Evolutionary mechanisms that promote cooperation may not promote social welfare

The Anh Han[1,⋆], Manh Hong Duong[2], Matjaz Perc[3,4,5,6]

[1] School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK
[2] School of Mathematics, University of Birmingham, Birmingham, UK
[3] Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia
[4] Community Healthcare Center Dr. Adolf Drolc Maribor, Maribor, Slovenia
[5] Complexity Science Hub Vienna, Vienna, Austria
[6] Department of Physics, Kyung Hee University, Seoul, Republic of Korea

⋆ Corresponding: The Anh Han (t.han@tees.ac.uk)

Abstract

Understanding the emergence of prosocial behaviours among self-interested individuals is an important problem in many scientific disciplines. Various mechanisms have been proposed to explain the evolution of such behaviours, primarily seeking the conditions under which a given mechanism can induce highest levels of cooperation. As these mechanisms usually involve costs that alter individual payoffs, it is however possible that aiming for highest levels of cooperation might be detrimental for social welfare – the later broadly defined as the total population payoff, taking into account all costs involved for inducing increased prosocial behaviours. Herein, by comparatively analysing the social welfare and cooperation levels obtained from stochastic evolutionary models of two well-established mechanisms of prosocial behaviour, namely, peer and institutional incentives, we demonstrate exactly that. We show that the objectives of maximising cooperation levels and the objectives of maximising social welfare are often misaligned. We argue for the need of adopting social welfare as the main optimisation objective when designing and implementing evolutionary mechanisms for social and collective goods.

Keywords: Social welfare, cost efficiency, reward, punishment, evolution of cooperation, social dilemma, evolutionary dynamics.

1 Introduction

Since Darwin, the challenge of explaining the evolution of cooperative behavior has been actively explored across various fields, including Evolutionary Biology, Ecology, Economics, and Multi-agent Systems (Han, 2022a, Nowak, 2006b, Paiva et al., 2018, Perc et al., 2017, Sigmund, 2010). Several mechanisms have been proposed to account for the evolution of cooperation, such as kin and group selection, direct and indirect reciprocity, structured populations, pre-commitments and incentives (Novak, 2006, Perc et al., 2017). Therein, the emphasis is often placed on the degree or level of cooperation that a given mechanism can induce.

However, these mechanisms typically involve costs that alter payoffs, either for the individuals involved in the interactions or for a third party (such as an institution) interested in promoting cooperation within the population. This can lead to a reduction in the overall social welfare of the population, broadly defined here as the total payoff of the population (Kaneko and Nakamura, 1979), including all costs associated with inducing behavioral changes. For example, let us consider peer incentives, where an agent can choose to pay a personal cost to decrease (peer punishment) or increase (peer reward) the payoff of the incentive recipient (Boyd et al., 2003, Fehr and Gächter, 2002, Han, 2016, Sigmund et al., 2001). Typically (and intuitively), peer punishment is considered more efficient than peer reward as the former can lead to a higher level of cooperation, since peer punishers are more advantageous than peer rewarders when playing against defectors (see also our results in Figure 1). However, given that cooperative players gain an increase in payoffs when playing with rewarders, compared to no increase when playing with punishers, the overall population payoffs might be higher under peer reward even when it has a lower level of cooperation. We discuss the importance of considering social welfare for other mechanisms of prosocial behaviours and for various real-world application domains in Discussion (Section 4).

In this paper, we demonstrate that it might be more important to optimise the social welfare, rather than focusing entirely on achieving highest levels of cooperation. Because the latter can lead to a misleading, undesirable outcome where a high cooperation level is achieved but social welfare decreases. We demonstrate these through analysing social welfare for two well-established classes of incentive mechanisms: peer and institutional incentives, for both positive (i.e. reward) and negative (i.e. punishment) types (Duong and Han, 2021, Sasaki et al., 2015, Sigmund et al., 2001, Van Lange et al., 2014).

We adopt Evolutionary Game Theory (EGT) (Imhof et al., 2005, Sigmund, 2010, Smith, 1974), a well-established mathematical framework for modelling and analysing cooperative behaviours and their emergence and stability (Nowak, 2006b, Perc et al., 2017). We derive close forms for the long-term expected social welfare, for population dynamics under varying mutation rates and selection intensities, which are key factors of Darwinian evolution (Nowak, 2006a). Our analysis is carried out using the one-shot Prisoner's Dilemma, a well adopted game for modelling a social dilemma of cooperation (Coombs, 1973, Sigmund, 2010).

