Social Manifestation of Guilt Leads to Stable Cooperation in Multi-Agent Systems

Luís Moniz Pereira
NOVA LINCS, Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
lmp@fct.unl.pt

Tom Lenaerts
Département d'Informatique, Université Libre de Bruxelles & Computer Science Department, Vrije Universiteit Brussel, Boulevard du Triomphe CP212, 1050 Brussels, Belgium
Tom.Lenaerts@ulb.ac.be

Luis A. Martinez-Vaquero
National Research Council of Italy, via San Martino della Battaglia 44, 00185 Rome, Italy
l.martinez.vaquero@gmail.com

The Anh Han
School of Computing, Teesside University, Borough Road, Middlesbrough, TS1 3BA, UK
T.Han@tees.ac.uk

ABSTRACT

Inspired by psychological and evolutionary studies, we present here theoretical models wherein agents have the potential to express guilt with the ambition to study the role of this emotion in the promotion of pro-social behaviour. To achieve this goal, analytical and numerical methods from evolutionary game theory are employed to identify the conditions for which enhanced cooperation emerges within the context of the iterated prisoners dilemma. Guilt is modelled explicitly as two features, i.e. a counter that keeps track of the number of transgressions and a threshold that dictates when alleviation (through for instance apology and self-punishment) is required for an emotional agent. Such an alleviation introduces an effect on the payoff of the agent experiencing guilt. We show that when the system consists of agents that resolve their guilt without considering the co-player's attitude towards guilt alleviation then cooperation does not emerge. In that case those guilt prone agents are easily dominated by agents expressing no guilt or having no incentive to alleviate the guilt they experience. When, on the other hand, the guilt prone focal agent requires that guilt only needs to be alleviated when guilt alleviation is also manifested by a defecting co-player, then cooperation may thrive. This observation remains consistent for a generalised model as is discussed in this article. In summary, our analysis provides important insights into the design of multi-agent and cognitive agent systems where the inclusion of guilt modelling can improve agents' cooperative behaviour and overall benefit.

Keywords

Guilt Emotion; Multi-Agent Systems; Evolution of Cooperation; Evolutionary Game Theory

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

1. INTRODUCTION

> "...what do you think, if a person does something very bad, do they have to be punished?"..."You know the reason I think they should be punished?"..."It's because of how bad they are going to feel, in themselves. Even if nobody did see them and nobody ever knew. If you do something very bad and you are not punished you feel worse, and feel far worse, than if you are."
>
> [In page 55 of "The love of a Good Woman", by Alice Munro (Nobel Prize in Literature 2013), in "Family Furnishings - Selected Stories, 1995-2014", Vintage Intl. Edition, 2015.]

Interest in machine ethics has significantly increased in recent years [42, 34]. One pertinent theme within that context addresses the computational modelling of human emotions [29] like guilt. Guilt is defined in the online Merriam-Webster dictionary as "The feeling of culpability especially for imagined offences or a sense of inadequacy" [13], which implies that guilt follows from introspection: An individual experiencing guilt will detect this emotional state, and can act upon it. Frank argued that guilt may provide a useful mechanism, if operationalised properly, to miminise social conflict and promote cooperation [9]. Notwithstanding the importance of this emotion for the evolution of cooperation, no in-depth numerical or analytical models have been provided to confirm or refute the hypothesis that this emotion has evolved to ensure stable social relationships.

Evolutionary work related to guilt has been addressed in [16, 30, 36]. These works focus on behaviours following from feeling guilty, i.e. apology and forgiveness, but do not however explicitly model guilt as a behavioural feature of each agent, which is the goal of the current work. In [16, 30], apology occurs within the context of long-term commitment behaviour, which assumes that prior agreements are made before the iterated Prisoner's Dilemma (IPD) is started and a compensation is given once the agreement is broken and

there is no apology. Moreover, apology leads to a payoff advantage for the co-player, which as we show here is not required for the evolutionary dynamics of guilt behaviour. Modelling guilt explicitly will allow us to go beyond this work and explore different aspects like the cumulative effect of wrongdoings or using anticipation to decide what to do in the context of a specific guilt level. Moreover, it will provide insight into when the computational modelling of emotions actually is relevant for autonomous agents in MAS, for instance, to enforce social norms or implement cognitive agent systems, as was argued in [29, 37, 45].

A non-evolutionary mechanistic model of guilt has been put forth in [1], within the context of lethal actions by semi-autonomous robots in the battlefield. A scalar value of guilt is increased, if an untoward lethal action is taken by the robot, whether by a human operator or by the robot itself (if it so recognises its mistake). When a predefined guilt threshold is reached, the robot is henceforth inhibited from deploying further lethal actions.

This behavioural quantification of guilt provides us with a basis to define our evolving agents (see Background section for psychological as wells as evolutionary theories on which this is based): Guilt is part of an agent's representation or *genotype*, i.e. they will all be equipped with a guilt threshold G, with $G \in [0, +\infty]$, and a transient guilt level, g ($g \geq 0$). Initially g is set to 0 for every agent. If an agent feels guilty after an action that she considers as wrong, then the agent's g is increased (by 1). When g reaches the agent's guilt threshold, i.e. $g \geq G$, the agent can (or not) act to alleviate her current guilt level. We assume here that guilt alleviation can be achieved through a sincere apology to the co-player or, otherwise, through self-punishment if it is not possible to apologise [12, 3]. Different from prior work [16, 30], we do not assume here that apology leads to a benefit for the co-player, considering it only as an honest signal of the experiencing of guilt. In general, the cost of guilt alleviation is modelled by a so-called *guilt cost* γ ($\gamma \geq 0$). Whenever the agent punishes herself, by paying γ, g is decreased (by 1). Using this genotype definition, one can imagine different types of agents with different G thresholds, such as those who never feel guilty (unemotional, with $G = +\infty$) or those who are very emotional, feeling guilty immediately after a wrongdoing (with $G = 0$).

