Co-evolution of social and non-social guilt in structured populations

Theodor Cimpeanu
Teesside University
Middlesbrough, United Kingdom
theodorcimpeanu@outlook.com

Luís Moniz Pereira
Universidade Nova de Lisboa
Caparica, Lisbon
lmp@fct.unl.pt

The Anh Han
Teesside University
Middlesbrough, United Kingdom
t.han@tees.ac.uk

ABSTRACT

Building ethical machines may involve bestowing upon them the emotional capacity to self-evaluate and repent on their actions. While reparative measures, such as apologies, are often considered as possible strategic interactions, the explicit evolution of the emotion of guilt as a behavioural phenotype is not yet well understood. Here, we study the co-evolution of social and non-social guilt of homogeneous or heterogeneous populations, including well-mixed, lattice and scale-free networks. Social guilt comes at a cost, as it requires agents to make demanding efforts to observe and understand others, while non-social guilt only requires the awareness of the agents' own state and hence incurs no social cost. Those choosing to be non-social are however more sensitive to exploitation by other agents due to their social unawareness. Resorting to methods from evolutionary game theory, we study analytically, and through extensive numerical and agent-based simulations, whether such social and non-social guilt can evolve, depending on the underlying structure of the populations or systems of agents. The results show that, in both lattice and scale-free networks, emotional guilt prone strategies are dominant for a larger range of the guilt and social costs incurred, compared to the well-mixed population setting, leading therefore to significantly higher levels of cooperation for a wider range of the costs. In structured population settings, both social and non-social guilt can evolve through clustering with emotional prone strategies, allowing them to be protected from exploiters, especially in case of non-social (less costly) strategies. Overall, our findings provide important insights into the design and engineering of self-organised and distributed cooperative multi-agent systems.

KEYWORDS

Guilt, emotion modelling, evolution of cooperation, social dilemma, evolutionary game theory, structured populations

ACM Reference Format:
Theodor Cimpeanu, Luís Moniz Pereira, and The Anh Han. 2023. Co-evolution of social and non-social guilt in structured populations. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 9 pages.

1 INTRODUCTION

"We are guilty for no reason, or just because we exist anyway, and are imperfect." - Peter J. Conradi [10]

Machine ethics involving the capacity for artificial intelligence (AI) to act morally is an open project for scientists and engineers [17, 44]. One important challenge is how to represent emotions that are thought to modulate human moral behaviour, such as guilt, in computational models [19, 31–33, 45, 46]. Upon introspection, guilt is present as a feeling of being worthy of blame for a moral offence. Burdened with guilt, an agent may then act to restore a blameless internal state in which this painful emotion is no longer present [58, 61].

The popular trend in research is to consider guilt more than shame leading to reparative actions. This has been looked at by de Hooge, Zeelenberg and Breugelmans [13], stating that guilt entails reparative action when there is conscious admission and accountability of the wrongdoing by the transgressor.

Sociocentric and egocentric cultures supposedly have different emotional expressions and experiences of shame and guilt. Sociocentric cultures, which are more social-looking, tend to generate more of a sense of character-intrinsic general shame, while more individualistic egocentric cultures lead more to a sense of specific action-intrinsic guilt in the transgressor [35]. Shame and guilt are often considered synonymous with one another, but shame is identified as a self-related emotion that motivates one to hide and escape, whereas guilt is identified as the emotion that motivates one to repair [6, 30]. When norms are well-established, societal members accept them as mandatory, internalise and comply with them, and experience guilt or shame when they violate them. When internal sanctions do not support compliance over extended periods of time, external sanctions may be necessary [7].

In social dilemmas such as the Prisoners' Dilemma (PD), where defection or cheating becomes the dominant strategy, defectors do better than cooperators regardless of whether their partners defect or cooperate [56]. In such a situation, it is rational for both parties to defect, even though mutual defection is often worse than reciprocal cooperation. Trivers [63] speculated that mutual evolution has promoted the emergence of guilt because it makes defection less attractive, with motivation from guilt becoming the dominant strategy due to attending social benefits. Individuals may gain materially by defecting, but guilt causes emotional suffering, and it is this suffering that encourages cooperation regardless of material gain. Nesse [37] sustains the temptation to defect arouses anxiety and defection arouses guilt, both aversive emotions that inhibit hasty selfishness. Guilt will motivate apologies, or self-punishment otherwise, and reparations are needed to reestablish trust.

From an evolutionary viewpoint, guilt is envisaged as an in-built mechanism that tends to prevent wrongdoing. Internal suffering and the need to alleviate it press an agent to their admission after wrongs are enacted, involving costly apology or penance, a change to correct behaviour, and an expectation of forgiveness to dispel

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

the guilt-induced suffering. The hypothesis then, is that within a population the emergence of guilt and its effects is evolutionary advantageous compared to a guilt free population. Moreover, the magnitude of the advantage presumably depends on the population's actual network structure, since it governs who is in touch with whom [3, 59], and determines the extent to which the social costs of guilt are globally worthwhile.

Inspired by the discussed psychological and evolutionary studies of guilt and cooperation in networks [3, 52, 60], this paper aims to provide a theoretical account of the evolution of costly guilt-prone behaviours in the context of distributed Multi-Agent Systems (MAS), with the overarching aim of achieving insights for the design and engineering of cooperative, self-organised systems. Resorting to methods from Evolutionary Game Theory (EGT) and agent-based simulations [43, 56], we study the evolution of social vs non-social aware guilt in differently structured populations.

We shall examine whether (non-)social guilt can evolve in such structured populations, e.g. through clustering of similarly emotionally prone individuals. Social guilt, and social emotions in general, depend upon awareness of the thoughts, feelings or actions of others in the environment [8, 25]. Thus, choosing to be social can be (much) more costly compared to being non-social, requiring efforts to understand others' thoughts and feelings and the context behind their actions; while non-sociality only requires awareness of one's own physical states. Hence, one might inquire whether and when such a more cost-efficient (but more easily exploitable as we will see) non-social strategy can evolve, depending on the underlying network structure.

This work fundamentally extends and generalises the work set forth in [45], which constructed theoretical models representing guilt to study its role in promoting pro-social behaviour, in the context of EGT using the Iterative PrisonersâĂŹ Dilemma (IPD) (see further discussion in Section 2, second paragraph). Guilt was modelled in terms of two features. Firstly, guilt involves a record of transgressions formalized as a counter tracking the number of offences. Secondly, guilt involves a threshold over which the guilty agent must alleviate it strained internal state, through deliberate change of behaviour and self-punishment, as required by the guilty feelings, both of which affect the game payoff for the guilty agent.