In the next section we describe the models and methods, including derivations of social welfare and institutional costs. Results and Discussion sections will follow. We also provide additional results in the Supporting Information.

2 Model and Methods

2.1 Prisoner's Dilemma

We consider a well-mixed population where all players interact with each other via the one-shot Prisoner's Dilemma (PD) game, choosing whether to cooperate (C) or to defect (D), with payoffs given by the following payoff matrix:

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

If both interacting players follow the same strategy, they receive the same payoff: R for mutual cooperation and P for mutual defection. If the agents play different strategies, the cooperator gets the sucker's payoff S, and the defector gets the temptation to defect T. The payoff matrix corresponds to the preferences associated with the PD when the parameters satisfy the ordering $T > R > P > S$ (Coombs, 1973).

2.2 Evolutionary processes

We consider an evolutionary process of a well-mixed, finite population of N interacting individuals (players). The players can adopt one of m strategies, 1, ..., m. The set of possible states of the population is

$$\Delta_N^m := \{\mathbf{n} = (n_1, \ldots, n_m) |\ 0 \le n_i \le N, \sum_{i=1}^{m} n_i = N\}, \tag{1}$$

where n_i is the number of players currently adopting strategy i ($i = 1, \ldots, m$). In each time step of the evolutionary process, an individual A is chosen at random to update their strategy. There are two ways to do so: with probability μ (mutation rate), A adopts a randomly selected strategy from the remaining $m - 1$ strategies. With probability $1 - \mu$, the update happens through social learning, whereby the most successful strategies tend to be imitated more often by other players (this process is equivalent to biological reproduction). That is, A adopts the strategy of another, randomly chosen from the population, player B, with a probability given by $P_{A,B}$. A popular approach is to use the Fermi distribution from statistical physics:

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

where f_A (f_B) denotes the fitness of individual A (B) and β measures the strength of the fitness contribution to the update process (a.k.a. intensity of selection). This approach leads to a unified framework for evolutionary dynamics at all intensities of selection, from random drift to imitation dynamics (Imhof et al., 2005, Traulsen and Nowak, 2006).

This elementary updating process, involving mutation and imitation, is then iterated over many time steps. As a result, we obtain an ergodic process on the space of all possible population states. This evolutionary process defines a Markov chain with state space Δ_N^m. The equilibrium of this Markov process, known as the mutation-selection distribution, is a fundamental object to quantify the evolutionary dynamics in finite populations describing the fraction of time the population spends in each population state in the long term. Understanding this equilibrium is a challenging problem due to the complexity of this calculation given the size of the transition matrix.

The number of states in the Markov chain is

$$S = |\Delta_N^m| = \binom{N+m}{m}.$$

For the transition probabilities of the Markov chain, for any two population states $\mathbf{n}$ and $\mathbf{n}'$ in an evolutionary process of size S, the transition probability to move from $\mathbf{n}$ to $\mathbf{n}'$ in one step of the process is given by

$$\omega_{\mathbf{n},\mathbf{n}'} = \begin{cases} \frac{n_i}{S}\left(\frac{\mu}{m-1} + (1-\mu)\frac{n_j}{S}P_{i,j}\right) & \text{if } n_i' = n_i - 1,\ n_j' = n_j + 1, n_l' = n_l \text{ for } l \notin \{i,j\}, \\ 1 - \sum_{j \neq i} \frac{n_i}{S}\left(\frac{\mu}{m-1} + (1-\mu)\frac{n_j}{S}P_{i,j}\right) & \text{if } \mathbf{n} = \mathbf{n}', \\ 0 \quad \text{otherwise.} \end{cases} \tag{2}$$

By computing the normalised left eigenvector of the transition matrix with respect to eigenvalue 1, we obtain the corresponding mutation-selection (stationary) distribution.

2.2.1 Strategy frequency

The frequency of strategy i (e.g. cooperation) is obtained by taking the average over all possible states $\mathbf{n}$ and weighting it with the corresponding stationary distribution $\bar{p}_{\mathbf{n}}$

$$f_i = \sum_{\mathbf{n}} \frac{\mathbf{n}_i \bar{p}_{\mathbf{n}}}{N}, \tag{3}$$

where $\mathbf{n}_i$ represents the quantity of individuals with strategy i in state $\mathbf{n}$.

