strategies of cooperative and uncooperative behaviour. Such driving strategies are associated with moral "emotions" that motivate moral discernment and substantiate ethical norms, leading to improved general conviviality on occasion, or not. To wit, we can model moral agency without explicitly representing embodied emotions, as we know them. Rather, such software-instantiated "emotions" are modelled as (un)conscious heuristics empowered in complex evolutionary games.

In the next two sections, starting with the ground breaking work of Alan Turing, functionalism is employed to scaffold a philosophical perspective on emotions and morality. The further five sections after those review materials from our EGT-based research in support of this perspective. This work has substantiated the philosophical viewpoint through an admixture of intention recognition, commitment, revenge, apology, and forgiveness. The final section conjectures on guilt, and its relationship with counterfactual reasoning, as a next natural step in our research program.

2. Turing is Among Us

Turing's relevance arises from the timelessness of the issues he tackled, and the innovative light he shed upon them (Pereira, 2012). He first defined the algorithmic limits of computability, via an *effective* well-specified mechanism, and showed the generality of his definition by proving its equivalence to other general, but less algorithmic and non-mechanical, more abstract formulations of computability. His originality lies on the essential simplicity of the mechanism invoked – the now dubbed Turing Machines (or programs), which he called A-Machines – and the proof of existence of a Universal A-Machine (i.e. the digital computer, known in academia as the Universal Turing Machine), which can simulate any other A-Machine, that is, execute any program.

Interestingly, he raised the issue of whether human beings are a measure for his "machines", and, in mechanizing human cognition, Turing implicitly introduced the modern perspective since known as “functionalism”. According to this paradigm, what counts is the realisation of function, independently of the hardware embodying it. Such "multiple realisation" is afforded by the very simplicity of his devised mechanism, relying solely on the manipulation of discrete information, where data and instructions are both represented just with symbols. The twain are stored in memory, instructions doubling as data and as rules for acting – the stored program. To this day, no one has invented a computational mechanical process with such general properties, which cannot be theoretically approximated with arbitrary precision by some A-Machine, where any interactions with the world outside are captured by Turing's innovative concept and definition of "oracle" – the very word employed by him for the purpose –, as a means to interrogate that world by posing queries to one or more outside oracles. This concept of oracle is regularly taught in computer science today, namely in the essential study of computation complexity, though not every student knows it came from Turing. In the midst of a computation a query may be posed to an outside oracle about the satisfaction of some truth, and the computation continued once an answer obtained, rather than the computer testing for an answer in a possibly infinite set of them.

Turing further claimed that his machines could simulate the effect of *any* activity of the mind, not just a mind engaged upon a "definite method of proceeding" or algorithm. He was clear that discrete state machines included those with learning or self-organising abilities, and stressed that these still fall within the scope of the computable. Turing drew attention to the apparent conflict between self-organisation and the definition of A-Machines as having fixed tables of behaviour, but sketched a proof that self-modifying machines are still definable by an unchanged instruction set (Hodges 1997; McDermott 2001).

The promise of this approach in studies of morality is that it represents a universal functionalism, the terms of which enable the bringing together of the ghosts in the several embodied machines (silicon-based, biological, extra-terrestrial or otherwise), to promote their symbiotic epistemic co-evolution, as they undertake moral action within a common moral theatre.

3. Functionalism and Emergence

The principle of the distinction between software and hardware appears clear-cut with the advent of the digital computer and its conceptual precursor, the Universal Turing Machine. The diversity of technologies employed to achieve the same function, confirms it ever since the first computers. One program is executable in physically different machines, precisely because the details of its execution below an ascertainable level of analysis are irrelevant, as long as an identical result at the level of discourse is produced. That said, however, the distinction between hardware and software is not so clear as it might seem. Hardware is not necessarily represented by physical things but rather by what, at some level of analysis, is considered fixed, given, and whose analysis or non-analysability is irrelevant for the purpose at hand. Historically, in the first computers, that level coincided with that of the physical parts of the machine. Subsequently, especially due to rapidly increasing computing power, "hardware" has become increasingly "soft", with the physical basis for the hardware/software distinction finally blurred by the concept of the "abstract machine": a fixed collection of mathematically defined instructions supporting a set of software functions, independently of the particular physical processes underlying the implementation of the abstract machine, that is, realising it.