The remainder of the paper is structured as follows. We start off with related work, proceed to our models and methods, present their results, and terminate with concluding remarks. Moreover, we provide additional results as Supporting Information (**SI**).

2 RELATED WORK

The problems of explaining the evolution and emergence of collective behaviours, such as cooperation, coordination and AI safety in dynamical populations or systems of self-interested agents, have been actively studied across disciplines, from Evolutionary Biology, Physics, Economics to AI and Multi-Agent Systems [1, 9, 15, 21, 23, 26, 27, 36, 43, 47, 53, 54, 65, 66]. Several mechanisms have been proposed to explain the dilemmas of cooperation, including kin selection, direct and indirect reciprocity, incentives or networked structures; see surveys in [38, 43, 56]. In contrast, there is a significant lack of studies looking at the role of cognitive and emotional mechanisms in behavioural evolution [12, 20]. Given that emotions play a crucial role in humans' decision making [33, 64], it is crucial to take into account these complex mechanisms to provide a more complete rendering of the evolution of cooperation, not just amongst humans, but between humans and machines. Our work attempts to bridge this gap and provides important insights into the design and engineering of self-organised and distributed MAS, especially in a hybrid human-AI setting (e.g. for cooperative AI) [2, 12, 42].

Most relevant to our work is the EGT model proposed in [45], showing that cooperation does not emerge when agents only alleviate their own guilt (i.e. non-social guilt), without considering their co-playersâĂŹ own attitudes about alleviation of guilt as well. In that case, guilt-prone agents are easily dominated by agents who do not express guilt or who have no motivation to alleviate their own guilt. However, when the tendency to alleviate guilt is mutual (i.e. social guilt), only then can cooperation thrive. This work did not consider that choosing to be social might require a cost (compared to being non-social), and thus the latter might have an evolutionary advantage against the former. Indeed, our (risk-dominance) analysis below shows that in a direct competition, a non-social guilt strategy is risk-dominant or advantageous against a social one. Because this work did not consider both guilt-prone strategies in co-presence within a population, it was not possible to address how this social cost might affect the evolutionary outcomes. The present work considers an extended model where all these strategies are in co-presence together with other non-emotional strategies in a population, to address these issues. Moreover, this prior work [45] only focused on the well-mixed population setting, therefore failing to assess how the structure of the underlying network of contacts among the agents in the population affects the evolutionary outcome and the design of cooperative societies. For example, our results below show that a spatial structure, even if homogeneous like square lattices, allows guilt-prone strategies and cooperation to prevail for a much wider range of the guilt and social costs (compared to the well-mixed setting). Heterogeneous (scale-free) networks, and to some extent square lattices, allow non-social guilt to evolve through clustering of guilt-prone individuals to avoid their exploiters.

Guilt has been considered implicitly in prior EGT models studying apology and forgiveness in social dilemma games [22, 34, 50]. These works do not look at guilt as part of agents' strategies, but rather play an implicit role leading agents to make an apology after wrongdoings. In our work, the modelling of guilt as a behavioural feature of a strategy enables exploration of new aspects related to feeling guilty, namely its social aspects and how it interacts with external factors like the network structure.

Our modelling work is inspired by a large number of works from psychological/sociological/philosophical literature. Ramsey and Deem [48] argue that the evolutionary emergence of the emotion of guilt needs support on the evolution of empathy. From a multi-agent perspective, including mixed social-technological communities encompassing potentially autonomous artificial agents, and invoking the so-called âĂIJvalue alignmentâĂİ problem (for a recent review cf. [18]). In line with Pereira et al. [45], the outcomes from our analyses below help confirm that conflicts can be avoided when morally salient emotions, like guilt, help guide participants toward acceptable behaviours. In this context, systems involving possible future artificial moral agents may be designed to include guilt, to align agent-level behaviour with human expectations, thereby resulting in overall social benefits through improved cooperation.

Finally, there exists a large body of computational modelling works of guilt in AI and MAS literature [11, 14, 16, 24, 41, 50, 55, 64]. Differently from our goal, these studies aim to formalise guilt as part of a MAS, such as those of virtual agent and cognitive agent systems, for the purpose of regulating social norms [11] or of improving agent decision making and reasoning processes [33, 64]. Beyond that, our results provide novel insights into the design and engineering 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 can drive the system to an overall more cooperative outcome where they are willing to take reparative actions after wrongdoings. Moreover, our analysis provides insights on how such guilt-capable agents should be distributed to optimise cooperative outcomes, depending on the specific MAS network structure [33, 55, 64].

3 MODELS AND METHODS

First we recall the Iterated Prisoner's Dilemma (IPD) game and the definition of guilt-prone strategies, as described in [45]. We next describe our model where social and non-social guilt strategies are in co-presence. Then, the methods for analysing the model, namely stochastic evolutionary dynamics in well-mixed populations and agent-based simulations in networks, are in turn described.

3.1 Iterated Prisoners' Dilemma (IPD)

In each round of the IPD, two players engage in an PD game interaction where its outcomes are defined by the following payoff matrix (for the row player)

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

A player who chooses to cooperate (C) with another 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 [27, 56]. In this work we are namely 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, Ω.

For a convenient interpretation of results, we also consider the simplified version of the PD, the Donation game [56], where the payoff entries are described through the cost c ($c > 0$) and benefit b ($b > c$) of cooperation, as follows: $T = b$, $R = b - c$, $P = 0$, $S = -c$.

3.2 Guilt modelling in IPD

We base our model and analysis on Pereira et al. [45]'s approach, which formalizes guilt as an aspect of an agentâĂŹs genotypical strategies, and is quantified in terms of a threshold, G. In this model, $G \in [0, +\infty]$, and guilt at a given time is characterized by a transient level of guilt, g ($g \geq 0$). As the experiment begins, g for every agent is set to 0. It increases by 1 after an action that the agent considers as wrong. After several wrongdoings result in g reaching that agentâĂŹs threshold of guilt, $g \geq G$, the agent can choose to (or not to) act to reduce guilt level g below that threshold. The model retains the mechanism of guilt alleviation described above, whereby guilt can be alleviated by apologising to offended partners, or by suffering guilt as self-punishment when apology to offended partners is not an option, which is what we will admit in the sequel. Either way the guilty party suffers a cost. The alleviation of guilt is costly, this cost being quantified in terms of γ ($\gamma \geq 0$), with which g is decreased by 1. According to this definition, agents can be characterized with respect to different guilt thresholds. Some may be incapable of suffering guilty feelings, so their $G = +\infty$. Others may be extremely prone to guilt, suffering guilty feelings with any first mistake, so for them $G = 0$. These are the only two cases we consider below.