The objective of this work is to show that agents expressing this emotion, despite the disadvantage of the costly guilt-alleviation acts, are evolutionary viable, can dominate agents not expressing the emotion and that they induce sustained social interactions, all of which will be shown in the context of the iterated Prisoner's Dilemma (IPD). To set the stage for future work we first focus on two extreme behaviours, i.e. $G = 0$ and $G = +\infty$, as will be explained in more detail in the Models & Methods section. Afterwards these results are generalised to situations where $G > 0$ yet less than the number of rounds in the IPD, since when G is larger this would correspond to $G = +\infty$. We use a stochastic evolutionary model incorporating frequency-dependent selection and mutation to identify when agents with guilt are evolutionary stable [38]. More importantly, we will show that for guilt to be evolutionary viable, a guilt prone player under focus should be reactive to the guilt-driven behaviour of its co-player: If this other party is not behaving properly and/or does not show guilt-alleviating behaviour then the focal agent's guilt is alleviated automatically or even non-existing. Pure self-punishment without social considerations will not allow for guilt to evolve at the individual level. In this sense, our work contrasts with for instance that of Gadou et al. [11] which takes an utilitarian perspective to model the behaviour resulting from guilt, not by introducing self-punishment but by introducing a guilt aversion level term into a player's utility function, which ignores the social role of guilt [9]. From a multi-agent perspective, considering socio-technical systems including autonomous agents, our results confirm that decision making conflicts can be reduced when including emotions to guide participants to socially acceptable behaviours.

2. BACKGROUND

The realisation of working computational models inspired by psychological theories, such as we strive to do here, permits scientific advances by forcing concrete if simple evolvable models, which reveal hidden assumptions and allow for empirical experimentation with dynamically pliable programmed artefacts [29]. Computational models of artificial emotion can play an important decision-making and control scheduling role in designing multi-agent autonomous systems [26]. Moreover, herein we reap inspiration from anthropological arguments specifically about the character, usage and evolutionary role of the emotion of guilt towards the enhancement of cooperation amongst autonomous agents, and propound how such evolution can be modelled by Evolutionary Game Theory (EGT) computational models.

Psychology conceives of shame and guilt as belonging to the family of self-conscious emotions [27, 7, 40], invoked through self-reflection and self-evaluation. Though both may have evolved to promote social relationships, guilt and shame can be treated separately. Guilt is an inward private phenomenon, though it can promote apology, and even spontaneous public confession. Shame is inherently public, though it may too lead to apology and to the request for forgiveness [39]. It hinges on being caught, failing to deceive, and the existence of a reputation mechanism. Guilt is also more directly associated with morality than shame [35]: It is closely associated with the idea of conscience, as an internal guide informing us when an action is wrong. Moreover, it is widely regarded as a fundamentally social emotion, which plays a positive prosocial role [9]. It arises especially when there is a threat of separation or exclusion. Guilt is an unpleasant emotion and, when experienced, people try to get rid of it: the most common coping strategies are confession, reparation, self-criticism, and punishment.

Guilt acts not only *a posteriori*, but functions as well *a priori*, preventing harm by wishing to avoid guilt and the necessity to alleviate it. For, guilt being unpleasant, people may resist doing things when anticipating feeling guilty about them. People will obey rules to avoid feeling guilty when breaking them. Anticipation of guilt leads to norm conformity even when retaliation won't arise (as when we might get away with being free-riders). Anticipating our own guilt can defeat the temptation to engage in harmful behaviour. If the cost of guilt was removed, then norm conformity might drop off dramatically. This lesson follows from the empirical research [35].

Evolutionary theorists studying guilt have argued that anticipatory guilt promotes cooperative behaviour by adding an emotional cost to defection [44, 9]. The emotion of anticipated guilt may function to deter the temptation to betray

a friend to reap a short-term gain because of the long-term cost of a lost friendship.

In many social dilemmas like the Prisoner's Dilemma (PD), defection is the dominant strategy: defectors do better than cooperators regardless of whether their trading partners defect or cooperate. But this fact makes it rational for both parties to defect, even though mutual defection is worse than mutual cooperation in many of these games. Trivers [44] speculated that mutual evolution has promoted the emergence of guilt because it makes defection less attractive. People may gain materially from defecting, but guilt makes them suffer emotionally, and that leads them to cooperate. Frank [9] noted that this tendency is so deeply engrained that people avoid defection even in cases where the other party is not a likely partner in future exchanges.

Both Trivers and Frank assume that guilt is the result of biological evolution. But it is equally possible that guilt emerged under cultural pressure as a tool for ensuring that people cooperate. If discovery of theft is inevitable, a set of emotional, psychological, and behavioural mechanisms should be activated, including genuine feelings and confessions of guilt and remorse to appease the victim and victim's kin, verbal attempts to exculpate oneself from blame, and the return of or reparations for stolen property [2].

Rephrasing [23]:

> To the extent we decide guilt is an innate mechanism we must conclude that humans have an innate capacity to judge certain actions to be transgressions of endorsed normative frameworks, meriting reparative or punitive response. If foreseen guilt prevents harm and absence of harm prevents possible retaliation and/or loss of reputation, then it would seem that a priori guilt would be evolutionarily advantageous. A posteriori guilt, on the other hand, would be evolutionarily advantageous because conducive to increased amount/possibility of apology, and we've seen apology is advantageous. Also apology reduces the pain of guilt.

Evolutionarily, guilt is envisaged as an in-built mechanism that tends to prevent wrong doing because of the internal self suffering it creates, and, should the wrong doing have taken place, it puts internal pressure on confession (admitting the wrong) and follow-up costly apology and penance, plus an expectation of forgiveness, so as to alleviate and dispel the internal suffering produced by guilt [6, 41]. Behavioural experiments on the iterated Prisoner's Dilemma and Ultimatum games have shown that guilt is a key factor in increasing cooperation among players [25].

Next, common sense stresses that feeling guilt for harm done to another makes sense only if you perceive the other is not attempting to harm you too. War is a case in point. Hence, recognising whether such is the case should be taken into account in any model of guilt. In fact, in our very first guilt model below, where recognising the intention of another is not considered [15, 17, 14], goes to show that feeling guilty about our defections regardless of what others feel about their defections is self-defeating. This view is in line with the well-adopted appraisal theories of emotion, in which emotions cannot be explained by solely focusing on the environment or by solely focusing on the individual; rather, they reflect the person-environment relationship [29].

Our work also contrasts with the model in [28], in that instead of self-punishment in their case an agent's payoff is simply reduced if it violates some norm's fairness with respect to a measure of the agent's sensitivity.

Finally, it is important to note the rich literature of computational modelling of guilt in AI and MAS literature [29, 5, 8, 37, 18, 45, 4]. But in contrast to our aim and approach, these studies aim to formalise guilt as part of a MAS, such as virtual agent and cognitive agent systems, for the purpose of regulating social norms (see survey in [4]) or improving agent decision making and reasoning processes [45]. However, our results and approach provide novel insights into the design of such MAS systems; for instance, if agents are equipped with the capacity of guilt feeling even if it might lead to costly disadvantage, that drives the system to an overall more cooperative outcome where they are willing to take reparative actions after wrongdoings.