Hence, "multiple realisation" stands for the thesis that a mental state can be "realised" or "implemented" by different physical states. Beings with different physical constitutions can thus be in the same mental state, and from these common grounds can cooperate, acting in mutual support (or not). According to classical functionalism, multiple realisation implies that psychology is autonomous: in other words, biological facts about the brain are irrelevant (Boden, 2008). Whether physical descriptions of the events subsumed by psychological generalisations have anything in common is irrelevant to the truth of the generalisations, to their interestingness, to their degree of confirmation, or, indeed, to any of their epistemological important properties (Fodor 1974).

Functionalism has continued to flourish, being developed into numerous versions by thinkers as diverse as David Marr, Daniel Dennett, Jerry Fodor, and David Lewis (Fodor 1974; Dennett 2005). It helped lay the foundations for modern cognitive science, being the dominant theory of mind in philosophy today. In the latter part of the 20th and early 21st centuries, functionalism stood as the dominant theory of mental states. It takes mental states out of the realm of the "private" or subjective, and gives them status as entities open to scientific investigation. Functionalism's characterisation of mental states in terms of their roles in the production of behaviour grants them the causal efficacy that common sense takes them to have. In permitting mental states to be multiply realised, functionalism offers an account of mental states compatible with materialism, without limiting the class of minds to creatures with brains like ours (Levin 2010).

Biological evolution is characterized by a set of highly braided processes, which produce a kind of extraordinarily complex combinatorial innovation. A generic term frequently used to describe this vast category of spontaneous, and weakly predictable, order-generating processes, is "emergence". This term became a sort of signal to refer to the paradigms of research sensitive to systemic factors. Complex dynamic systems can spontaneously assume patterns of ordered behaviours not previously imaginable from the properties of their constitutive elements or from their interaction patterns. There is unpredictability in self-organising phenomena – preferably called "evolutionary" (Turing 1950) – with considerably variable levels of complexity, where "complexity" refers to the emergence of collective properties in systems with many interdependent components. These components can be atoms or macromolecules in physical or biological contexts, and people, machines or organisations in socioeconomic contexts.

What does emerge? The answer is not something defined physically but rather something like a shape, pattern, or function. The concept of emergence is applicable to phenomena in which the relational properties predominate over the properties of the compositional elements in the determination of the ensemble's characteristics. Emergence processes are due to starting configurations and interaction topologies, not intrinsic to the components themselves (Deacon 2003). This functionalism is, almost by definition, anti substance-essence, anti vital-principle, anti monopoly of *qualia*.

Building intelligent machines may seek a partial understanding of the emergence of higher-level properties, like morality. Here, functionalism affirms the salience of the results of this work in assessing, for example, human morality. Again, functionalism holds that the material substrate is not of the essence, and that it suffices to realise equivalent functionality albeit by way of a different material vehicle. Moreover, distinct roads to the same behaviour may be had, thereby adding to our understanding of what, say, "general intelligence" or "mind" means. Thus, on our estimation, the most fruitful inquires into the nature of "mind" or "general intelligence" will certainly include the use of Artificial Intelligence aided in time by the embryonic field of artificial emotions, qua strategies, to simulate complex mental operations, as already foreseen (Turing 1950).

4. Learning to recognise intentions and committing resolve cooperation dilemmas

Few problems have motivated the amalgamation of so many seemingly unrelated research fields as has the evolution of cooperation (Nowak, 2006; Sigmund, 2010). Several mechanisms have been identified as catalysers of cooperative behaviour (see survey in Nowak (2006) and Sigmund (2010)). Yet these studies, mostly grounded on evolutionary dynamics and game theory, have neglected the important role, which is played by intention recognition (Han & Pereira, 2013) in behavioural evolution. In our work (Han et al., 2011; Han et al., 2012a), 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 (e.g. win-stay-lose-shift, and tit-for-tat like strategies), and promote a significantly high level of cooperation, even in the presence of noise plus the reduction of fitness associated with the cognitive costs of performing intention recognition. Thus, 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.

Moreover, our recent research (Han, et al., 2015a, Han, et al., 2015b) into the synergy between intention recognition and cooperative 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 co-operation.