3.3 Social vs. non-social guilt in co-presence

In this setting, a strategy is described by three factors or components:

(I) Guilt threshold G**.** Since we shall focus in the current work on understanding the evolution of social (or social guilt) behaviours as well as the impact of network structures, we consider two basic types of guilt thresholds

- $G = +\infty$: In this type of agent the guilt level g will never reach the threshold no mater how many times they defect; hence, they never need to reduce g, and consequently never pay the guilt cost γ. In other words, this type of agent experiences no guilt feeling. They are dubbed (guilt-)unemotional agents.
- $G = 0$: whenever this type of agent defects, it becomes immediately true that $g > G$; hence, the agent needs to act right away to reduce g, thus paying γ. In other words, this type of agent always feels guilty subsequent to a single a wrongdoing, viz. defection. They are dubbed (guilt-)emotional agents.

(II) Decision making in the IPD. Given the agent's guilt threshold G, she can choose to play either C or D in a PD and, if the ongoing guilt level g reaches G, whether to change her behaviour from D to C (to avoid further emotional pain and cost).

(III) Social vs non-social about when to feel guilty. The emotional agents can choose to be non-social or social, regarding the way they express their emotion. To be social agents need an extra effort such as signalling guilt or observing guilt. Hence, we assume there is an additional cost, γ_S, to being social.

Overall, since we do not consider noise in IPD (i.e. non-deliberate mistakes) in this work, there are in total six possible strategies [1], denoted as follows

[1]There can be other strategies such as emotional (i.e. $G = 0$) cooperators who always cooperate and thus never feel guilty. But as we are not modelling noise in this work, this strategy is equivalent to C in all interactions, and can be removed from our analysis.

(1) Unemotional cooperator (C): always cooperates, unemotional (i.e. $G = +\infty$)
(2) Unemotional defector (D): always defects, unemotional (i.e. $G = +\infty$)
(3) Emotional non-adaptive defector that is non-social (DGDN): always defects, feels guilty after one wrongdoing (i.e. $G = 0$) regardless of what others feel, but does not change its behaviour.
(4) Emotional adaptive defector that is non-social (DGCN): defects initially, feels guilty after one wrongdoing (i.e. $G = 0$) regardless of what others feel, and changes its behaviour from D to C.
(5) Emotional non-adaptive defector that is social (DGDS): always defects, feels guilty after one wrongdoing (i.e. $G = 0$) only if her co-player would also feel guilty after a wrongdoing, but does not change its behaviour.
(6) Emotional adaptive defector that is social (DGCS): defects initially, feels guilty after one wrongdoing (i.e. $G = 0$) only if her co-player would feel also guilty after a wrongdoing, and changes its behaviour from D to C.

We can derive the payoff matrix for the six strategies (for row player), as follows

$$
\begin{array}{c}
\\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6
\end{array}
\begin{array}{c}
\begin{array}{cccccc}
\quad C \quad & \quad D \quad & DGDN & DGCN & DGDS & DGCS
\end{array}
\\
\begin{pmatrix}
R & S & S & \frac{S+R\Theta}{\Omega} & S & \frac{S+R\Theta}{\Omega} \\
T & P & P & \frac{P+T\Theta}{\Omega} & P & P \\
T-\gamma & P-\gamma & P-\gamma & \frac{P+T\Theta}{\Omega}-\gamma & P-\gamma & \frac{P+T\Theta}{\Omega}-\gamma \\
\frac{T-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S\Theta}{\Omega} & \frac{P-\gamma+S\Theta}{\Omega} & \frac{P-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S\Theta}{\Omega} & \frac{P-\gamma+R\Theta}{\Omega} \\
T-\gamma-\gamma_S & P-\gamma_S & P-\gamma-\gamma_S & \frac{P+T\Theta}{\Omega}-\gamma-\gamma_S & P-\gamma-\gamma_S & \frac{P+T\Theta}{\Omega}-\gamma-\gamma_S \\
\frac{T-\gamma-\gamma_S+R\Theta}{\Omega} & P-\gamma_S & \frac{P-\gamma-\gamma_S+S\Theta}{\Omega} & \frac{P-\gamma-\gamma_S+R\Theta}{\Omega} & \frac{P-\gamma-\gamma_S+S\Theta}{\Omega} & \frac{P-\gamma-\gamma_S+R\Theta}{\Omega}
\end{pmatrix}
\end{array},
\tag{1}
$$

where we employ $\Theta = \Omega - 1$ just for the purpose of a neater representation.

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 be able to disseminate throughout the population.

Previous work shows that an emotional guilt-based response only makes sense when the other is not attempting to harm you too, or attempting to harm you but feeling guilty too [45]. That is, guilt needs to be social to prevail in social dynamics. The main reason is that players who feel guilty after a wrongdoing, regardless of others' behaviours, would be exploited by non-emotional defectors (i.e. D strategy). We argue that, since being social might be costly as agents need to observe and understand others' actions and feelings, non-social guilt might be more cost-efficient and prevail in environments where they are protected from such exploiters. As previous guilt modelling work only looked at well-mixed populations wherein all individuals in the population interact with each other, it was not possible to consider such protection. To bridge this gap, in this work, we address structured populations where players interact with their direct neighbours.

3.4 Evolutionary Dynamics in Well-Mixed Populations

Individuals' payoff represents their *fitness* or social *success*, and evolutionary dynamics is shaped by social learning [28, 57], whereby the most successful agents will tend to be imitated more often by the other agents. In the current work, social learning is modeled using the so-called pairwise comparison rule [62], a standard approach in EGT, assuming that an agent A with fitness f_A adopts the strategy of another agent B with fitness f_B with probability p given by the Fermi function,

$$p_{A,B} = \left(1 + e^{-\beta(f_B - f_A)}\right)^{-1}. \tag{2}$$

The parameter β represents the 'imitation strength' or 'intensity of selection', i.e., how strongly the agents base their decision to imitate on fitness difference between themselves and the opponents. For $\beta = 0$, we obtain the limit of neutral drift – the imitation decision is random. For large β, imitation becomes increasingly deterministic. In line with previous works and human behavioural experiments [49, 59, 67], we set $\beta = 1.0$ in the main text, which also allow us to compare directly with the previous guilt model in [45].

In the absence of 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 agent switches randomly to a different strategy without imitating another agent. In the limit of small mutation rates, the dynamics will proceed with, at most, two strategies in the population, such that the behavioral 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 [29, 39]. The resulting Markov Chain has a stationary distribution, which characterizes the average time the population spends in each of these monomorphic end states (see some examples in Figure 1).