3. MODELS AND METHODS

Considering the foregoing, an attempt to introduce guilt in EGT models of cooperation seems unavoidable. The issue concerning guilt within such models is whether its presence is more worthwhile than its absence, with respect to an advantageous emergence of cooperation. One can introduce guilt explicitly in models to show that it is worthwhile, in further support of its appearance on the evolutionary scene. Indeed, one may focus on emotions, like guilt, as being strategies in abstract evolutionary population games, sans specific embodiment nor subjective *quale* [33].

3.1 Iterated Prisoner's Dilemma

Social interactions are modeled in this article as symmetric two-player games defined by the payoff matrix

$$\begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} R,R & S,T \\ T,S & P,P \end{pmatrix} \end{array}$$

A player who chooses to cooperate (C) with someone who defects (D) receives the sucker's payoff S, whereas the defecting player gains the temptation to defect, T. Mutual cooperation (resp., defection) yields the reward R (resp., punishment P) for both players. Depending on the ordering of these four payoffs, different social dilemmas arise [20, 38]. Namely, in this work we are concerned with the PD, where $T > R > P > S$. In a single round, it is always best to defect, because less risky, but cooperation may be rewarding if the game is repeated. In IPD, it is also required that mutual cooperation is preferred over an equal probability of unilateral cooperation and defection ($2R > T + S$); otherwise alternating between cooperation and defection would lead to a higher payoff than mutual cooperation. The PD is repeated for a number of rounds, where the number of rounds is modelled by Ω.

3.2 Guilt modelling in IPD

Starting from the definition of the agent-based guilt feature in the Introduction, we will focus in the current work only on two basic types of (extreme) guilt thresholds (a more generalised model for non-extreme guilt levels shall be analysed in Section 4.3):

- $G = +\infty$: In this type of agents the guilt level g will never reach the threshold no mater how many

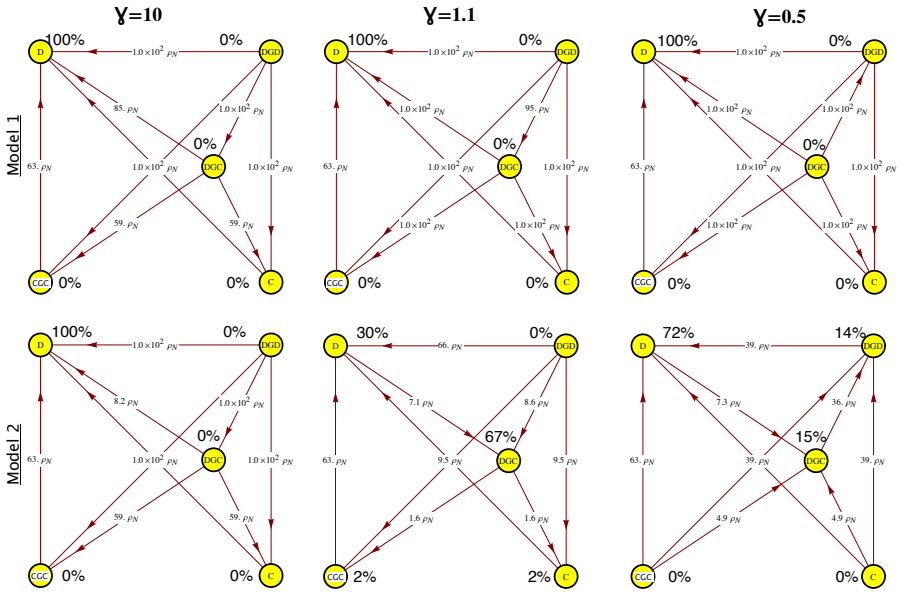

Figure 1: Stationary distribution and transition probabilities directions for the two models (top row for model 1 and bottom row for model 2) and for various values of γ (first, second, and third columns correspond to $\gamma =$ 10, 1.1, 0.5, respectively). The arrows identify the transitions that are stronger than neutral, annotated with the corresponding transition probability. In the first model, D always dominates the population, regardless of the value of γ. In the second model, for an intermediate value of γ, DGC performs well against defective strategies, taking over their population. When this value is too large, DGC is dominated by D (who does not feel guilty) while when this value is too small, DGC is dominated by DGD (who feels guilty when defecting but still can benefit more from the PD). Parameters: $T = 2,\ R = 1,\ P = 0,\ S = -1$; $\beta = 1$; $N = 100$; $\Omega = 10$.

times they defect; hence, they never need to reduce g, and consequently never pay the guilt cost γ. In other words, this type of agents experiences no guilt feeling. They are dubbed (guilt-)unemotional agents.

- $G = 0$: whenever this type of agents defects, it becomes true that $g > G$; hence, the agents need to act immediately to reduce g, thus paying γ. In other words, this type of agents always feel guilty after a wrongdoing, viz. defection. They are dubbed (guilt-) emotional agents.

Besides the guilt threshold, an agent's strategy is described by what she plays in a PD (C or D) and, when the agent's ongoing guilt level g reaches the threshold G, by whether the agent changes her behaviour from D to C. Hence, there are five possible strategies, labeled as follows

1. Unemotional cooperator (C): always cooperates, unemotional (i.e. $G = +\infty$)
2. Unemotional defector (D): always defects, unemotional (i.e. $G = +\infty$)
3. Emotional cooperator (CGC): always cooperates, emotional (i.e. $G = 0$)
4. Emotional non-adaptive defector (DGD): always defects, feels guilty after a wrongdoing (i.e. $G = 0$), but does not change behaviour.
5. Emotional adaptive defector (DGC): defects initially, feels guilty after a wrongdoing (i.e. $G = 0$), and changes behaviour from D to C.

In order to understand when guilt can emerge and promote cooperation, our EGT modelling study below analyses whether and when emotional strategies, i.e. those with $G = 0$, can actually overcome the disadvantage of the incurred costs or fitness reduction associated with the guilt feeling and its alleviation, and in consequence disseminate throughout the population. Namely, in the following we aim to show that, in order to evolve, guilt alleviation through self-punishment can only be evolutionarily viable when only the focal agent misbehaves. In other words, an emotional guilt-based response only makes sense when the other is not attempting to harm you too. To that purpose, we analyse two different models, which differ in the way guilt influences the preferences of the focal agents, where the preferences are determined by the payoffs in the matrices (1) and (2).