An ability to assess the intentions of others based on experience and observations facilitates cooperative behaviour without resort to formal commitments like contracts. Our research found that the synergy between intention recognition and commitment strongly depends on the confidence and accuracy of the intention recognition. To reach high levels of cooperation, commitments may be unavoidable if intentions cannot be assessed with sufficient confidence and accuracy. Otherwise, it is advantageous to wield intention recognition to avoid arranging costly commitments.

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. This research demonstrates that reaching prior explicit agreement about the consequences of not honouring a deal provide 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 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 if the deal is not honoured.

In our study, EGT is used to show that when the cost of arranging commitments (for example, to hire a lawyer to make a contract) is justified with respect to the benefit of the joint endeavour (for instance buying a house), and when the compensation is set sufficiently high, commitment proposers become prevalent, leading to a significant level of cooperation. Commitment proposers can get rid of fake co-operators that agree to cooperate with them yet act differently, also avoiding interaction with the bad guys that only aim to exploit the efforts of the cooperative ones.

But what happens if the cost of arranging the commitments is too high compared to the benefit of cooperation? Would you make a legal contract for sharing a cake? Our results show that in that case those that free ride on the investment of others will "immorally" and inevitably benefit. Establishing costly agreements only makes sense for specific kinds of projects. Our study shows that insisting that your partner share in the cost of setting up a deal leads to even higher levels of cooperation, suggesting the evolution of cooperation for a larger range of arrangement costs and compensations. This makes sense, as equal investment will ensure the credibility of the pledge by both partners. Agreements based on shared costs result in better friends.

We also compared this behaviour with costly punishment, a strategy that does not make any prior agreements and simply punishes afterwards. Previous studies show that by punishing strongly enough bad behaviour cooperation can be promoted in a population of self-interested individuals (Fehr & Gachter, 2002). Yet these studies also show that the punishment must sometimes be quite excessive in order to obtain significant levels of cooperation. Our study shows that arranging prior agreements can significantly reduce the impact-to-cost ratio of punishment. Higher levels of cooperation can be attained through lower levels of punishment. Good agreements make good friends indeed.

5. Emergence of Cooperation in Groups: Avoidance vs. Restriction

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 when 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. Hence, contributors always gain less than free riders, 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). Prior agreements help clarify the intentions and preferences of other players (Han et al., 2012a). Refusing agreements may be conceived as intending or preferring not to cooperate (the non-committers).

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, 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.

6. Why is it so hard to say sorry?

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 go 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 iterated Prisoner's Dilemma (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 show 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 involving in long-term repeated interactions (such as a marriage). An apology can resolve a conflict without having to 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 sellercustomer 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).

The Iterated Prisoner's Dilemma (IPD) has been the standard model to investigate conflict resolution and the problem of the evolution of cooperation in repeated interaction settings (Axelrod, 1984; Sigmund, 2010). The IPD game is usually known as a story of tit-for-tat (TFT), which won both Axelrod's tournaments (Axelrod, 1984). TFT cooperates if the opponent cooperated in the previous round, and defects if the opponent defected. But if there can be erroneous moves due to noise (i.e. an intended move is wrongly performed), the performance of TFT declines, because an erroneous defection by one player leads to a sequence of unilateral cooperation and defection. A generous version of TFT, which sometimes cooperates even if the opponent defected (Nowak & Sigmund, 1992), can deal with noise better, yet not thoroughly. For these TFT-like strategies, apology is modelled implicitly as one or more cooperative acts after a wrongful defection.

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 an appropriate amount (the higher the more sincere), 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. in 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*".

In Artificial Intelligence and Computer Science, apology (Tzeng, 2004; Utz et al., 2009) and commitment (Winikoff, 2007; Wooldridge & Jennings, 1999) have been widely studied, namely how their mechanisms can be formalized, implemented, and used to enhance cooperation in human-computer interactions and online market systems (Tzeng, 2004; Utz et al., 2009), as well as general multi-agent systems (Winikoff, 2007; Wooldridge & Jennings, 1999). Our study provides important insights for the design and deployment of such 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) 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 Iterated Prisoners Dilemma. 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 (Hofbauer & Sigmund, 1998; Sigmund, 2010), we provide analytical and numerical insight into the viability of commitment strategies in repeated social interactions, modelled through the Iterated Prisoners Dilemma (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 take into account how to behave when the agreement is present and when it is absent, on top of proposing, accepting or rejecting such agreements in the first place. Second, as 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 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).

