To Comply or Not: A Social Dynamics Analysis of Institutional Reward and Punishment for Commitment Compliance

The Anh Han

School of Computing, Engineering and Digital Technologies, Teesside University, UK

Introduction

Commitments, such as contracts and agreements, are fundamental components of many social and economic interactions, ranging from personal, to institutional, to political or religious ones, in order to ensure a mutually beneficial outcome for the parties involved (Nesse, 2001; Han et al., 2013; Sasaki et al., 2015; Sosis, 2000; Irons, 2001; Akdeniz and van Veelen, 2021; Ogbo et al., 2022; Han, 2013). Commitments are also important in the context of computerised multi-agent systems, where they are formalised as a tool for regulating agents' interactions and collective behaviour (Singh, 2013; Chopra and Singh, 2009).

It has been suggested that human specialised capacity for commitment might have been shaped by natural selection (Nesse, 2001; Sterelny, 2012). Arranging a commitment from all parties involved prior to an interaction can increase the chance of reaching mutual cooperation, enabling individuals to clarify preferences or intentions from their partners before committing to a potentially costly course of actions (Han et al., 2015; Tomasello et al., 2005; Chen and Komorita, 1994; Cherry and McEvoy, 2013; Nesse, 2001; Han et al., 2012, 2011). Since individuals can decide whether not to honour an adopted commitment—there being abundant evidence of commitment breaching in both controlled experiments and real-world scenarios—it is important to understand what mechanisms, such as positive and negative incentives, are more efficient at ensuring compliance and the cooperation-promoting benefit provided by commitments.

With this motivation, we highlight recent findings from our recent theoretical modelling analysis which comparatively explored institutional punishment of commitment violators and reward of commitment fulfillers as potential mechanisms to enhance commitment compliance and thus the overall cooperation in the population (Han, 2022). Using Evolutionary Game Theory (Sigmund, 2010; Maynard-Smith, 1982; Nowak, 2006) in the context of the one-shot Prisoner's Dilemma, the work investigated whether and when participating in a costly commitment, and complying with it, is an evolutionary stable strategy, and also results in high levels of cooperation.

Model

The model assumes a finite, well-mixed population of constant size N. At each time step, or generation, a random pair of players are chosen to play with each other. Interactions are modelled as a symmetric one-shot two-player Prisoner's Dilemma (PD) game; see below for an instance of the payoff matrix for the row player

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix}. \end{array}$$

A full set of eight possible strategies regarding commitment formation was considered in (Han, 2022). Namely, before each PD interaction, players have three decisions to make: i) whether to accept (A) or not (N) to join a prior commitment before a PD game, ii) to cooperate (C) or defect (D) in the PD if the commitment is formed, and iii) to cooperate (C) or defect (D) in the PD if the commitment is not formed.

The eight strategies are denoted as ACC, ACD, ADC, ADD, NCC, NCD, NDC and NDD. A commitment is formed when both players in a PD commit and in that case the committed players share a participation cost ϵ (Han, 2022; Han et al., 2017).

For providing institutional incentives, a per capita budget u is made available (by the institution). A fraction α of the budget can be used to reward those who are willing to participate in a commitment (i.e. players who adopt either ACC, ACD, ADC or ADD), hoping to increase the chance a commitment being formed. The remaining budget, i.e. $(1-\alpha)u$, is used for rewarding commitment compliant players (i.e. ACC and ACD players) or punishing non-compliant ones (i.e. ADC and ADD players). When $\alpha = 0$, it means the budget is used only for incentivising commitment compliant behaviours (i.e., pure reward and pure punishment scenarios, see Figure 1B).

Results and Discussion

The results show that, given a sufficient budget for providing incentives (i.e. sufficient u), rewarding of commitment