In the first model, an agent's ongoing guilt level g increases whenever the agent defects, regardless of what the co-player does. The payoff matrix for the five strategies C, D, CGC, DGD, and DGC, can be written as follows

$$\begin{array}{c} \\ C \\ D \\ CGC \\ DGD \\ DGC \end{array} \begin{array}{c} \begin{array}{ccccc} C & D & CGC & DGD & DGC \end{array} \\ \begin{pmatrix} R & S & R & S & \frac{S+R\Theta}{\Omega} \\ T & P & T & P & \frac{P+T\Theta}{\Omega} \\ R & S & R & S & \frac{S+R\Theta}{\Omega} \\ T-\gamma & P-\gamma & T-\gamma & P-\gamma & \frac{P+T\Theta}{\Omega}-\gamma \\ \frac{T-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S\Theta}{\Omega} & \frac{T-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S(\Theta)}{\Omega} & \frac{P-\gamma+R(\Theta)}{\Omega} \end{pmatrix} \end{array}, \quad (1)$$

where we use $\Theta = \Omega - 1$ just for the purpose of a neater representation. Note that the actions C and CGC are essentially equivalent; both considered for the sake of completeness of the strategies set.

In the second model, an agent feels guilty when defecting if the co-player acted pro-socially or was observed to feel guilty after defection, viz. through exercising self-punishment or apologising. Thus in this second model, guilt has a particular social aspect that is missing from the first model. In particular, DGC does not change behaviour to C if the co-player played D and did not try to alleviate her guilt as a result of her bad behaviour. Now, the payoff matrix is rewritten as follows:

$$\begin{array}{c} \\ C \\ D \\ CGC \\ DGD \\ DGC \end{array} \begin{array}{c} \begin{array}{ccccc} C & D & CGC & DGD & DGC \end{array} \\ \begin{pmatrix} R & S & R & S & \frac{S+R\Theta}{\Omega} \\ T & P & T & P & P \\ R & S & R & S & \frac{S+R\Theta}{\Omega} \\ T-\gamma & P & T-\gamma & P-\gamma & \frac{P+T\Theta}{\Omega}-\gamma \\ \frac{T-\gamma+R\Theta}{\Omega} & P & \frac{T-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S\Theta}{\Omega} & \frac{P-\gamma+R\Theta}{\Omega} \end{pmatrix} \end{array}. \quad (2)$$

Notice the differences in the payoff matrices for the interactions between the emotional strategies that defect, i.e. DGD and DGC, and the unemotional defector D. Note also that these payoff matrices are conceptually different from those used in the situation where commitment and costly apology are used (see [30]): in the current work, apology does not induce a benefit for the co-player.

3.3 Evolutionary Dynamics in Finite Populations

Our analysis of the two models above is based on EGT methods for finite populations [32, 22]. In such a setting, individuals' payoff represents their *fitness* or social *success*, and evolutionary dynamics is shaped by social learning [20, 38], whereby the most successful individuals will tend to be imitated more often by the others. In the current work, social learning is modelled using the so-called pairwise comparison rule [43], assuming that an individual A with fitness f_A adopts the strategy of another individual B with fitness f_B with probability given by the Fermi function, $\left(1+e^{-\beta(f_B-f_A)}\right)^{-1}$. The parameter β represents the 'imitation strength' or 'intensity of selection', i.e., how strongly the individuals base their decision to imitate on fitness comparison. For $\beta=0$, we obtain the limit of neutral drift – the imitation decision is random. For large β, imitation becomes increasingly deterministic.

In the absence of strategy mutations or exploration, the end states of evolution are inevitably monomorphic: once such a state is reached, it cannot be escaped through imitation. We thus further assume that, with a certain mutation probability, an individual switches randomly to a different strategy without imitating another individual. In the limit of small mutation rates, the behavioural dynamics can be conveniently described by a Markov Chain, where each state represents a monomorphic population, whereas the transition probabilities are given by the fixation probability of a single mutant [10, 22, 19]. The resulting Markov Chain has a stationary distribution, which characterises the average time the population spends in each of these monomorphic end states.

Let N be the size of the population. Suppose there are at most two strategies in the population, say, k individuals using strategy A ($0 \leq k \leq N$) and $(N-k)$ individuals using strategy B. Thus, the (average) payoff of the individual that

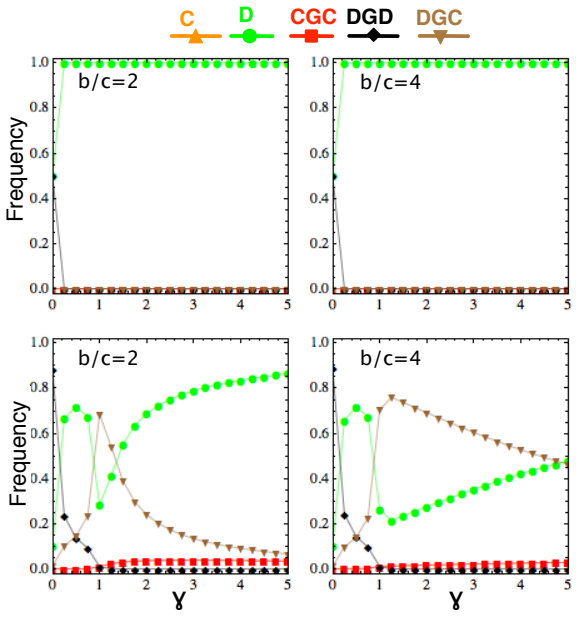

Figure 2: Frequency of each strategy as a function of the guilt cost, γ, for the two models (top row for model 1 and bottom row for model 2), and for different benefit-to-cost ratios b/c. In the first model, D always dominates the population. In the second model, for an intermediate value of γ, DGC is the most frequent strategy; but when it is too small or too large, DGD is dominant. Parameters: $\beta=1$; $N=100$; $\Omega=10$.

uses A and uses B can be written as follows, respectively,

$$\begin{aligned} \Pi_A(k) &= \frac{(k-1)\pi_{A,A}+(N-k)\pi_{A,B}}{N-1}, \\ \Pi_B(k) &= \frac{k\pi_{B,A}+(N-k-1)\pi_{B,B}}{N-1}, \end{aligned} \quad (3)$$

where $\pi_{X,Y}$ stands for the payoff an individual using strategy X obtained in an interaction with another individual using strategy Y.

