Fear of Punishment Promotes the Emergence of Cooperation and Enhanced Social Welfare in Social Dilemmas

Extended Abstract

Theodor Cimpeanu
Teesside University
T.Cimpeanu@tees.ac.uk

The Anh Han
Teesside University
T.Han@tees.ac.uk

ABSTRACT

Social punishment has been suggested as a key approach to ensuring high levels of cooperation and norm compliance in one-shot interactions. However, it has been shown that it only works when punishment is highly cost-efficient. On the other hand, signalling retribution hearkens back to medieval sovereignty, insofar as the very word for gallows in French stems from the Latin word for *power* and serves as a grim symbol of the ruthlessness of high justice. Here we introduce the mechanism of signalling an act of punishment and a special type of defector emerges, one who can recognise this signal and avoid punishment by way of fear. We perform extensive agent-based simulations so as to confirm and expand our understanding of the external factors that influence the success of social punishment. We show that our suggested mechanism serves as a catalyst for cooperation, even when signalling and punishment are very costly. We observe the preventive nature of advertising retributive acts and we contend that the resulting social prosperity is a desirable outcome in the contexts of AI and multi-agent systems. Overall, we suggest that fear acts as an effective stimulus to pro-social behaviour.

ACM Reference Format:
Theodor Cimpeanu and The Anh Han. 2020. Fear of Punishment Promotes the Emergence of Cooperation and Enhanced Social Welfare in Social Dilemmas. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

Punishment has been suggested as one of the most relevant explanations to understanding how selfish individuals self-organise and enforce cooperation or compliance to social norms in various societies [1, 9, 10, 12, 14, 16]. Numerous empirical studies show human proclivity towards punishing unjust behaviour or violations of social norms, often at great cost to their own selves [8, 9, 12]. Although in modern societies sanctioning systems have been widely implemented in the hopes of enforcing laws, many social norms continue to be upheld by the effects of private sanctions [9]. Moreover, third-party punishment has also been implemented in various online systems, such as virtual agent societies [15] or vendor marketplaces [13], as a method of enhancing pro-social behaviour and norms compliance, by both customers and sellers [13].

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

In this paper, we propose and analyse a novel approach towards explaining the evolutionary advantage of punishers in the context of anonymous interactions [16] (without relying on reputation). We make use of evolutionary game theoretic models to show that signalling acts of punishment can promote the emergence of cooperation in the selfish environment of the one-shot Prisoner's Dilemma (PD) [16]. This game is a popular underlying agent interaction framework for studying self-regarding agents and it is also the most difficult pairwise social dilemma for cooperation to emerge in. Threat of punishment can reduce defection from others without having to punish and we show that social welfare in this regime is much higher than what can be achieved with the traditional social punishment models. We provide below key results, where a comprehensive view of the outcomes of external factors, such as cost of signalling or effectiveness of punishment, can be found in [7]. We also show that expensive signalling can still provide meaningful gains to cooperation when punishing others is costly.

The effect of threat of punishment by costly signalling may provide key insights into policy making in the context of distributed systems or artificial intelligence. Indeed, it has been concluded that increasing the probability of developing super-intelligent agents is incompatible with using safety methods that incur delays or limit performance [5]. Moreover, when technological supremacy can be achieved in the short to medium term, the significant advantage gained from underestimating or even ignoring ethical and safety precautions could lead to serious negative consequences[2, 6, 11]. Our results show that threat signalling may serve as one intrinsic factor to prevent catastrophic consequences in that regard.

2 MODELS AND METHODS

We adapt the Prisoner's Dilemma (PD), first by integrating the option of costly punishment as a benchmark and we follow by describing the main model and the different configurations which we explore using agent-based simulations. Players experience, in pairs, a cooperation dilemma. In an interaction, individuals can decide whether to cooperate (play C) or defect (play D). Mutual cooperation (mutual defection) yields the reward R (penalty P), whereas unilateral defection provides a defector with the temptation T and the cooperator with the sucker's payoff S ($T > R > P > S$) ([16]). The game is considered one-shot, in other words there is no memory of past actions or prior knowledge about the interaction.