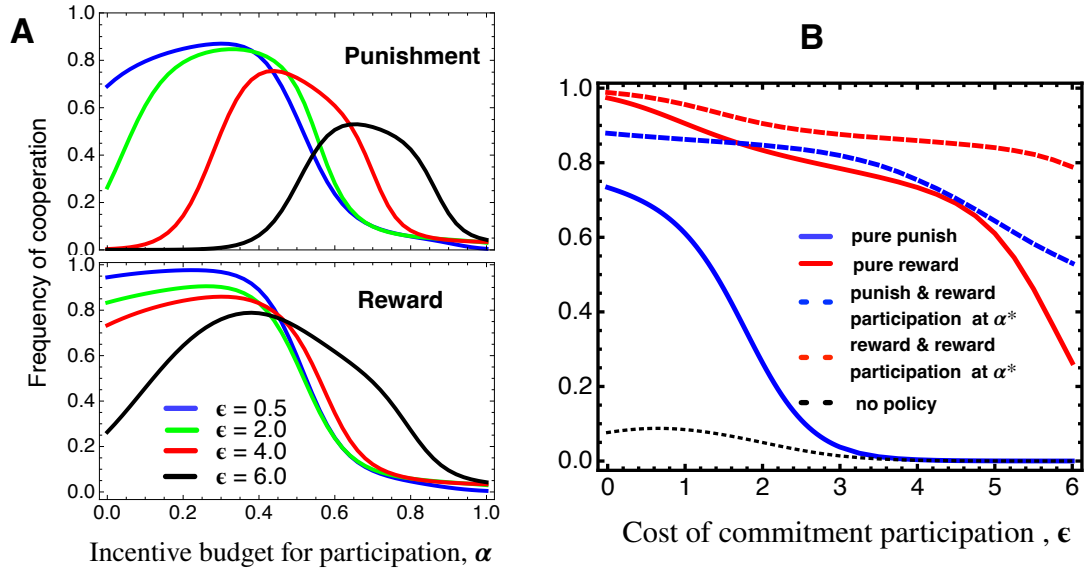

Figure 1: **(A) Rewarding participation can improve cooperation despite reducing the budget for incentivising commitment compliant behaviour.** Shown is the frequency of cooperation as a function of the fraction of the budget for rewarding of participation (α), for different values of the cost of commitment participation ϵ. **(B) Reward promotes higher levels of cooperation than punishment**. The value $\alpha^\star$ denote the optimal fraction of the budget for rewarding participation. Other parameters: Population size $N = 100$; incentive budget, $u = 2$. See Han (2022) for a robust analysis and analytical conditions.

compliant behaviours better promotes cooperation than punishment of non-compliant ones, see Figure 1. Reward can ensure commitment compliance to be evolutionarily viable for a larger range of the commitment cost (ϵ). This observation also holds for varying the benefit-to-cost ratio defined by the PD payoff matrix, see (Han, 2022) for details. This finding has useful implications for the design of institutional mechanisms for promoting pro-social behaviour, especially when communication is allowed to establish prior commitments (Nesse, 2001).

Moreover, participating in a commitment can be quite costly and that might discourage players from joining the commitment in the first place. Examples of incentives for encouraging participation are many, and have been shown to be crucial for ensuring a positive outcome, as in the contexts of climate change agreements and healthcare programmes (Tappin et al., 2015; Bruni et al., 2009). The results show that, by sparing part of the incentive budget ($\alpha > 0$) for rewarding those willing to participate in a commitment, the overall level of cooperation can be significantly enhanced for both reward and punishment, see Figure 1A-B. Another key finding from is that, the presence of mistakes in deciding to participate in a commitment favours evolutionary stability of commitment compliance and cooperation.

Relevant to the presented model, evolutionary modelling and analysis of voluntary participation has been considered in several studies (De Silva et al., 2010; Mathew and Boyd, 2009; Sasaki et al., 2012; Hauert et al., 2007; Sigmund et al., 2010), showing that cooperation can evolve even in one-shot cooperation dilemmas if players have the option to opt out. However, these works did not consider strategies conditioned on the formation of a commitment, nor incentives for encouraging the participation in it. Thus, the presented model here offers a more complete picture of how prior commitments such as formal and informal contracts and agreements, provides an efficient mechanism for promoting the evolution of cooperation.

In short, we have summarised here a analysis of different forms of institutional incentive for promoting participation and compliance in interactions with a prior commitment formation (Han, 2022). The results highlight the efficiency of the reward mechanism in contrast to the punishing one, and also the importance of incentivising participation, in promoting the evolution and stability of cooperation in social dilemmas. That said, using incentives for ensuring participation is as important as for enhancing compliance.

It is noteworthy that institutional approaches have been widely adopted to study biological and artificial life systems (Andras et al., 2018; Powers et al., 2018; Duong and Han, 2021; Ostrom, 1990; Han et al., 2022; Perc et al., 2017; Cimpeanu et al., 2019). The summarised analysis provides new, fundamental insights into a cost-efficient design of commitment- and institution-based solutions for promoting pro-social behaviours and behavioural compliance in biological, social as well as artificial systems (Singh, 2013; Nesse, 2001; Ostrom, 1990).

References

Akdeniz, A. and van Veelen, M. (2021). The evolution of morality and the role of commitment. *Evolutionary Human Sciences*, pages 1–53.

Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S. T., et al. (2018). Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine*, 37(4):76–83.

Bruni, M. L., Nobilio, L., and Ugolini, C. (2009). Economic incentives in general practice: the impact of pay-for-participation and pay-for-compliance programs on diabetes care. *Health policy*, 90(2-3):140–148.