Now, the probability to change the number k of individuals using strategy A by ± 1 in each time step can be written as

$$T^{\pm}(k) = \frac{N-k}{N}\frac{k}{N}\left[1+e^{\mp\beta[\Pi_A(k)-\Pi_B(k)]}\right]^{-1}. \quad (4)$$

The fixation probability of a single mutant with a strategy A in a population of $(N-1)$ individuals using B is given by [43, 10]

$$\rho_{B,A} = \left(1+\sum_{i=1}^{N-1}\prod_{j=1}^{i}\frac{T^-(j)}{T^+(j)}\right)^{-1}. \quad (5)$$

In the limit of neutral selection (i.e. $\beta=0$), $\rho_{B,A}$ equals the inverse of population size, $1/N$.

Considering a set $\{1,...,q\}$ of different strategies, these fixation probabilities determine a transition matrix $M=\{T_{ij}\}_{i,j=1}^{q}$, with $T_{ij,j\neq i}=\rho_{ji}/(q-1)$ and $T_{ii}=1-\sum_{j=1,j\neq i}^{q}T_{ij}$, of a Markov Chain. The normalized eigenvector associated with the eigenvalue 1 of the transposed of M provides the stationary distribution described above [10, 22], describing the relative time the population spends adopting each of the strategies.

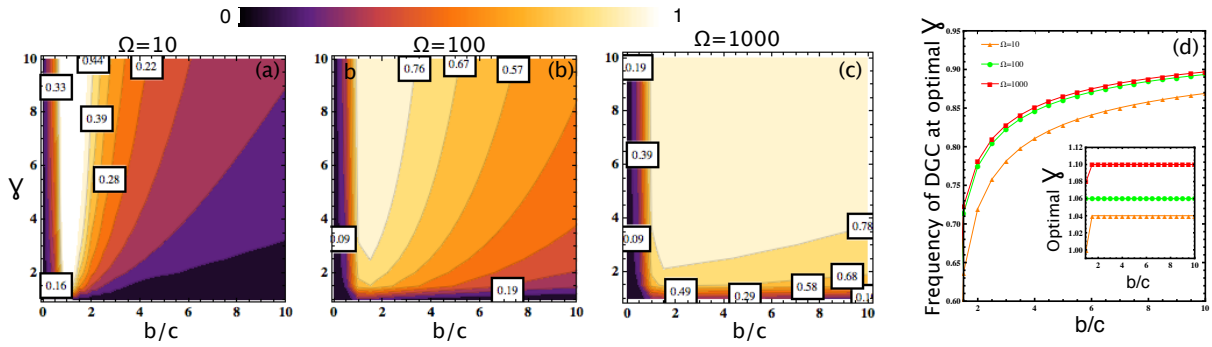

Figure 3: Panels a-c: Frequency of DGC in Model 2 as a function of γ and b/c, for different values of Ω. In all cases, DGC performs best for an intermediate value of γ, reaching its highest frequency. In addition, DGC is more frequent for larger Ω. Panel d: Frequency of DGC at optimal value of γ (outer panel) and the optimal value of γ itself (inner panel), both as a function of b/c, and for different values of Ω. In all cases, the frequency of DGC increases with b/c. It also slightly increases with Ω. The optimal value of γ increases with Ω (see inner panel). For a fixed value of Ω, it does not depend on b/c when this ratio is sufficiently high. Parameters: $\beta = 1$; $N = 100$.

3.4 Analytical condition for risk-dominance

An important criteria for pairwise comparison of strategies in finite population dynamics is *risk-dominance*, that is, whether it is more probable for an A mutant fixating in a homogeneous population of individuals adopting B than a B mutant fixating in a homogeneous population of individuals adopting A. When the first is more likely than the latter (i.e. $\rho_{B,A} > \rho_{A,B}$), A is said to be *risk-dominant* against B [24, 31], which holds for any intensity of selection and in the limit of large N when

$$\pi_{A,A} + \pi_{A,B} > \pi_{B,A} + \pi_{B,B}. \tag{6}$$

4. RESULTS

This section starts by deriving analytical conditions for when DGC can be a viable strategy, being risk-dominant when playing against defective strategies (i.e. D and DGD). We show that this strategy is always dominated by defective strategies in the first model, while there is a wide range of parameters in which it dominates both defective strategies in the second model, resulting in high levels of cooperation therein. We then provide numerical simulation results to support the analytical observations. Furthermore, the results are generalised to consider non-radical guilt modelling (i.e. $0 < G < \infty$) , showing that the obtained results are robust beyond the context of radical guilt strategies.

4.1 Analytical conditions for risk-dominance of DGC against other strategies

To begin with, using inequality (6), we examine whether and under which conditions DGC is risk-dominant against other strategies. In the first model defined by Matrix (1), DGC is always dominated by D, which follows from:

$$\frac{P - \gamma + S(\Omega - 1)}{\Omega} < P \text{ (since } S < P \text{ and } \gamma > 0)$$

and

$$\frac{P - \gamma + R(\Omega - 1)}{\Omega} < \frac{P + T(\Omega - 1)}{\Omega} \text{ (since } R < T \text{ and } \gamma > 0).$$

To the contrary, in the model defined by Matrix (2), DGC is risk-dominant against D and DGD, respectively, when

$$\gamma < (\Omega - 1)(R - P) \tag{7}$$

and

$$\gamma > \frac{T + P - R - S}{2}. \tag{8}$$

These two inequalities indicate that the guilt cost γ needs to be within certain limits for DGC to be dominant. To make this more understandable one can simplify the PD to a Donation game [38] — a famous special case of the PD: $T = b$, $R = b - c$, $P = 0$, $S = -c$, satisfying that $b > c > 0$, where b and c stand respectively for benefit and cost of cooperation. We thus obtain

$$c < \gamma < (\Omega - 1)(b - c). \tag{9}$$

That is, in order for DGC to be a viable strategy against defective strategies (i.e. D and DGD), the cost γ needs to exceed the cost of cooperation c as otherwise it will be exploited by DGD players who do not change to cooperation when feeling guilty (and having to pay only a small cost); and, on the other hand, it cannot be too large (less than the benefit of mutual cooperation obtained in the rounds after alleviating guilt) as otherwise it will be dominated by D players who never pay the guilt cost after defecting.