Let N be the size of the population. Denote $\pi_{X,Y}$ the payoff a strategist X obtains in a pairwise interaction with strategist Y (defined in the payoff matrices). Suppose there are at most two strategies in the population, say, k agents using strategy A ($0 \leq k \leq N$) and $(N-k)$ agents using strategies B. Thus, the (average) payoff of the agent that uses A (similarly for B) is

$$\Pi_A(k) = \frac{(k-1)\pi_{A,A} + (N-k)\pi_{A,B}}{N-1}. \tag{3}$$

Now, the probability to change the number k of agents using strategy A by $\pm$ 1 in each time step can be written as [62]

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

The fixation probability of a single mutant with a strategy A in a population of $(N-1)$ agents using B is given by [39, 62]

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

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

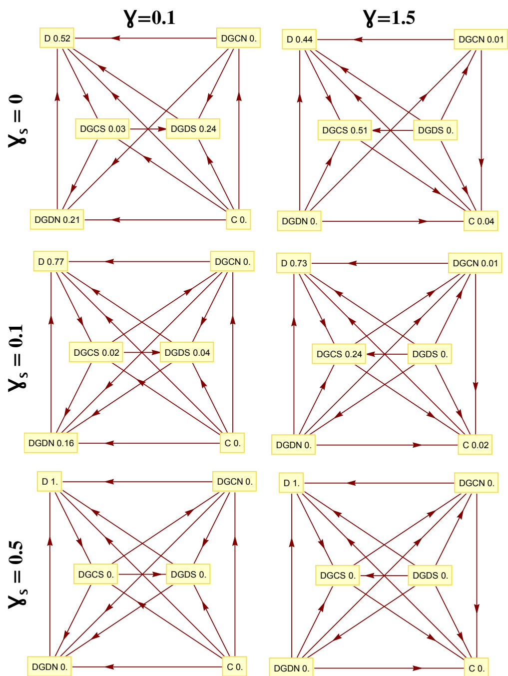

Figure 1: Markov diagrams and stationary distributions (Well-mixed Populations). Transitions direction among strategies, where the arrows show the direction where the transition probability is stronger than the reverse. The results are in line with risk-dominance analysis (in Section 4.1). Other parameters: $N = 100$, $\Omega = 10$, $R = 1$, $S = -1$, $T = 2$, $P = 0$.

the transposed of M provides the stationary distribution described above [29], depicting the relative time the population spends adopting each of the strategies.

Risk-dominance. An important measure to compare the two strategies A and B is which direction the transition is stronger or more probable, an A mutant fixating in a population of agents using B, $\rho_{B,A}$, or a B mutant fixating in the population of agents using A, $\rho_{A,B}$. It can be shown that the former is stronger, in the limit of large N, if [39, 57]

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

3.5 Agent-based Simulations and Network Structures

3.5.1 Network Topologies. Links in the network describe a relationship of proximity both in the interactional sense (whom the agents can interact with), but also observationally (whom the agents can imitate). Ergo, the network of interactions coincides with the imitation network [40]. As each network type converges at different rates and naturally presents various degrees of heterogeneity, we choose different population sizes in the various experiments to account for this while optimising run-time.

Parameter	Symbol
Population size	N
Cost of cooperation	c
Benefit of cooperation	b
Intensity of selection	β
Guilt cost	γ
Social cost of guilt	γ_S
Number of rounds in IPD	Ω
Guilt threshold	G

Table 1: Model parameters

Well-mixed populations offer a convenient baseline scenario, where interaction structure is absent. By studying structured populations, we go one step beyond and ask whether network properties and structural heterogeneity can foster the evolution of guilt-prone behaviours. To begin with, we study square lattice (SL) populations of size $N = 30 \times 30$, with periodic boundary conditions — a widely adopted population structure in population dynamics and evolutionary games (for a survey, see [59]), wherein each agent can only interact with its four immediate edge neighbours. With the SL we introduce a network structure, yet one where all nodes can be envisaged as equivalent.

Going further still, we explore complex networks in which the network portrays a heterogeneity that mimics the power-law distribution of wealth (and opportunity) characteristic of real-world settings. The Barabási and Albert (BA) model [5] is one of the most famous models used in the study of such heterogeneous, complex networks. The main features of the BA model are that it follows a *preferential attachment* rule, has a small clustering coefficient, and a typical *power-law degree distribution.* In order to explain preferential attachment, let us describe the construction of a BA network. Starting from a small set of m_0 interconnected nodes, each new node selects and creates a link with m older nodes according to a probability proportional to their degree (number of edges). The procedure stops when the required network size of N is reached. This will produce a network characterised by a power-law distribution, $p_k \sim k^{-\gamma}$, where the exponent γ is its *degree exponent* [4]. There is a high degree correlation between nodes, and the degree distribution is typically skewed with a long tail. There are few hubs in the network that attract an increasing number of new nodes which attach as the network grows (in a typical *"rich-get-richer"* scenario). The power-law distribution exhibited by BA networks resembles the heterogeneity present in many real-world networks. The average connectivity of the resulting scale-free network is $z = 2m$. For all of our experiments, we pre-seed 10 different scale-free networks of size $N = 1000$, with an average connectivity of $z = 4$, to coincide with the number of neighbours in a square lattice.

3.5.2 Computer Simulations. Initially each agent is designated as one of the six strategies (i.e., C, D, DGDN, DGCN, DGDS, DGCS), with equal probability. At each time step, each agent plays the PD with its immediate neighbours. The score for each agent is the sum of the payoffs in these encounters. At the end of each step, an agent A with fitness f_A chooses to copy the strategy of a randomly selected neighbour agent B with score f_B, with a probability given by the Fermi function [59], as given in Equation 2. Similar to the well-mixed setting above, we set $\beta = 1$ in our simulations.