Chen, X.-P. and Komorita, S. S. (1994). The effects of communication and commitment in a public goods social dilemma. *Organizational Behavior and Human Decision Processes*, 60(3):367–386.

Cherry, T. L. and McEvoy, D. M. (2013). Enforcing compliance with environmental agreements in the absence of strong institutions: An experimental analysis. *Environmental and Resource Economics*, 54(1):63–77.

Chopra, A. K. and Singh, M. P. (2009). Multiagent commitment alignment. In *AAMAS'2009*, pages 937–944.

Cimpeanu, T., Han, T. A., and Santos, F. C. (2019). Exogenous rewards for promoting cooperation in scale-free networks. In *Artificial Life Conference Proceedings*, pages 316–323. MIT Press.

De Silva, H., Hauert, C., Traulsen, A., and Sigmund, K. (2010). Freedom, enforcement, and the social dilemma of strong altruism. *Journal of Evolutionary Economics*, 20(2):203–217.

Duong, M. H. and Han, T. A. (2021). Cost efficiency of institutional incentives for promoting cooperation in finite populations. *Proceedings of the Royal Society A*, 477(2254):20210568.

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

Han, T. A. (2022). Institutional incentives for the evolution of committed cooperation: ensuring participation is as important as enhancing compliance. *Journal of The Royal Society Interface*, 19(188):20220036.

Han, T. A., Lenaerts, T., Santos, F. C., and Pereira, L. M. (2022). Voluntary safety commitments provide an escape from over-regulation in ai development. *Technology in Society*, 68:101843.

Han, T. A., Pereira, L. M., and Lenaerts, T. (2017). Evolution of commitment and level of participation in public goods games. *Autonomous Agents and Multi-Agent Systems*, pages 1–23.

Han, T. A., Pereira, L. M., and Santos, F. C. (2011). Intention recognition promotes the emergence of cooperation. *Adaptive Behavior*, 19(3):264–279.

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

Han, T. A., Pereira, L. M., Santos, F. C., and Lenaerts, T. (2013). Good agreements make good friends. *Scientific reports*, 3(2695).

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

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

Irons, W. (2001). Religion as a hard-to-fake sign of commitment. In Nesse, R. M., editor, *Evolution and the capacity for commitment*, pages 292–309. New York: Russell Sage.

Mathew, S. and Boyd, R. (2009). When does optional participation allow the evolution of cooperation? *Proceedings of the Royal Society B: Biological Sciences*, 276(1659):1167–1174.

Maynard-Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press, Cambridge.

Nesse, R. M. (2001). *Evolution and the capacity for commitment*. Foundation series on trust. Russell Sage.

Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the Equations of Life*. Harvard University Press, Cambridge, MA.

Ogbo, N. B., Elragig, A., and Han, T. A. (2022). Evolution of coordination in pairwise and multi-player interactions via prior commitments. *Adaptive Behavior*, 30(3):257–277.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.

Perc, M., Jordan, J. J., Rand, D. G., Wang, Z., Boccaletti, S., and Szolnoki, A. (2017). Statistical physics of human cooperation. *Phys Rep*, 687:1–51.

Powers, S. T., Ekárt, A., and Lewis, P. R. (2018). Modelling enduring institutions: The complementarity of evolutionary and agent-based approaches. *Cognitive Systems Research*, 52:67–81.

Sasaki, T., Brännström, Å., Dieckmann, U., and Sigmund, K. (2012). The take-it-or-leave-it option allows small penalties to overcome social dilemmas. *Proceedings of the National Academy of Sciences*, 109(4):1165–1169.

Sasaki, T., Okada, I., Uchida, S., and Chen, X. (2015). Commitment to cooperation and peer punishment: Its evolution. *Games*, 6(4):574–587.

Sigmund, K. (2010). *The Calculus of Selfishness*. Princeton University Press.

Sigmund, K., Silva, H. D., Traulsen, A., and Hauert, C. (2010). Social learning promotes institutions for governing the commons. *Nature*, 466:7308.

Singh, M. P. (2013). Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):21.

Sosis, R. (2000). Religion and intra-group cooperation: preliminary results of a comparative analysis of utopian communities. *Cross-Cultural Research*, 34:70–87.

Sterelny, K. (2012). *The evolved apprentice*. MIT Press.

Tappin, D., Bauld, L., Purves, D., Boyd, K., Sinclair, L., MacAskill, S., McKell, J., Friel, B., McConnachie, A., De Caestecker, L., et al. (2015). Financial incentives for smoking cessation in pregnancy: randomised controlled trial. *Bmj*, 350.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(05):675–691.