This result shows that there is some intermediate value of γ that would lead to an optimal performance of DGC, i.e. reaching the highest frequency. Moreover, inequality (9) also implies that for a fixed cost of cooperation c, the range in which DGC can outperform defective strategies is larger for less harsh PD (i.e. with a larger benefit-to-cost ratio, b/c) and when the PD is repeated longer (i.e. the larger Ω) as the cost γ used to sustain a (long-term) relationship is more beneficial.

4.2 Numerical results for the radical guilt emotions

Fig. 1 shows stationary distributions and transitions probabilities directions among strategies in the two models. In the first model, D always dominates the population. In the second model, for an intermediate value of γ, DGC performs well against defective strategies, taking over their population. When this value is too large, DGC is dominated by D

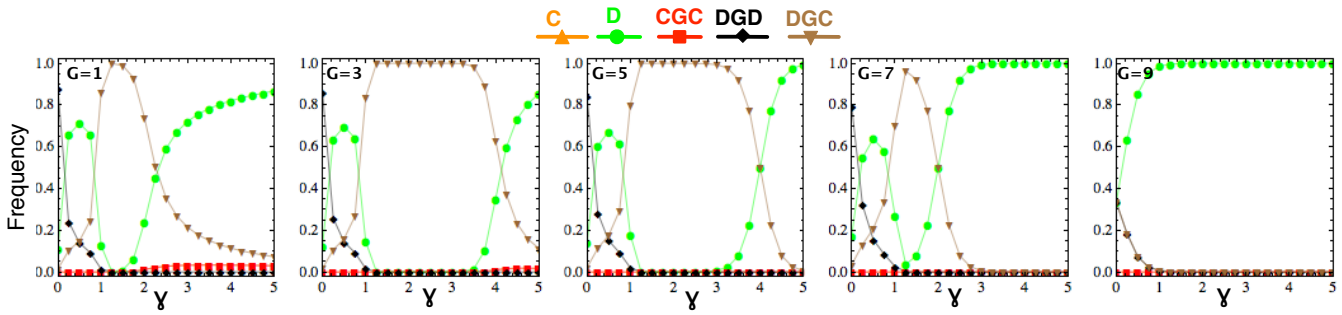

Figure 4: Frequency of each strategy as a function of the guilt cost, γ, and for different values of G, as well as the cooperation level in the population (dotted lines with label C). Parameters: $\beta = 1$; $N = 100$; $\Omega = 10$.

(who does not feel guilty) while when this value is too small, DGC is dominated by DGD (who feels guilty when defecting but still can benefit more from the PD). The results are in accordance with the analytical conditions above, noting that the risk-dominance conditions in inequalities (7) and (8) for the current PD payoff matrix can be simplified just to $1 < \gamma < 9$.

These observations are robust for varying γ, as can be seen in Fig. 2, which shows the frequency of each strategy in the two models as a function of γ. In the first model, D always dominates, leading to no cooperation in the population. In the second model, for an intermediate value of γ, DGC is the dominant strategy. The range in which DGC is the most frequent strategy is larger for a less harsh PD (i.e. larger benefit-to-cost ratio, b/c, comparing left and right columns: namely, the range is roughly $[1.0, 1.5]$ for $b/c = 2$ and $[1.0, 5.0]$ for $b/c = 4$). When γ is too large, D dominates while when this value is too small, DGD is dominant as it can exploit DGC better than D (paying a small cost for self-punishment, like in pretending to feel guilty).

Next we focus on analysing the second model as it is clear by now that guilt can emerge only when guilt-capable players take into account whether their co-players are observed to express similar emotions immediately prior to or after a wrongdoing. Furthermore, to understand when guilt enhances cooperation we can focus on the DGC strategy as it is the main strategy that generates cooperation in the population. Indeed, in Fig. 3 (panels a, b and c) we compute the frequency of DGC in the second model as a function of

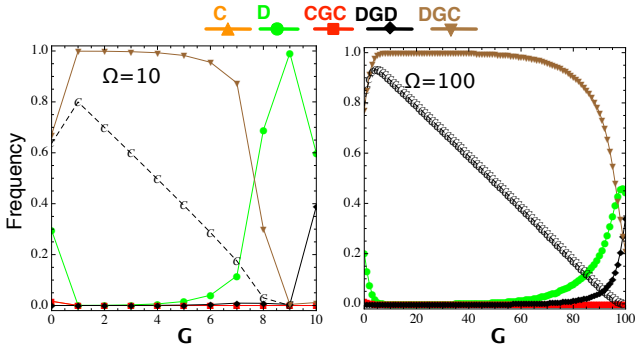

Figure 5: Frequency of each strategy and the actual cooperation level, as a function of G, and for different values of Ω. Parameters: $\beta = 1$; $N = 100$; $\gamma = 1.1$.

γ and b/c, and for different values of Ω. In general, DGC performs best for an intermediate value of γ (whenever b/c is not too small), reaching its highest frequency. In addition, comparing the three panels we can see that DGC is more frequent for larger Ω. This is because for larger Ω the guilt cost γ is better justified, which is in accordance with the analytical conditions above.

Furthermore, in panel d of this figure we show the frequency of DGC (in the second model) at optimal value of γ and the optimal value of γ itself, both as a function of b/c, and for different values of Ω. In all cases, the frequency of DGC increases with b/c and also slightly increases with Ω. The optimal value of γ increases with Ω but, most interestingly, when considering a concrete value of Ω, it does not depend on b/c when this ratio is sufficiently high (namely when ≥ 1.5). This independence suggests a way to measure the optimal value of γ for a given Ω, although future work will need to check analytically whether this independence holds in general and what is the threshold of b/c (above which the independence holds).