The importance of apology and forgiveness for sustaining long-term relationships has been brought out in different experiments (Abeler et al., 2010; Takaku et al., 2001; Okamoto & Matsumura, 2000; Ohtsubo & Watanabe, 2009). Apology and forgiveness is of interest as they remove the interference of external institutions (which can be quite costly to all parties involved), in order to ensure cooperation.

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 for 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 as revenge by withholding the 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 offered in the 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 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 whenever mistakes are made.

Ohtsubo's experiment (Ohtsubo & Watanabe, 2009) shows that a costlier apology is better at communicating sincerity, and as a consequence will be more often forgiven. This observation is shown to be valid across cultures (Takaku et al., 2001). In another laboratory experiment (Fischbacher & Utikal, 2013), the authors showed apologies work because they can help reveal the intention behind a wrongdoer's preceding offence. In compliance with this observation, in our model, apology best serves those who intended to cooperate but defect by mistake.

Despite the fact that "to err is human" (Pope, 1711), our research results demonstrate that behaviours like revenge and forgiveness can evolve to cope with mistakes, even when they occur at high rates. Complicating matters is that mistakes are not necessarily intentional, and that even if they are then it might still be worthwhile to continue a mutually beneficial agreement. Here, a sincerity threshold exists whereby the cost of apologising should exceed the cost of cooperating if the encouragement of cooperation is the goal.

8. Future Work: Emotional and Counterfactual Guilt

A natural extension of our work on intention recognition, commitment, revenge, apology, and forgiveness involves adding guilt, shame, and confession with surplus apology. We leave shame alone for now as it involves reputation, which we did not address above so as to concentrate on the more basic model of pairwise interactions, without the intrusion of reputational hearsay. Though both have ostensibly evolved to promote cooperation, we believe that guilt and shame can be treated separately. Guilt is an inward phenomenon that can foster apology, and even spontaneous public confession. Shame is inherently public, and it too may lead to apology and request for forgiveness. Shame, however, hinges on being caught, on failing to deceive, and on a mechanism being in place that lets one fall into disrepute.

From an evolutionary viewpoint, guilt is envisaged as an in-built mechanism that tends to prevent wrong doing because of internal suffering that pressures an agent to confess when wrongs are enacted, alongside a costlier apology and penance, plus an expectation of forgiveness to alleviate or dispel the guilt-induced suffering.

The hypothesis, consequently, is that the emergence of guilt within a population is evolutionarily advantageous as it represents an extra-costly apology compared to a non-guilty one, enacted as it is in order to decrease the added suffering. We can test this hypothesis by adapting our existing model comprising commitment, revenge, apology, and forgiveness, via piggybacking guilt onto it. To do so, one introduces a present/absent guilt parameter such that, on defection by a guilt-ridden player, not only is thereby increased the probability of apology (confession), but also the player spontaneously pays a costlier apology, as a means to atone internal guilt. On the other hand, the co-player will more readily accept a guilty extra-valued apology, and forgive. In addition, this co-player's attitude, if copied, will contribute to favour his own forgiveness by others in the population, in case his own super-apologetic confession of guilt replaces of the standard one in the absence of acknowledged guilt. The prediction is that guilt will facilitate and speed-up the emergence of cooperation, in spite of its heavier cost. One reason behind this prediction is that costs of cooperation are compensated for by the costlier guilt apology paid by others. Another reason is that it is in general more conducive to forgiveness, especially in the border cases where the standard apology is outright insufficient.

We know that guilt is alleviated by private confession, e.g. to a priest or psychotherapist, with cost in prayers or fees, plus the renunciation of past failings. In the context of our research, such ersatz confessions and atonements, precisely by exacting a cost, should render temptation to defect less probable – a preference reversal (Correia, 2015) – with the proceeds appertained to some common good (e.g. in a Public Goods Game, or like through charity).

