Co-evolution of Social and Non-Social Guilt in Structured Populations

Extended Abstract

Theodor Cimpeanu
University of St Andrews,
St Andrews, United Kingdom
tic1@st-andrews.ac.uk

Luís Moniz Pereira
Universidade Nova de Lisboa
Lisbon, Portugal
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 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:
Theodor Cimpeanu, Luís Moniz Pereira, and The Anh Han. 2023. Co-evolution of Social and Non-Social Guilt in Structured Populations: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

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 a 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 (DGCN), social emotionally non-adaptive defector (DGDS), or social emotionally adaptive defector (DGCS). The resulting payoff matrix (for row player), is as follows [4]

$$\begin{pmatrix} C & D & DGDN & DGCN & DGDS & DGCS \\ R & S & S & \frac{S+R\Theta}{\Omega} & S & \frac{S+R\Theta}{\Omega} \\ T & P & P & \frac{P+T\Theta}{\Omega} & P & P \\ T-\gamma & P-\gamma & P-\gamma & \frac{P+T\Theta}{\Omega}-\gamma & P-\gamma & \frac{P+T\Theta}{\Omega}-\gamma \\ \frac{T-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S\Theta}{\Omega} & \frac{P-\gamma+S\Theta}{\Omega} & \frac{P-\gamma+R\Theta}{\Omega} & \frac{P-\gamma+S\Theta}{\Omega} & \frac{P-\gamma+R\Theta}{\Omega} \\ T-\gamma-\gamma_S & P-\gamma_S & P-\gamma-\gamma_S & \frac{P+T\Theta}{\Omega}-\gamma-\gamma_S & P-\gamma-\gamma_S & \frac{P+T\Theta}{\Omega}-\gamma-\gamma_S \\ \frac{T-\gamma-\gamma_S+R\Theta}{\Omega} & P-\gamma_S & \frac{P-\gamma-\gamma_S+S\Theta}{\Omega} & \frac{P-\gamma-\gamma_S+R\Theta}{\Omega} & \frac{P-\gamma-\gamma_S+S\Theta}{\Omega} & \frac{P-\gamma-\gamma_S+R\Theta}{\Omega} \end{pmatrix}, \quad (1)$$

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.

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 (γ_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

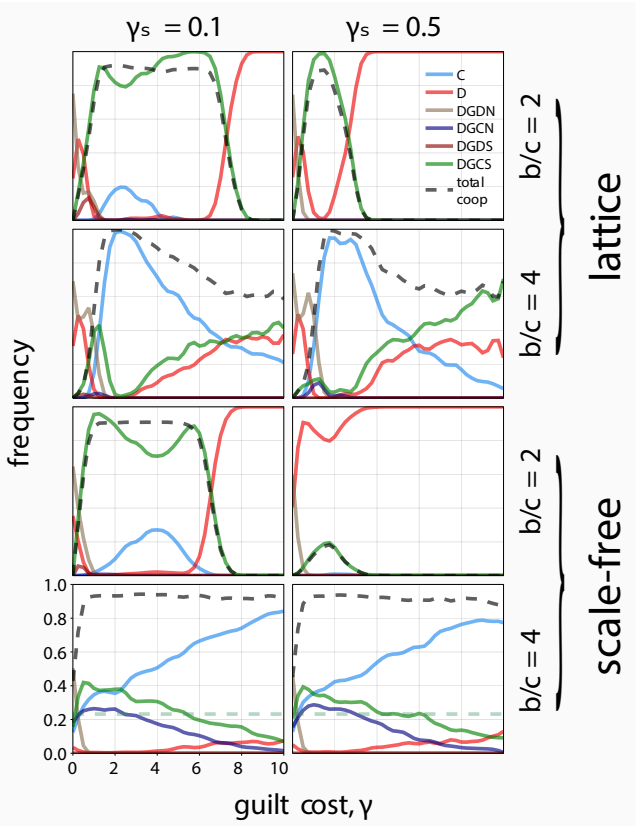

Figure 1: Strategies' frequency and the total cooperation level as a function of the guilt cost, γ.

