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 whether such social and non-social guilt can evolve, depending on the underlying structure of the populations or systems of agents. 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; evolution of cooperation; evolutionary game theory; social dilemma; structured populations

ACM Reference Format:

1 INTRODUCTION

Machine ethics involving the capacity for artificial intelligence (AI) to act morally is an open project for scientists and engineers [6, 22]. One important challenge is how to represent emotions that are thought to modulate human moral behaviour, such as guilt, in computational models [7, 12, 14–16, 23, 24]. 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 an blameless internal state in which this painful emotion is no longer present [29, 32].

From an evolutionary viewpoint, guilt is envisaged as a built-in 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 the guilt-induced suffering [9, 10, 13, 17–19, 25]. 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 [2, 3, 5, 26, 30], 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 [2, 21, 27, 31], 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. This work fundamentally extends and generalises the work set forth in [23], which constructed theoretical models representing guilt to study its role in promoting pro-social behaviour, by examining whether guilt can evolve in such structured populations, for instance through clustering of similarly emotionally prone individuals.

2 MODELS AND METHODS

We base our model and analysis on Pereira et al. [23]’s approach, which formalizes guilt as an aspect of an agent’s genotypical strategies. We consider that an agent might play C or D in an Iterated Prisoner’s Dilemma, and given an on-going guilt level, they might change their behaviour from D to C (to avoid further emotional pain and cost). They can also express their emotion socially, which requires an extra effort, such as signalling or observing guilt. In brief, in our model, an agent can be one of six strategies: unemotional cooperator (C), unemotional defector (D), non-social emotionally non-adaptive defector (DGDN), non-social emotionally adaptive defector (DGDS), or socially emotionally adaptive defector (DGCS). The resulting payoff matrix (for row player), is as follows [4]

\[
\begin{align*}
C & | D & DGDN & DGCS \\
R & T & P & P & P & P \\
S & T & P & P & P & P \\
DGDN & T & P & P & P & P \\
DGDN & T & P & P & P & P \\
DGDN & T & P & P & P & P \\
DGCS & T & P & P & P & P \\
DGCS & T & P & P & P & P \\
DGCS & T & P & P & P & P \\
DGCS & T & P & P & P & P \\
\end{align*}
\]
Given the model and payoff matrix described above, we derive analytical conditions for when guilt-prone strategies can be viable and promote the evolution of enhanced cooperation. Furthermore, we obtain simulated numerical results for the well-mixed population setting, validating the analytical conditions. For structured populations, we run extensive agent-based simulations. For a full description of the model, as well as details on the studied network topologies and the simulations, please refer to the full version of this paper (see [4]).

3 RESULTS AND CONCLUSION

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. 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 ($\gamma_s$) is low enough (e.g., $\gamma_s = 0$ and $\gamma_s = 1$, as highlighted in Figure 1).

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) [26, 27, 30]. 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. When benefit-to-cost ratios are high, we find higher levels of cooperation in scale-free networks than in square lattices, 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 MAS of self-interested agents.

Based on psychological and evolutionary accounts of guilt and social emotions, the present paper adopts 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 work 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.

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 [2].

Overall, the present investigation has resulted in a rigorous, game-theoretical based account, of how 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. Our analysis provides novel insights into the design and engineering of self-organised and distributed cooperative multi-agent systems and how guilt-capable agents should be distributed to optimise cooperative outcomes, depending on the specific MAS network structure [1, 8, 11, 16, 20, 28, 33].

4 ACKNOWLEDGEMENTS

T.C. is supported by the John Templeton Foundation (grant no. 62281).

REFERENCES

[1] Peter Andras, Lukas Esterle, Michael Guckert, The Anh Han, Peter R Lewis, Kristina Milanovic, Terry Payne, Cedric Perret, Jeremy Pitt, Simon T Powers,