After introducing the mechanisms of punishment and costly signalling, we derive the average payoffs for each strategy in the population based on the two possible sequences of events for each

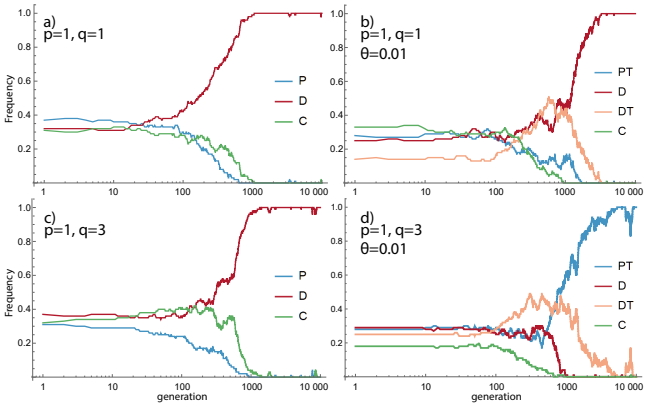

Figure 1: Typical time evolution of strategies' frequency, with and without the signalling of threat mechanism. Parameters: $T = 2$; $R = 1$; $P = 0$; $S = -1$; $\mu = 0.001$; $\beta = 1$.

agent acting out a conditional strategy (either PT players who signal acts of punishment or DT players who respond to the threat of punishment). Note that punishment incurs a cost p in order to inflict punishment q. We denote θ to be the cost of signalling an act of punishment. Contingent on which type of defector each signalling punisher encounters first, we observe that each time a PT player encounters a DT player, before encountering a normal defector, the punisher loses out on one possible act of cooperation and is therefore forced to punish a fearful defector, and vice-versa. The probability of either sequence happening first varies according to the composition of the population. Let n_1, n_2 and n_3 denote the numbers of PT, D and DT players in the population, respectively. We have $n_1 + n_2 + n_3 = N$. We denote $\Pi_{(A,B)}$ the payoff received by a player following the strategy A when facing players following strategy B (some payoffs are equivalent e.g. $\Pi_{(C,C)} = \Pi_{(PT,C)} = \Pi_{(C,PT)} = \Pi_{(PT,PT)} = R$). The average payoff can be derived, for instance, for PT: $\frac{1}{N-1}\left((n_1+n_3-1)*\Pi_{(C,C)} + n_2*\Pi_{(PT,D)} + \frac{n_3*(\Pi_{(PT,D)}-\Pi_{(C,C)})}{n_2+n_3}\right)$. Similarly, the average payoffs for other strategies, namely, C, D, and DT, can be derived. More details can be found in the full version in [7].

3 RESULTS AND CONCLUSIONS

Our comprehensive study of the external factors under which cooperation emerges, in regards to efficiency of punishment and the cost of signalling shows that fear of punishment enhances cooperation for almost all configurations (with the notable exception of highly efficient punishment coupled with expensive signalling). The results suggest that the transparency of social punishment, specifically the awareness agents have regarding acts of retribution, coupled with the ease of advertising said acts, behaves as a fulcrum towards cooperation. Fear of punishment, therefore, is most effective when awareness of who is or is not a punisher is high. On the other hand, the more deleterious an act of punishment is, the more likely it becomes for standard costly punishment to lead towards satisfactory outcomes.

We show that signalling acts as a catalyst for the emergence of cooperation when defectors are fearful of the punishers who advertise themselves as such. Furthermore, we argue that exhibiting deeds of

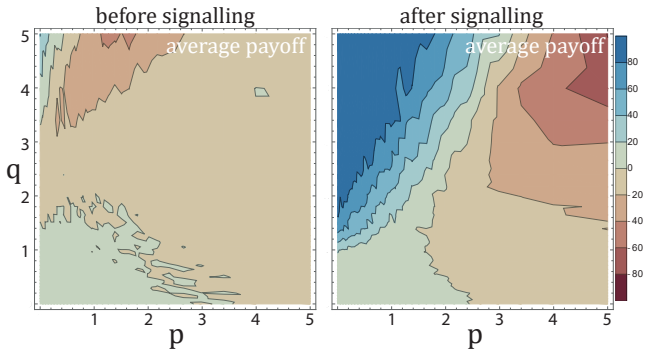

Figure 2: Social welfare of the population with varying efficiency of punishment. Parameters as in Figure 1.

punishment can explain the success of punishers, when traditional social punishment mechanisms would otherwise fail due to external factors, such as lowly efficient acts of punishment. Indeed, it seems to be the case that fearing punishment can discourage future defectors even more than the evolutionary dynamics associated with inexpensive, deleterious deeds of retribution. Moreover, we show how the traditionally damaging effects of social punishment upon social welfare can be mitigated by way of threat. Because signalling punishers cooperate indiscriminately, they outperform fearful defectors who are always vying for higher status at the expense of others, including themselves.