4.3 Numerical results for non-radical guilt emotions

Above we analysed the competition of the radical guilt-emotional strategy ($G = 0$) against the unemotional one ($G = \infty$). We now consider that the emotional strategy is not extreme, i.e. $G > 0$. Also note that if $G \geq \Omega$ it would be equivalent to $G = \infty$ as players would never feel guilty within the Ω rounds of the IPD. Therefore we assume that $G < \Omega$. The new payoff matrix is the previous one plus the following (noting the same payoff matrix as in the second model is reproduced by substituting $G = 0$)

$$\begin{array}{c} \\ C \\ D \\ CGC \\ DGD \\ DGC \end{array} \begin{array}{c} \begin{array}{ccccc} C & D & CGC & DGD & DGC \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & \frac{G(S-R)}{\Omega} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{G(S-R)}{\Omega} \\ \frac{G\gamma}{\Omega} & 0 & \frac{G\gamma}{\Omega} & \frac{G\gamma}{\Omega} & \frac{G(P-T+\gamma)}{\Omega} \\ \frac{G(T-R)}{\Omega} & 0 & \frac{G(T-R)}{\Omega} & \frac{G(P-S)}{\Omega} & \frac{G(P-R)}{\Omega} \end{pmatrix} \end{array}. \quad (10)$$

Let us consider for example DGD playing against C. Now DGD keeps defecting and only at the $(G+1)$-th round it feels guilty and pays γ. The rest of the rounds would be the same as before (DGD keeps defecting and paying γ without changing behaviour. Hence, its payoff against C is

$$\begin{aligned} \Pi_{DGD,C} &= \frac{1}{\Omega}\left((G+1)T - \gamma + (\Omega - G - 1)(T - \gamma)\right) \\ &= T - \gamma + \frac{G\gamma}{\Omega}. \end{aligned}$$

Similarly for DGC playing against C, just that now it is the case that after the $(G+1)$-th round this type of players would keep cooperating for the rest of the IPD. Hence,

$$\Pi_{DGC,C} = \frac{1}{\Omega}\left((G+1)T - \gamma + (\Omega - G - 1)R\right) = \frac{T - \gamma + R(\Omega - 1)}{\Omega} + \frac{G(T-R)}{\Omega}.$$

Numerical results in Fig. 4 show that for not too large values of G, DGC dominates the population. Moreover, as before, we observe that this strategy performs best for an intermediate value of γ, reaching its highest frequency. When G is too large, D players dominate the population. We obtain a similar conditions for which DGC is risk-dominant against the defective strategies (D and DGD)

$$\frac{T + P - R - S}{2} < \gamma < (\Omega - G - 1)(R - P) \qquad (11)$$

Now the range is smaller for larger value of G.

Next, we also plot the actual cooperation level in the population, which is obtained as the total of the frequency of C, CGC and $(\Omega - G - 1)/\Omega \times$ the frequency of DGC (as this type of players only cooperates with each other after the first $G+1$ rounds of the IPD). Similarly to DGC, the frequency of cooperation reaches its highest possible frequency for intermediate values of γ.

These observations become clearer by looking at Fig. 5, where we show the strategies frequency and the total cooperation level as a function of G. In general, DGC reaches a high frequency for a wide range of intermediate G. However, given that the larger G, the more rounds it takes for DGC players to start cooperating with each other (both feel guilty and exhibit alleviation acts after $G+1$ rounds), the value of G leading to the optimal level of cooperation is the smallest G that provides (close to) the highest frequency of DGC. Next, comparing the two panels in Fig. 5 we observe that a higher level of cooperation is achieved for larger Ω.

5. CONCLUSIONS AND FUTURE WORKS

On the basis of psychological and evolutionary understandings of guilt, and inspired by these, this paper proffers and studies, for the first time, two analytical models of guilt, within a system of multi-agents adopting a combination of diverse guilty and non-guilty strategies. To do so, it employs the methods and techniques of EGT, in order to identify the conditions when there does emerge an enhanced cooperation, improving on the case when guilt is absent.

Guilt, depending on an agent's strategy, may result in self-punishment, with effect on fitness, and on a change in behaviour. In the first model of guilt, a guilt prone agent is insensitive to whether the co-player also feels guilt on defection. This model does not afford cooperation enhancement because guilt prone agents are then free-ridden by non-guilt prone ones. In our second model, guilt is not triggered in an agent sensitive to the defecting co-player not experiencing guilt too. It is this latter model that shows the improvement on cooperation brought about by the existence of guilt in the population, and how it becomes pervasive through the usual EGT phenomenon of social imitation. Another successful variation of this model allows to stipulate guilt accumulation coupled with a triggering threshold.

Our results provide important insights for the design of self-organised and distributed MAS: if agents are equipped with the capacity for guilt feeling even if it might appear to lead to disadvantage, that drives the system to an overall more cooperative outcome wherein agents become willing to take reparative actions after wrongdoings.

In future research, the model shall be complicated via our existing EGT models comprising apology, revenge, and forgiveness, by piggybacking guilt onto them [33, 16, 30].

In the IPD and other models of cooperation, players judge others by their actions: whether they cooperate or defect. However, we not only care about whether others cooperate, but also about their decision-making process: We place more trust in cooperators who never even considered defecting. To quote Kant, "In law a man is guilty when he violates the rights of others. In ethics he is guilty if he only thinks of doing so." [21]. Hence, detecting another's proclivity to cheat, albeit checked by guilt, allots intention recognition an important role to play even when the intention is not carried out [15, 17, 14].

Last but not least: Currently we only consider one type of emotional strategy playing against unemotional strategy. It is possible that strategies with multiple guilt threshold are co-present in the population. We envisage that different types might dominate in different game configurations, which we will analyse in future work.

Acknowledgments

LMP acknowledges support from FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013. LAMV and TL from Fonds voor Wetenschappelijk Onderzoek - FWO through grant nr. G.0391.13N. TL also from Fondation de la Recherche Scientifique - FNRS through grant FRFC nr. 2.4614.12. TAH from Teesside URF funding (11200174).

REFERENCES

[1] R. C. Arkin and P. Ulan. An ethical adaptor: Behavioral modification derived from moral emotions. Technical Report GIT-GVU-09-04, College of Computing at the Georgia Institute of Technology, Atlanta, GA, 30332, 2009.

[2] K. Asao and D. M. Buss. The tripartite theory of machiavellian morality: Judgment, influence, and conscience as distinct moral adaptations. In *The Evolution of Morality*, pages 3–25. Springer, 2016.

[3] B. Brown. Face saving and face restoration in negotiation. In D. Druckman, editor, *Negotiations: Social-Psychological Perspectives*, pages 275–300. SAGE Publications, 1977.

[4] N. Criado, E. Argente, and V. Botti. Open issues for normative multi-agent systems. *AI Communications*, 24(3):233–264, 2011.

[5] C. M. De Melo, P. Carnevale, S. Read, D. Antos, and J. Gratch. Bayesian model of the social effects of emotion in decision-making in multiagent systems. In *AAMAS'2012*, pages 55–62, 2012.