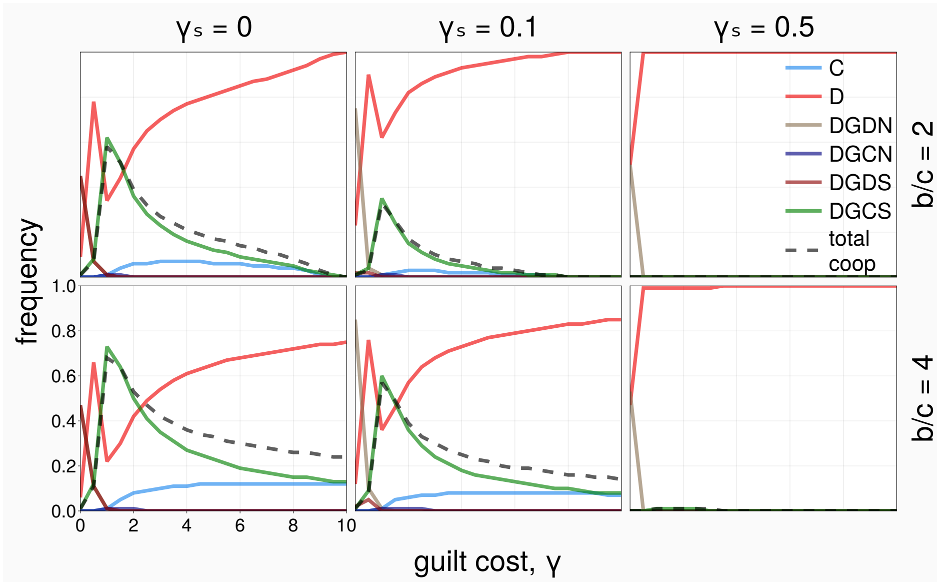

Figure 2: Strategies' frequency and total cooperation level as a function of the guilt cost, γ (well-mixed, $N = 100$, $\Omega = 10$).

We simulate this evolutionary process until a stationary state or a cyclic pattern is reached. For the sake of a clear and fair comparison, all simulations are run for 10^6 steps. Moreover, for each simulation, the results are averaged over the final 10^5 generations, in order to account for the fluctuations characteristic of these stable states. Furthermore, to improve accuracy, for each set of parameter values, the final results are obtained from averaging 30 independent realisations (20 for scale-free networks due to computational overheads and the additional pre-seeded networks, i.e. 200 replicates for SF networks).

4 RESULTS

Given the model and methods described above (see Table 1 for a summary of the parameters), we first derive analytical conditions for when guilt-prone strategies can be viable and promote the evolution of enhanced cooperation. Next, we obtain numerical results for the well-mixed population setting, validating the analytical conditions. We then show results from our extensive agent-based simulations in structured population settings.

4.1 Risk dominance of guilt-prone strategies

To start with, we obtain analytical conditions for when guilt-prone strategies can be evolutionarily viable against other strategies. For that, we apply the risk-dominance criteria in Equation 6 to the payoff matrix given in Equation 1.

First, DGCS is risk-dominant against DGDS if

$$\gamma + \gamma_S > \frac{T - R + P - S}{2} = c. \tag{7}$$

The condition for DGCS to be risk-dominant against C is the reverse of that of against DGDS above. DGCS is risk-dominant against DGDN if

$$(\Omega - 1)\gamma - \gamma_S > (\Omega - 1)\frac{T - R + P - S}{2} = (\Omega - 1)c. \tag{8}$$

It can be seen that this condition subsumes the one for risk-dominance against DGDS above. Also, for this inequality to hold the necessary condition is $\gamma > c$.

Now, DGCS is risk-dominant against D if

$$\gamma + (\Omega + 1)\gamma_S < (\Omega - 1)(R - P) = (\Omega - 1)(b - c). \tag{9}$$

DGCS is risk-dominated by DGCN whenever $\gamma_S > 0$. They are neutral when $\gamma_S = 0$. However, DGCN is always risk-dominated by D. Thus, there is a cyclic pattern from DGCS (social guilt), to DGCN (non-social guilt), to D (non-emotional defectors), and back to DGCS, whenever the condition in Equation 9 holds. That occurs when γ and γ_S are sufficiently small. Fixing c, the latter condition is more easily satisfied for a more beneficial PD (i.e. large b).

Moreover, DGCS to be risk-dominant against all the defective strategies (i.e. all but C and DGCN), the guilt cost γ needs to be sufficiently large; that is, at least the cost of cooperation, c. Given that, the smaller the social cost, the easier it is for these conditions to be satisfied. The upper bound of this cost is $\frac{(\Omega-1)(b-c)}{\Omega+1}$.

4.2 Well-mixed populations: Evolution of social guilt and the eradication of non-social guilt

To illustrate the above obtained analytical observations, Figure 1 shows the stationary distribution and transition directions in a well-mixed population of the six strategies (see Methods). We can see that the directions of transition, showing risk-dominance of the strategy at the end of the transition or arrow, corroborate the analytical conditions.

Figure 2 shows the long-term frequencies of the strategies and the total level of cooperation in the population, for varying the guilt cost γ, for different benefits $b = 2$ (first row) and $b = 4$ (second row), and for different social costs γ_S. We observe that, when the social cost γ_S is sufficiently small, there is an intermediate value of the guilt cost γ (around $\gamma = c$), which leads to an optimal frequency

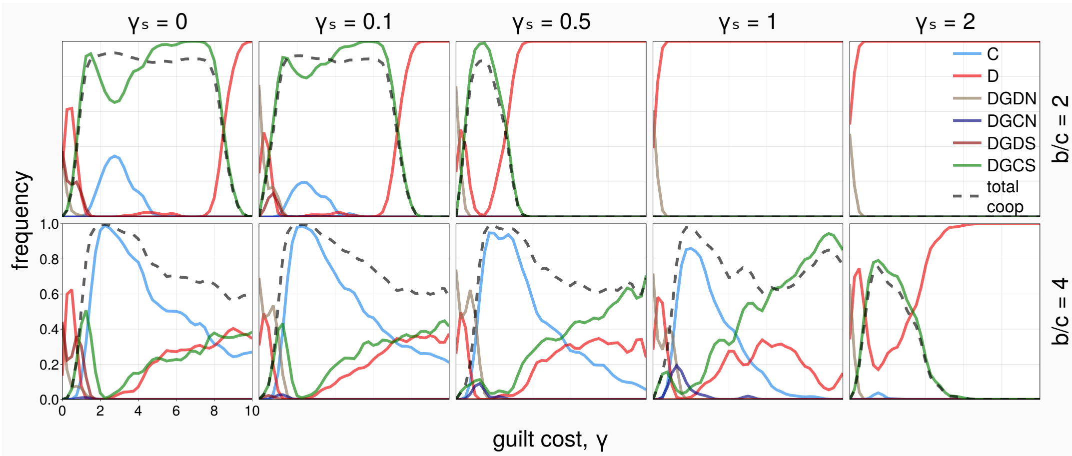

Figure 3: Strategies' frequency and the total cooperation level as a function of the guilt cost, γ (square lattice, $N = 900$, $\Omega = 10$)