In summary, future research will attempt to show, by simulation if not analytically, that guilt naturally connects with apology and forgiveness mechanisms because of its emergent evolutionary advantage. It seems not too difficult to incorporate into the present framework, by splitting each strategy into one variant experiencing guilt in case of defection, plus a guiltless one. The population at the start would now contain, instead, an admixture of all of both types, for a given fixed cost and extra cost of guilty apology, plus the usual other parameters, namely a forgiveness threshold. The prediction again is that guilt is evolutionarily advantageous, within a range of the overall parameters defining a starting population composition, via EGT evolution with the usual social imitation of strategies with high payoff success.

This further opens the way to treatment of emotions modelled as strategies, guilt being a widely acknowledged one. It should show that one does not need a specific kind of body (namely an anthropomorphic one) for guilt to serve the role of a moral emotion, useful as it is in population settings where moral cooperation attains good value for all regardless of means of embodiment.

Finally, counterfactual reasoning (Byrne, 2007; Collins et al., 2004; Pereira & Saptawijaya, 2015a) could be wielded to prime and tune guilt. Presupposing that the agent can reason counterfactually, e.g. given the by now known sequence of plays by its co-players it might reason: "Had I before felt guilty instead, and played according to such guilt, then I would have fared better." As a consequence, the player would then meta-reflectively (Mendonça, 2015) modify its "feeling level" of guilt for the future.

One could envisage the whole of our above approach as purveying a form of fiction, though recognisably a rather abstract one, yet still adumbrated as per the "Moral Feelings from Rocky Fictional Ground" (John, 2015), the next chapter in this volume. Indeed, our abstract mathematical and computational fictional simulations might be construed and stretched to fit a bill whereby such fiction would not necessarily offer theorists of emotion or morality immediate embodied evidence, as in novels, say. In contradistinction, it can possibly offer interesting, challenging and conjectural ideas that might benefit the theorising in these domains. A computer scientist friend bemusedly jokes about my "soap opera" research, what with intention recognition, commitment proposal, defection, guilt, apology, forgiveness, revenge...

Acknowledgements

Profound thanks are due to my co-authors of joint published work herein cited, without which the personal summing up and specific philosophical viewpoint above would not have been possible at all. Alphabetically: Ari Saptawijaya, Francisco C. Santos, Luis Martinez-Vaquero, The Anh Han, and Tom Lenaerts. Moreover, the author thanks Ari Saptawijaya, Sara Silva and Jeffrey White for comments on prior drafts, and the support from grant FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013.

References

Abeler, J., Calaki, J., Andree, K., & Basek, C. (2010). The power of apology. *Economics Letters, 107(2)*, 233-235.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, R. & Hamilton, W. D. (1981). The evolution of cooperation. *Science* 211, 1390–1396.

Back, I. & Flache, A. (2008). The Adaptive Rationality of Interpersonal Commitment. *Rationionality and Society* 20, 65–83.

Boden, M. A. (2008). Information and Cognitive Science. *Philosophy of Information*. P. Adriaans & J. van Bentham (eds.), pp. 741-761. Amsterdam: North-Holland, Elsevier.

Byrne, R. M. J (2007). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge MA, MIT Press, 2007.

Cherry, T. L., & McEvoy, D. M. (2013). Enforcing compliance with environmental agreements in the absence of strong institutions: An experimental analysis. *Environmental and Resource Economics*, 54(1):63–77.

Collins, J., Hall, N. & L. A. Paul, L. A. (Eds.). *Causation and Counterfactuals*. Cambridge MA, MIT Press, 2004.

Correia, V. (2015). Weakness of Will and Self-control: the role of emotions in impulsive behaviour. In this volume.

Deacon, T. W. (2003). The Hierarchic Logic of Emergence: Untangling the Interdependence of Evolution and Self-Organization. *Evolution and Learning: The Baldwin Effect Reconsidered*. H. W. Weber, D. J. Depew (eds.), pp. 273-308. Cambridge: The MIT Press.

Dennett, D. C. (2005). *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*. Cambridge: The MIT Press.

Fehr, E. & Gachter, S. (2002). Altruistic punishment in humans. *Nature* 415: 137–140.