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,

et al. 2018. Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine* 37, 4 (2018), 76–83.
[2] Albert-Laszlo Barabasi. 2014. *Linked-how Everything is Connected to Everything Else and what it Means F.* Perseus Books Group.
[3] Theodor Cimpeanu, Alessandro Di Stefano, Cedric Perret, and The Anh Han. 2023. Social diversity reduces the complexity and cost of fostering fairness. *Chaos, Solitons & Fractals* 167 (2023), 113051.
[4] Theodor Cimpeanu, Luis Moniz Pereira, and The Anh Han. 2023. Co-evolution of Social and Non-Social Guilt. *arXiv preprint arXiv:2302.09859* (2023).
[5] Theodor Cimpeanu, Francisco C Santos, Luís Moniz Pereira, Tom Lenaerts, and The Anh Han. 2022. Artificial intelligence development races in heterogeneous settings. *Scientific Reports* 12, 1 (2022), 1–12.
[6] S. A. Frank. 1998. *Foundations of social evolution.* Princeton Univ. Press, Princeton.
[7] Benoit Gaudou, Emiliano Lorini, and Eunate Mayor. 2014. Moral guilt: An agent-based model analysis. In *Advances in social simulation.* Springer, 95–106.
[8] The Anh Han. 2022. Emergent Behaviours in Multi-agent Systems with Evolutionary Game Theory. *AI Communications* 35, 4 (2022), 327 – 337.
[9] T. A. Han, T. Lenaerts, F. C. Santos, and L. M. Pereira. 2015. Emergence of cooperation via intention recognition, commitment and apology–A research summary. *AI Communications* (2015), 1–7. Issue 4.
[10] 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.
[11] The Anh Han, Cedric Perret, and Simon T. Powers. 2021. When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games. *Cognitive Systems Research* 68 (2021), 111–124.
[12] 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.
[13] B. Ho. 2012. Apologies as signals: with evidence from a trust game. *Management Science* 58, 1 (2012), 141–158.
[14] 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.
[15] Kingson Man and Antonio Damasio. 2019. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence* 1, 10 (2019), 446–452.
[16] Stacy Marsella and Jonathan Gratch. 2014. Computationally modeling human emotion. *Commun. ACM* 57, 12 (2014), 56–67.
[17] Luis A Martinez-Vaquero, The Anh Han, Luis Moniz Pereira, and Tom Lenaerts. 2017. When agreement-accepting free-riders are a necessary evil for the evolution of cooperation. *Scientific reports* 7, 1 (2017), 2478.
[18] Michael McCullough. 2008. *Beyond revenge: The evolution of the forgiveness instinct.* John Wiley & Sons.
[19] Y. Ohtsubo and E. Watanabe. 2009. Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior* 30, 2 (2009), 114–123.
[20] Ana Paiva, Fernando P Santos, and Francisco C Santos. 2018. Engineering pro-sociality with autonomous agents. In *Thirty-second AAAI conference on artificial intelligence.*
[21] Matjaž Perc, Jesús Gómez-Gardenes, Attila Szolnoki, Luis M Floría, and Yamir Moreno. 2013. Evolutionary dynamics of group interactions on structured populations: a review. *Journal of the royal society interface* 10, 80 (2013), 20120997.
[22] Luís Moniz Pereira, The Anh Han, and António Barata Lopes. 2021. Employing AI to Better Understand Our Morals. *Entropy* 24, 1 (2021), 10.
[23] 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.
[24] Luís Moniz Pereira, Ari Saptawijaya, et al. 2016. *Programming machine ethics.* Vol. 26. Springer.
[25] 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.
[26] 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.
[27] 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.
[28] Bastin Tony Roy Savarimuthu, Maryam Purvis, and Martin Purvis. 2008. Social Norm Emergence in Virtual Agent Societies. In *AAMAS '08.* 1521–1524.
[29] Nick Smith. 2008. I was wrong: The meanings of apologies. (2008).
[30] G. Szabó and G. Fáth. 2007. Evolutionary games on graphs. *Phys Rep* 97-216, 4-6 (2007).
[31] Péter Szabó, Tamás Czárán, and György Szabó. 2007. Competing associations in bacterial warfare with two toxins. *J. theor. Biol.* 248 (2007), 736–744.
[32] 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.
[33] 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