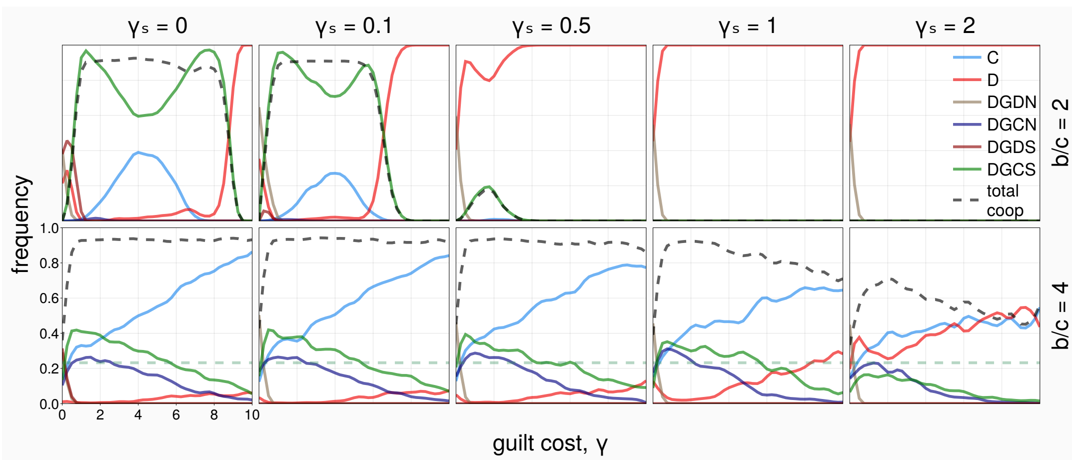

Figure 4: Strategies' frequency and the total cooperation level as a function of the guilt cost, γ (scale-free, $N = 1000$, $\Omega = 10$). When shown, the dashed green line marks the baseline level of cooperation achieved solely through network reciprocity.

Figure 5: Structured populations foster clustering in mixed strategy outcomes. The stacked bars represent the mean fraction of strategists in the neighbourhood for each focal strategist. The percentage shown on the bar represents the total fraction of those players in the population. Left column reports results for square lattice ($N = 900$) and right one for scale-free networks ($N = 1000$). Typical runs selected to show mixed strategy outcomes if available (more replicates and different parameter values in SI). Parameters: $\Omega = 10$; $b = 2$, $\gamma = 4$, $\gamma_S = 0$ (A and B); $b = 4$, $\gamma = 1$, $\gamma_S = 1$ (C and D); $b = 4$, $\gamma = 7$, $\gamma_S = 0$ (E and F).

of DGCS and the total cooperation in the population. When γ is too small, DGCS is dominated by DGDN (and DGDS) (see also Figure 1, first column). When γ is larger, D frequency increases and dominates the population, despite being still dominated by DGCS (see Figure 1, second column). There is now a transition from DGCS to C which is strongly dominated by D. Comparing the first and second rows of Figure 2, a higher level of cooperation is achieved for a larger benefit of cooperation b.

In short, we can observe that social guilt (DGCS) can evolve in the well-mixed population setting when the social cost is sufficiently small, reaching its peak around $\gamma \approx c$. Non-social guilt does not evolve at all in this setting, even when it dominates DGCS (whenever $\gamma_S > 0$, see Figure 1, second and third rows), as DGCN is always strongly dominated by D.

4.3 Structured populations enhance social guilt and enable the emergence of non-social guilt

We study the effect of spatial or structured populations on the evolutionary dynamics and outcomes of guilt-prone strategies (both social and non-social), as well as cooperation. Firstly, we consider results in the square lattice (SL) network, a regular (homogeneous) structure, see Figure 3. We observe that, for a small benefit of cooperation $b = 2$ (top row), for sufficiently small social costs γ_S (0 and 0.1), DGCS dominates the population over a wide range of γ, between approximately $1 < \gamma < 8$. Interestingly, there is also a chance for C to emerge. Moreover, when b is larger (bottom row), C even dominates the population for a wide range of γ and γ_S. DGCS dominates when γ is sufficiently high. Interestingly, in such networked populations, even non-social guilt strategy can survive with some frequency when the social cost is non-negligible, see $\gamma_S = 0.1,\ 0.5$ and 1 at intermediate ranges of γ. Overall, we observe significantly higher levels of cooperation and guilt-prone strategies for a wider range of both guilt and social costs, compared to well-mixed populations.

Importantly, we see a shift in the cyclic dynamics previously encountered in well-mixed populations. This property can be clarified by observing the clustering behaviours typical of structured populations, even in the case of homogeneous graphs (see Figure 5, left column). Typically, we see that unemotional cooperators (C) are better protected against unemotional defectors (D) when spatiality allows for network reciprocity, especially when evolutionary dynamics lead to mixed strategy outcomes (no one strategy fully dominates the others). Through such clusters, emotionally adaptive strategists (DGCN and DGCS) can often survive in the face of D players. Moreover, this can allow for the co-existence of guilt-prone individuals in communities of other like-minded strategists and C players, especially if the cost of being social (γ_s) is low enough (e.g., $\gamma_s = 0$ and $\gamma_s = 1$, as highlighted in Figure 5).

We now consider a more complex network structure, the scale-free (SF) network, heterogeneous and highly diverse in the number and distribution of connections. Previous works studying the evolution of cooperation on different networks showed that SF properties can markedly promote cooperation in one-shot social dilemmas, as heterogeneity in the network structure allows cooperators to form clusters around highly connected nodes (hubs) [51, 52, 59]. Our aim is to study whether this property would also allow pro-social behaviours to evolve; strategies which would not have had a chance to do so previously. To this end, we investigate whether non-social guilt strategies can emerge, leading to even higher levels of (less-costly) cooperation overall.

We observe similar outcomes to SL when $b = 2$, with a slight decrease of cooperation when $\gamma_S = 0.5$. When $b = 4$, we find higher levels of cooperation in SF than in SL, across a wide range of guilt and social costs. This improvement can be attributed to the success of non-social guilt, which becomes rather abundant across the entire parameter space. This is a remarkable observation, whereby the easily exploitable non-social individuals (which are also desirably cost efficient) can evolve and co-exist with other strategies in an evolving population/MAS of self-interested agents.

To further explain this finding and confirm our intuitions, we show the clustering behaviours typical of scale-free populations in Figure 5, right column. Given a low social cost γ_s, social guilt can thrive even in cases when the cost of guilt γ is very large (see Figure 5 panels B and F). Communities of emotionally adaptive individuals co-evolve and co-exist, surviving in the face of the predictions of evolutionary dynamics in homogeneous populations. That is, emotionally sacrificial strategies are empowered through heterogeneous environments, even in an incipient form that does not require costly monitoring of the surrounding contexts.

5 CONCLUDING REMARKS

Based on psychological and evolutionary accounts of guilt and social emotions, the present paper studies an evolutionary game theoretical model with social and non-social guilt-prone strategies in co-presence, in the context of structured populations (or distributed MASs). The paper considered several important population structures, from homogeneous ones, in the forms of well-mixed and square lattices, to heterogeneous, scale-free networks, showing that the evolutionary outcomes of social and non-social guilt strategies are highly dependent on population structure.