[6] D. M. Fessler and K. J. Haley. The strategy of affect: Emotions in human cooperation. *The Genetic and Cultural Evolution of Cooperation, P. Hammerstein, ed*, pages 7–36, 2003.

[7] K. Fischer and J. Tangney. Self-conscious emotions and the affect revolution: Framework and introduction. *Self-conscious emotions: Shame, guilt,*

embarrassment, and pride. New York: Guilford.

[8] J. Fix, C. von Scheve, and D. Moldt. Emotion-based norm enforcement and maintenance in multi-agent systems: Foundations and petri net modeling. In *AAMAS '06*, pages 105–107. ACM, 2006.

[9] R. H. Frank. *Passions Within Reason: The Strategic Role of the Emotions*. Norton and Company, 1988.

[10] D. Fudenberg and L. A. Imhof. Imitation processes with small mutations. *Journal of Economic Theory*, 131:251–262, 2005.

[11] B. Gaudou, E. Lorini, and E. Mayor. Moral guilt: An agent-based model analysis. In *Advances in Social Simulation*, volume 229 of *Advances in Intelligent Systems and Computing*, pages 95–106. Springer, 2014.

[12] E. Goffman. *Interaction Ritual: : essays in face-to-face behavior*. Random House, 1967.

[13] "guilt". *The Merriam-Webster Dictionary*. Merriam-Webster.com, retrieved on 8 November 2016.

[14] T. A. Han. *Intention Recognition, Commitments and Their Roles in the Evolution of Cooperation: From Artificial Intelligence Techniques to Evolutionary Game Theory Models*, volume 9. Springer, 2013.

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

[16] T. A. Han, L. M. Pereira, F. C. Santos, and T. Lenaerts. Why Is It So Hard to Say Sorry: The Evolution of Apology with Commitments in the Iterated Prisoner's Dilemma. In *IJCAI'2013*, pages 177–183, 2013.

[17] T. A. Han, F. C. Santos, T. Lenaerts, and L. M. Pereira. Synergy between intention recognition and commitments in cooperation dilemmas. *Scientific reports*, 5(9312), 2015.

[18] T. A. Han, A. Saptawijaya, and L. M. Pereira. Moral reasoning under uncertainty. In *Proceedings of the 18th International Conference on Logic for Programming, Artificial Intelligence and Reasoning (LPAR-18)*, pages 212–227. Springer LNAI 7180, 2012.

[19] C. Hauert, A. Traulsen, H. Brandt, M. A. Nowak, and K. Sigmund. Via freedom to coercion: The emergence of costly punishment. *Science*, 316:1905–1907, 2007.

[20] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

[21] M. Hoffman, E. Yoeli, and C. D. Navarrete. Game theory and morality. In *The Evolution of Morality*, pages 289–316. Springer, 2016.

[22] L. A. Imhof, D. Fudenberg, and M. A. Nowak. Evolutionary cycles of cooperation and defection. *Proc. Natl. Acad. Sci. U.S.A.*, 102:10797–10800, 2005.

[23] R. Joyce. *The evolution of morality*. MIT press, 2007.

[24] M. Kandori, G. J. Mailath, and R. Rob. Learning, mutation, and long run equilibria in games. *Econometrica*, 61:29–56, 1993.

[25] T. Ketelaar and W. Tung Au. The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition & Emotion*, 17(3):429–453, 2003.

[26] Z. Kowalczuk and M. Czubenko. Computational approaches to modeling artificial emotion – an overview of the proposed solutions. *Frontiers in Robotics and AI*, 19 April 2016.

[27] M. Lewis. Thinking and feeling: The elephant's tail. In M. Schwebel, C. A. Maher, and N. S. Fagley, editors, *Promoting cognitive growth over the life span*, pages 89–110. Lawrence Erlbaum Associates, Inc, 1990.

[28] E. Lorini and R. Mühlenbernd. The long-term benefits of following fairness norms: A game-theoretic analysis. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 301–318. Springer, 2015.

[29] S. Marsella and J. Gratch. Computationally modeling human emotion. *Communications of the ACM*, 57(12):56–67, 2014.

[30] L. A. Martinez-Vaquero, T. A. Han, L. M. Pereira, and T. Lenaerts. Apology and forgiveness evolve to resolve failures in cooperative agreements. *Scientific reports*, 5(10639), 2015.

[31] M. A. Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560, 2006.

[32] M. A. Nowak, A. Sasaki, C. Taylor, and D. Fudenberg. Emergence of cooperation and evolutionary stability in finite populations. *Nature*, 428:646–650, 2004.

[33] L. M. Pereira. Software sans emotions but with ethical discernment. In S. G. D. D. Silva, editor, *Morality and Emotion: (Un)conscious Journey to Being*, pages 83–98. Routledge, 2016.

[34] L. M. Pereira and A. Saptawijaya. *Programming Machine Ethics*, volume 26 of *SAPERE series*. Springer, 2016.

[35] J. J. Prinz and S. B. Nichols. Moral emotions. In *Oxford University Press*, 2010.

[36] S. Rosenstock and C. O'Connor. When it's good to feel bad: Evolutionary models of guilt and apology. *Philosophy of Science*, 64(6):637–658, 2016.

[37] B. T. R. Savarimuthu, M. Purvis, and M. Purvis. Social norm emergence in virtual agent societies. In *AAMAS '08*, pages 1521–1524, 2008.

[38] K. Sigmund. *The Calculus of Selfishness*. Princeton University Press, 2010.

[39] N. Smith. *I was wrong: The meanings of apologies*. Cambridge University Press, 2008.

[40] J. P. Tangney and R. L. Dearing. *Shame and guilt*. Guilford Press, 2003.

[41] J. P. Tangney, J. Stuewig, E. T. Malouf, and K. Youman. 23 communicative functions of shame and guilt. *Cooperation and Its Evolution*, page 485, 2013.

[42] TheEconomist. March of the machines - a special report on artificial intelligence, June 25, 2016.

[43] A. Traulsen, M. A. Nowak, and J. M. Pacheco. Stochastic dynamics of invasion and fixation. *Phys. Rev. E*, 74:11909, 2006.

[44] R. L. Trivers. The evolution of reciprocal altruism. *Quaterly Review of Biology*, 46:35–57, 1971.

[45] P. Turrini, J.-J. C. Meyer, and C. Castelfranchi. Coping with shame and sense of guilt: a dynamic logic account. *Autonomous Agents and Multi-Agent Systems*, 20(3):401–420, 2010.