Fischbacher, U., & Utikal, V. (2013). On the acceptance of apologies. *Games and Economic Behavior, 82*, 592 – 608.

Fodor, J. A. (1974). Special Sciences, or the Disunity of Science as a Working Hypothesis. *Synthèse,* 28: 77-115.

Gaspar, A. (2015). Morality and Empathy *vs* Empathy and Morality: a quest for the source of goodness in phylogenetic and ontogenetic contexts. In this volume.

Han, T. A. (2013). Intention Recognition, Commitments and Their Roles in the Evolution of Cooperation: From Artificial Intelligence Techniques to Evolutionary Game Theory Models. *SAPERE series*, 9. Berlin: Springer-Verlag.

Han, T. A., & Pereira, L. M. (2013). State-of-the-art of intention recognition and its use in decision making. AI Communication, 26, 237–246.

Han, T. A., Pereira, L. M., & Santos, F. C. (2011). Intention recognition promotes the emergence of co-operation. Adaptive Behavior, 19, 264–279.

Han, T. A., Pereira, L. M., and Santos, F. C. (2012a). Corpus-based intention recognition in cooperation dilemmas. *Artificial Life*, 18(4):365–383.

Han, T. A., Pereira, L. M., & Santos, F. C. (2012b). Intention Recognition, Commitment, and The Evolution of Cooperation. In *Proceedings of IEEE Congress on Evolutionary Computation* (pp. 1–8). IEEE Press John Wiley & Sons, Hoboken, NJ, USA.

Han, T. A., Pereira, L. M., & Santos, F. C. (2012c). The emergence of commitments and cooperation. In *Proceedings of the Eleventh International Conference on Autonomous Agents and Multiagent Systems* (pp. 559-566). International Foundation for Autonomous Agents and Multiagent Systems.

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

Han, T. A., Pereira, L. M., Santos, F. C., and Lenaerts, T. (2013a). Good agreements make good friends. doi:10.1038/srep02695, *Scientific Reports*, 3(2695).

Han, T. A., Pereira, L. M., Santos, F. C., and Lenearts, T. (2013b). Why Is It So Hard to Say Sorry: The Evolution of Apology with Commitments in the Iterated Prisoner's Dilemma. In *Proceedings of the 23nd* *international joint conference on Artificial intelligence (IJCAI'2013)*. AAAI Press, Palo Alto, CA, USA, 2013.

Han, T. A., Pereira, L. M., Santos, F. C., and Lenearts, T. (2015a). Emergence of Cooperation via Intention Recognition, Commitment, and Apology -- A Research Summary, *AI Communications*, doi:10.3233/AIC-150672, vol. 28, preprint online June 2015.

Han, T. A., Santos, F. C., T. Lenaerts, T., and Pereira, L. M. (2015b). Synergy between intention recognition and commitments in cooperation dilemmas, doi:10.1038/srep09312, *Nature Scientific Reports*, *Sci. Rep.* 5:9312.

Han, T. A., Saptawijaya, A., & Pereira, L. M. (2012). Moral reasoning under uncertainty. In N. Bjørner, & A. Voronkov (Eds.), *Proceedings of the Eighteenth International Conference on Logic for Programming Artificial Intelligence and Reasoning (LNCS)* (Vol. 7180, pp. 212-227). Berlin: Springer-Verlag.

Hodges, A. (1997). *Alan Turing: one of The Great Philosophers*. London: Phoenix.

Hofbauer, J., & Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. New York, NY: Cambridge University Press.

John, E. (2015). Moral Feelings from Rocky Fictional Ground. In this volume.

Levin, J. (2010). Functionalism. *The Stanford Encyclopaedia of Philosophy*, E.N. Zalta (ed.), http://plato.stanford.edu/archives/sum2010/entries/functionalism/

Martinez-Vaquero, L. A. & Cuesta, J. A. (2013). Evolutionary stability and resistance to cheating in an indirect reciprocity model based on reputation. *Physical Reviews E* 87, 052810.

Martinez-Vaquero, L. A. & Cuesta, J. A. (2014). Spreading of intolerance under economic stress: Results from a reputation-based model. *Phys. Rev. E* 90, 022805.