We showed, in the context of the Iterated Prisoner's Dilemma, that only social guilt can evolve in the well-mixed population context, which is in line with previous findings in the literature [45] (see **SI** for additional analyses where social and non-social guilt strategies are considered separately). Spatial structures, even homogeneous ones (e.g. square lattices), allow guilt-prone strategies and cooperation to prevail for a much wider range of the guilt and social costs (compared to the well-mixed setting). Interestingly, heterogeneous networks (i.e. scale-free), and to a lesser extent square lattices, allow non-social guilt to evolve through the formation of clusters with other emotional agents to defend against exploitation.

This finding is remarkable, as it showed that costly guilt-prone strategies can prevail in spatial environments, even in an incipient form which does not require expensive monitoring of the context behind others' actions. This is especially true when the underlying networks mirror realistic, heterogeneous structures [3].

Overall, the present investigation has resulted in a rigorous, game-theoretical based account, of how together the social costs and underlying network structures of a population, or distributed MAS, allow for the co-evolution and co-existence of diverse forms of social and non-social emotions. As a result, this strengthens cooperation, though their beholders incur a significant emotional cost to achieve this.

REFERENCES

[1] Stéphane Airiau, Sandip Sen, and Daniel Villatoro. 2014. Emergence of conventions through social learning. *Autonomous Agents and Multi-Agent Systems* 28, 5 (2014), 779–804.

[2] Peter Andras, Lukas Esterle, Michael Guckert, The Anh Han, Peter R Lewis, Kristina Milanovic, Terry Payne, Cedric Perret, Jeremy Pitt, Simon T Powers, et al. 2018. Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine* 37, 4 (2018), 76–83.

[3] Albert-Laszlo Barabasi. 2014. *Linked-how Everything is Connected to Everything Else and what it Means F.* Perseus Books Group.

[4] Albert-László Barabási. 2016. *Network Science.* Cambridge University Press. 474 pages pages.

[5] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

[6] Coralie Bastin, Ben J Harrison, Christopher G Davey, Jorge Moll, and Sarah Whittle. 2016. Feelings of shame, embarrassment and guilt and their neural correlates: A systematic review. *Neuroscience & Biobehavioral Reviews* 71 (2016), 455–471. https://doi.org/10.1016/j.neubiorev.2016.09.019

[7] Paul Billingham and Tom Parr. 2020. Online Public Shaming: Virtues and Vices. *Journal of Social Philosophy* 51, 3 (sep 2020), 371–390. https://doi.org/10.1111/josp.12308

[8] Stephanie Burnett, Geoffrey Bird, Jorge Moll, Chris Frith, and Sarah-Jayne Blakemore. 2009. Development during adolescence of the neural processing of social emotion. *Journal of cognitive neuroscience* 21, 9 (2009), 1736–1750.

[9] Theodor Cimpeanu, Cedric Perret, and The Anh Han. 2021. Cost-efficient interventions for promoting fairness in the ultimatum game. *Knowledge-Based Systems* 233 (2021), 107545.

[10] Peter J Conradi. 2010. Laughing at Something Tragic: Murdoch as Anti-Moralist BT - Iris Murdoch and Morality. Palgrave Macmillan UK, London, 56–69. https://doi.org/10.1057/9780230277229_5

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

[12] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, Thore Graepel, et al. 2021. Cooperative AI: machines must learn to find common ground. *Nature* 593, 7857 (2021), 33–36.

[13] Ilona E De Hooge, Marcel Zeelenberg, and Seger M Breugelmans. 2010. Restore and protect motivations following shame. *Cognition and Emotion* 24, 1 (2010),

111–127.

[14] Celso M De Melo, Peter Carnevale, Stephen Read, Dimitrios Antos, and Jonathan Gratch. 2012. Bayesian model of the social effects of emotion in decision-making in multiagent systems. In *AAMAS'2012*. 55–62.

[15] Elias Fernández Domingos, Juan Carlos Burguillo, and Tom Lenaerts. 2017. Reactive versus anticipative decision making in a novel gift-giving game. In *Thirty-First AAAI Conference on Artificial Intelligence*. 4399–4405.

[16] Julia Fix, Christian von Scheve, and Daniel Moldt. 2006. Emotion-based Norm Enforcement and Maintenance in Multi-agent Systems: Foundations and Petri Net Modeling. In *AAMAS '06*. ACM, 105–107.

[17] S. A. Frank. 1998. *Foundations of social evolution*. Princeton Univ. Press, Princeton.

[18] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30, 3 (2020), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[19] Benoit Gaudou, Emiliano Lorini, and Eunate Mayor. 2014. Moral guilt: An agent-based model analysis. In *Advances in social simulation*. Springer, 95–106.

[20] The Anh Han. 2022. Emergent Behaviours in Multi-agent Systems with Evolutionary Game Theory. *AI Communications* 35, 4 (2022), 327 âĂŞ 337.

[21] The Anh Han, Luís Moniz Pereira, and Tom Lenaerts. 2017. Evolution of commitment and level of participation in public goods games. *Autonomous Agents and Multi-Agent Systems* (2017), 1–23.

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

[23] The Anh Han, Luis Moniz Pereira, Francisco C. Santos, and Tom Lenaerts. 2020. To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race. *Journal of Artificial Intelligence Research* 69 (2020), 881–921.

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

[25] Shlomo Hareli and Brian Parkinson. 2008. What's social about social emotions? *Journal for the Theory of Social Behaviour* 38, 2 (2008), 131–156.

[26] Mohammad Rashedul Hasan and Anita Raja. 2013. Emergence of Cooperation using Commitments and Complex Network Dynamics. In *IEEE/WIC/ACM Intl Joint Conferences on Web Intelligence and Intelligent Agent Technologies*. 345–352.

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

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

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

[30] Shian-Ling Keng and Jun Xian Tan. 2017. Effects of brief mindful breathing and loving-kindness meditation on shame and social problem solving abilities among individuals with high borderline personality traits. *Behaviour Research and Therapy* 97 (2017), 43–51. https://doi.org/10.1016/j.brat.2017.07.004

[31] Zdzisław Kowalczuk and Michał Czubenko. 2016. Computational approaches to modeling artificial emotion–an overview of the proposed solutions. *Frontiers in Robotics and AI* 3 (2016), 21.

[32] Kingson Man and Antonio Damasio. 2019. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence* 1, 10 (2019), 446–452.