The prosperity of the population observed under threat of punishment speaks for the preventive nature of advertising acts of justice. Undeniably, it is a beneficial outcome for wicked ventures not to occur in the first place, but contexts such as the development of AI or climate change provide us with unparalleled incentive to prevent potentially disastrous consequences. Given the importance of intrinsic factors that guide the decisions of researchers and policy makers in the field [4], we aim to explore further how the concept of threat, and the self-preservation associated with it, could help guide the current literature on this issue. Additionally, implementing this type of signal response could improve safety conditions in multi-agent systems such as artificial societies [3, 17], especially in cases where the transparency of interactions is reduced.

These observations raise important questions around the coexistence of various types of punishers with different proclivities to signalling, as well as thresholds under which they decide that advertising their deeds of punishment would be appropriate. Reciprocally, defectors could evolve to decide when avoiding punishment is worth the act of justice and which punishers they can exploit even as they signal their propensity towards justice. Perhaps having a loud voice would be more conducive to the emergence of cooperation than the ease with which one can act revenge upon their enemies. Analytically, we have suggested the synergistic mutuality between signalling punishers and fearful defectors. We reason that the high sensitivity of defectors to signals of threat may allow less expensive signalling, whereas lowly responsive defectors require more obvious (and inherently costly) displays.

4 ACKNOWLEDGMENT

This work was supported by the Future of Life Institute (grant RFP2-154).

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] Stuart Armstrong, Nick Bostrom, and Carl Shulman. 2015. Racing to the precipice: a model of artificial intelligence development. *AI & SOCIETY* (08 2015). https://doi.org/10.1007/s00146-015-0590-y

[3] Tina Balke and Daniel Villatoro. 2012. *Operationalization of the Sanctioning Process in Utilitarian Artificial Societies*. 167–185. https://doi.org/10.1007/978-3-642-35545-5_10

[4] Seth Baum. 2016. On the promotion of safe and socially beneficial artificial intelligence. *AI & SOCIETY* 32 (09 2016). https://doi.org/10.1007/s00146-016-0677-0

[5] Nick Bostrom. 2017. Strategic Implications of Openness in AI Development. *Global Policy* (02 2017). https://doi.org/10.1111/1758-5899.12403

[6] Stephen Cave and Seán S. ÓhÉigeartaigh. 2018. An AI Race for Strategic Advantage: Rhetoric and Risks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 36–40. https://doi.org/10.1145/3278721.3278780

[7] Theodor Cimpeanu and The Anh Han. 2020. Making an Example: Signalling Threat in the Evolution of Cooperation. arXiv:cs.GT/2001.08245

[8] Dominique JF De Quervain, Urs Fischbacher, Valerie Treyer, Melanie Schellhammer, et al. 2004. The neural basis of altruistic punishment. *Science* 305, 5688 (2004), 1254.

[9] Ernst Fehr and Simon Gachter. 2002. Altruistic punishment in humans. *Nature* 415 (2002), 137–140.

[10] The Anh Han. 2016. Emergence of Social Punishment and Cooperation through Prior Commitments. In *Proceedings of the Conference of the American Association of Artificial Intelligence (AAAI'2016)*. Phoenix, Arizona, USA, 2494–2500.

[11] The Anh Han, Luís Moniz Pereira, and Tom Lenaerts. 2019. Modelling and Influencing the AI Bidding War: A Research Agenda. In *Proceedings of the AAAI/ACM conference AI, Ethics and Society*. 5–11.

[12] Benedikt Herrmann, Christian Thöni, and Simon Gächter. 2008. Antisocial Punishment Across Societies. *Science* 319 (2008), 1362–1367.

[13] Tomasz Michalak, Joanna Tyrowicz, Peter McBurney, and Michael Wooldridge. 2009. Exogenous coalition formation in the e-marketplace based on geographical proximity. *Electronic Commerce Research and Applications* 8, 4 (2009), 203–223.

[14] Simon T Powers, Daniel J Taylor, and Joanna J Bryson. 2012. Punishment can promote defection in group-structured populations. *Journal of theoretical biology* 311 (2012), 107–116.

[15] Bastin Tony Roy Savarimuthu, Maryam Purvis, Martin Purvis, and Stephen Cranefield. 2009. Social norm emergence in virtual agent societies. In *Declarative Agent Languages and Technologies VI*. Springer, 18–28.

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

[17] Daniel Villatoro, Giulia Andrighetto, Jordi Sabater-Mir, and Rosaria Conte. 2011. Dynamic sanctioning for robust and cost-efficient norm compliance. In *IJCAI*, Vol. 11. 414–419.