Martinez-Vaquero, L. A., Han, T. A., Pereira, L. M., & Lenaerts, T. (2015). Apology and Forgiveness Evolve to Resolve Failures in Cooperative Agreements. doi:10.1038/srep10639, *Nature Scientific Reports, Sci. Rep. 5:10639*.

Mendonça, D. (2015). Emotions and Akratic Feelings - Insights into Morality through Emotions. In this volume.

McCullough, M. E. (2008). Beyond Revenge, the evolution of the forgiveness instinct. Jossey-Bass, San Francisco, CA, USA.

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

McCullough, M. E., Pedersen, E. J., Tabak, B. A. & Carter, E. C. (2014). Conciliatory gestures promote forgiveness and reduce anger in humans. *Proceedings of The National Academy of Sciences U. S. A. (PNAS)* 111, 11211–11216.

McDermott, D. (2001). *Mind and Mechanism*. Cambridge: The MIT Press.

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

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science, 314(5805)*, 1560-1563. doi: 10.1126/science.1133755.

Nowak, M. A., and Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature, 355*, 250–253.

Ohtsubo, Y., & Watanabe, E. (2009). Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior, 30(2)*, 114–123.

Okamoto, K. & Matsumura, S. (2000). The evolution of punishment and apology: an iterated prisoner's dilemma model. *Evolutionary Ecology* 14, 703–720.

Pereira, L. M. (2012). Turing is Among Us. *Journal of Logic and Computation, 22(6)*, 1257-1277.

Pereira, L. M., & Saptawijaya, A. (2011). Modelling Morality with Prospective Logic. In M. Anderson and S. L. Anderson (Eds.), *Machine Ethics* (pp. 398-421). New York, NY: Cambridge University Press.

Pereira, L. M., & Saptawijaya, A. (2015a). Abduction and Beyond in Logic Programming with Application to Morality. In *IfCoLog Journal of Logics and Their Applications (Special Issue on "Frontiers of Abduction"*. London: College Publications. (To appear.)

Pereira, L. M., & Saptawijaya, A. (2015b). Bridging Two Realms of Machine Ethics. In J. White, & R. Searl (Eds.), *Rethinking Machine Ethics in the Age of Ubiquitous Technology* (pp. 197-224), Hershey, PA: IGI Global.

Pope, A. (1711). *An Essay on Criticism, part II.* W. Lewis, Russel Street, Covent Garden.

Prinz, J. (2015). Emotions, Morality, and Identity: An Empirical Approach. In this volume.

Saptawijaya, A., & Pereira, L. M. (2015a). Logic Programming Applied to Machine Ethics. In Pereira, F., Machado, P., Costa, E., Cardoso, A. (Eds.), *Proceedings of the Seventeenth Portuguese Intl. Conf. on Artificial Intelligence* (6 pages). LNCS vol. 9273, ISBN 978-3-319-23485-4. Berlin: Springer-Verlag.

Saptawijaya, A., & Pereira, L. M. (2015b). The Potential of Logic Programming as a Computational Tool to Model Morality. In R. Trappl (Ed.), *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations* (40 pages). Lecture Notes in Cognitive Technologies ISBN 978-3-319-21548-8. Berlin: Springer-Verlag.

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

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

Smith, N. (2008). *I was wrong: The meanings of apologies, vol. 8*. New York, NY: Cambridge University Press New York.

Sterelny, K. (2012). *The evolved apprentice*. MIT Press.

Takaku, S., Weiner, B. & Ohbuchi, K. (2001). A cross-cultural examination of the effects of apology and perspective taking on forgiveness. *Journal of Language and Social Psychology* 20, 144–166.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology* 46: 35–57.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind* 59:433-460.

Tzeng, J.-Y. (2004). Toward a more civilized design: studying the effects of computers that apologize. *International Journal of Human-Computer Studies, 61(3)*, 319 – 345.

Utz, S., Matzat, U., & Snijders, C. (2009). 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(3)*, 95–118.

Winikoff, M. (2007). Implementing commitment-based interactions. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems* (pp. 868–875), Association for Computing Machinery (ACM), New York, NY, USA.

Wooldridge, M., & Jennings, N. R. (1999). The cooperative problem-solving process. *Journal of Logic and Computation, 9(4)*, 563-592.