[33] Stacy Marsella and Jonathan Gratch. 2014. Computationally modeling human emotion. *Commun. ACM* 57, 12 (2014), 56–67.

[34] Luis A Martinez-Vaquero, The Anh Han, Luís Moniz Pereira, and Tom Lenaerts. 2015. Apology and forgiveness evolve to resolve failures in cooperative agreements. *Scientific reports* 5, 10639 (2015).

[35] Claude-Hélène Mayer and Elisabeth Vanderheiden. 2021. Naming and Shaming in Cyberspace: Forms, Effects and Counterstrategies BT - Shame 4.0: Investigating an Emotion in Digital Worlds and the Fourth Industrial Revolution. Springer International Publishing, Cham, 389–412. https://doi.org/10.1007/978-3-030-59527-2_18

[36] Ramona Merhej, Fernando P. Santos, Francisco S. Melo, and Francisco C. Santos. 2022. Cooperation and Learning Dynamics under Wealth Inequality and Diversity in Individual Risk. *J. Artif. Int. Res.* 74 (sep 2022), 32. https://doi.org/10.1613/jair.1.13519

[37] Randolf M Nesse. 2019. *Good Reasons for Bad Feelings: Insights from the Frontier of Evolutionary Psychiatry*. Allen Lane. 384 pages.

[38] M. A. Nowak. 2006. *Evolutionary Dynamics*. Harvard University Press, Cambridge, MA.

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

[40] Hisashi Ohtsuki, Martin A Nowak, and Jorge M Pacheco. 2007. Breaking the symmetry between interaction and replacement in evolutionary dynamics on graphs. *Physical review letters* 98, 10 (2007), 108106.

[41] Cailin OâĂŹConnor. 2016. The evolution of guilt: a model-based approach. *Philosophy of Science* 83, 5 (2016), 897–908.

[42] Ana Paiva, Fernando P Santos, and Francisco C Santos. 2018. Engineering prosociality with autonomous agents. In *Thirty-second AAAI conference on artificial intelligence*.

[43] Matjaž Perc, Jillian J Jordan, David G Rand, Zhen Wang, Stefano Boccaletti, and Attila Szolnoki. 2017. Statistical physics of human cooperation. *Phys Rep* 687 (2017), 1–51.

[44] Luís Moniz Pereira, The Anh Han, and António Barata Lopes. 2021. Employing AI to Better Understand Our Morals. *Entropy* 24, 1 (2021), 10.

[45] Luís Moniz Pereira, Tom Lenaerts, Luis A Martinez-Vaquero, and The Anh Han. 2017. Social manifestation of guilt leads to stable cooperation in multi-agent systems. In *AAMAS*. 1422–1430.

[46] Luís Moniz Pereira, Ari Saptawijaya, et al. 2016. *Programming machine ethics*. Vol. 26. Springer.

[47] Steve Phelps, Peter McBurney, and Simon Parsons. 2010. Evolutionary mechanism design: a review. *Autonomous Agents and Multi-Agent Systems* 21, 2 (2010), 237–264.

[48] Grant Ramsey and Michael J Deem. 2022. Empathy and the Evolutionary Emergence of Guilt. *Philosophy of Science* 89, 3 (2022), 434–453. https://doi.org/DOI:10.1017/psa.2021.36

[49] David G. Rand, Corina E. Tarnita, Hisashi Ohtsuki, and Martin A. Nowak. 2013. Evolution of fairness in the one-shot anonymous Ultimatum Game. *Proc. Natl. Acad. Sci. USA* 110 (2013), 2581–2586.

[50] Sarita Rosenstock and Cailin O'Connor. 2016. When it's Good to Feel Bad: Evolutionary Models of Guilt and Apology. *Philosophy of Science* 64, 6 (2016), 637–658.

[51] F. C. Santos and J. M. Pacheco. 2005. Scale-free networks provide a unifying framework for the emergence of cooperation. *Phys. Rev. Lett.* 95 (2005), 098104.

[52] F. C. Santos, M. D. Santos, and J. M. Pacheco. 2008. Social diversity promotes the emergence of cooperation in public goods games. *Nature* 454 (2008), 214–216.

[53] Fernando P Santos, Samuel Mascarenhas, Francisco C Santos, Filipa Correia, Samuel Gomes, and Ana Paiva. 2020. Picky losers and carefree winners prevail in collective risk dilemmas with partner selection. *Autonomous Agents and Multi-Agent Systems* 34, 2 (2020), 1–29.

[54] Bastin Tony Roy Savarimuthu and Stephen Cranefield. 2011. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems* 7, 1 (2011), 21–54.

[55] Bastin Tony Roy Savarimuthu, Maryam Purvis, and Martin Purvis. 2008. Social Norm Emergence in Virtual Agent Societies. In *AAMAS '08*. 1521–1524.

[56] K. Sigmund. 2010. *The calculus of selfishness*. Princeton Univ. Press.

[57] Karl Sigmund. 2010. *The Calculus of Selfishness*. Princeton University Press.

[58] Nick Smith. 2008. I was wrong: The meanings of apologies. (2008).

[59] G. Szabó and G. Fáth. 2007. Evolutionary games on graphs. *Phys Rep* 97-216, 4-6 (2007).

[60] Péter Szabó, Tamás Czárán, and György Szabó. 2007. Competing associations in bacterial warfare with two toxins. *J. theor. Biol.* 248 (2007), 736–744.

[61] June P Tangney, Jeffrey Stuewig, Elizabeth T Malouf, and Kerstin Youman. 2013. 23 Communicative Functions of Shame and Guilt. *Cooperation and its evolution* (2013), 485.

[62] A. Traulsen, M. A. Nowak, and J. M. Pacheco. 2006. Stochastic Dynamics of Invasion and Fixation. *Phys. Rev. E* 74 (2006), 11909.

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

[64] Paolo Turrini, John-Jules Ch. Meyer, and Cristiano Castelfranchi. 2010. Coping with shame and sense of guilt: a Dynamic Logic Account. *Autonomous Agents and Multi-Agent Systems* 20, 3 (2010), 401–420. https://doi.org/10.1007/s10458-009-9083-z

[65] Karl Tuyls and Simon Parsons. 2007. What evolutionary game theory tells us about multiagent learning. *Artificial Intelligence* 171, 7 (2007), 406–416.

[66] Jason Xu, Julian Garcia, and Toby Handfield. 2019. Cooperation with bottom-up reputation dynamics. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 269–276.

[67] Ioannis Zisis, Sibilla Di Guida, The Anh Han, Georg Kirchsteiger, and Tom Lenaerts. 2015. Generosity motivated by acceptance - evolutionary analysis of an anticipation games. *Scientific reports* 5, 18076 (2015).