2.2.2 Social welfare

Similarly, the total population payoff (social welfare), SW, is given as follows

$$SW = \sum_{\mathbf{n}} \frac{SW(\mathbf{n})\bar{p}_{\mathbf{n}}}{N}, \tag{4}$$

where $SW(\mathbf{n})$ is the population total payoff when the population is in state $\mathbf{n}$.

2.3 Social welfare with external intervention

We assume that there is an external party (i.e. an institution) that aims to promote a certain behavioural profile (Duong and Han, 2021, Han and Tran-Thanh, 2018, Wang et al., 2019). Now, when optimising social welfare one also needs to take into account the costs spent by the third party.

To keep it general (institutional incentives that we analyse in the present work are a special case), we assume that at state $\mathbf{n} = (n_1, \ldots, n_m)$, an institutional incentive budget $\theta_{\mathbf{n}}$ can be used to promote a certain objective (e.g. maximising or ensure a certain threshold of the total frequency of cooperation). We write $\theta_{\mathbf{n}} = \sum_{i=1}^{m} n_i \theta_i$, where θ_i is the per capita budget for strategist i at state $\mathbf{n}$, which can be used for either reward or punishment.

We denote $\Theta = \{\theta_{\mathbf{n}}\}_{\mathbf{n} \in \Delta_N^m}$ the overall incentive policy. The expected cost for providing incentive per evolutionary step is given by

$$E(\Theta) = \sum_{\mathbf{n}} \theta_{\mathbf{n}} \bar{p}_{\mathbf{n}}. \tag{5}$$

Thus, the total social welfare can be rewritten as follows:

$$SW(\Theta) - E(\Theta).$$

Note that the population payoff depends on incentive policy Θ. Namely, it alters the transition probabilities given in Equation 2 (more concretely, the terms $P_{i,j}$) and thus the stationary distribution.

3 Results

3.1 Peer incentives

A peer (social) punisher (SP) and peer (social) rewarder (SR) cooperates in the PD, and after the PD game, they pay a cost ϵ to punish a defective co-player or reward a cooperative one, respectively. The rewarded/punished player receives an increase/decrease of δ in their payoff.

We consider minimal models of peer incentives in the one-shot PD game, with three strategies: unconditional cooperator (C), unconditional defector (D), and either SP or SR. The payoff matrices for peer punishment and peer reward cases are given as follows, respectively

$$\begin{array}{c} \\ C \\ D \\ SP \end{array} \begin{array}{c} \begin{array}{ccc} C & D & SP \end{array} \\ \begin{pmatrix} R & S & R \\ T & P & T-\delta \\ R & S-\epsilon & R \end{pmatrix} \end{array}, \quad \begin{array}{c} \\ C \\ D \\ SR \end{array} \begin{array}{c} \begin{array}{ccc} C & D & SR \end{array} \\ \begin{pmatrix} R & S & R+\delta \\ T & P & T \\ R-\epsilon & S & R-\epsilon+\delta \end{pmatrix} \end{array}.$$

We now comparatively study the efficiency of the two peer incentive approaches, for promoting cooperation and enhancing the social welfare, focusing on whether these two objectives are aligned. Indeed, Figure 1 (left column) shows that, as expected, peer punishment usually surpasses peer reward in effectively promoting cooperation, especially when selection is weaker and the impact to cost ratio of incentive is sufficiently high. However, interestingly, reward leads to higher social welfare than punishment in most cases

For peer reward, increasing efficiency leads to increase in social welfare in general. It is however not the case for punishment, where social welfare usually decreases with the impact to cost ratio. That is, applying peer punishment is often detrimental for social welfare.

Overall, our results have shown that, in case of peer punishment, the objective of promoting the evolution of high levels of cooperation can be detrimental for social welfare. Peer reward, on the other hand, is more efficient in promoting social welfare, even though it leads to lower levels of cooperation than punishment. The observations are robust for varying mutation rates, see Figure S3.

It is noteworthy that our analysis focused on the minimal models of peer incentives, where cooperation is generally promoted for favourable conditions of incentives (i.e. sufficiently high impact to cost ratio). These settings are suitable for the purpose of our study, as we aim to demonstrate that achieving a high level cooperation could potentially be detrimental to social

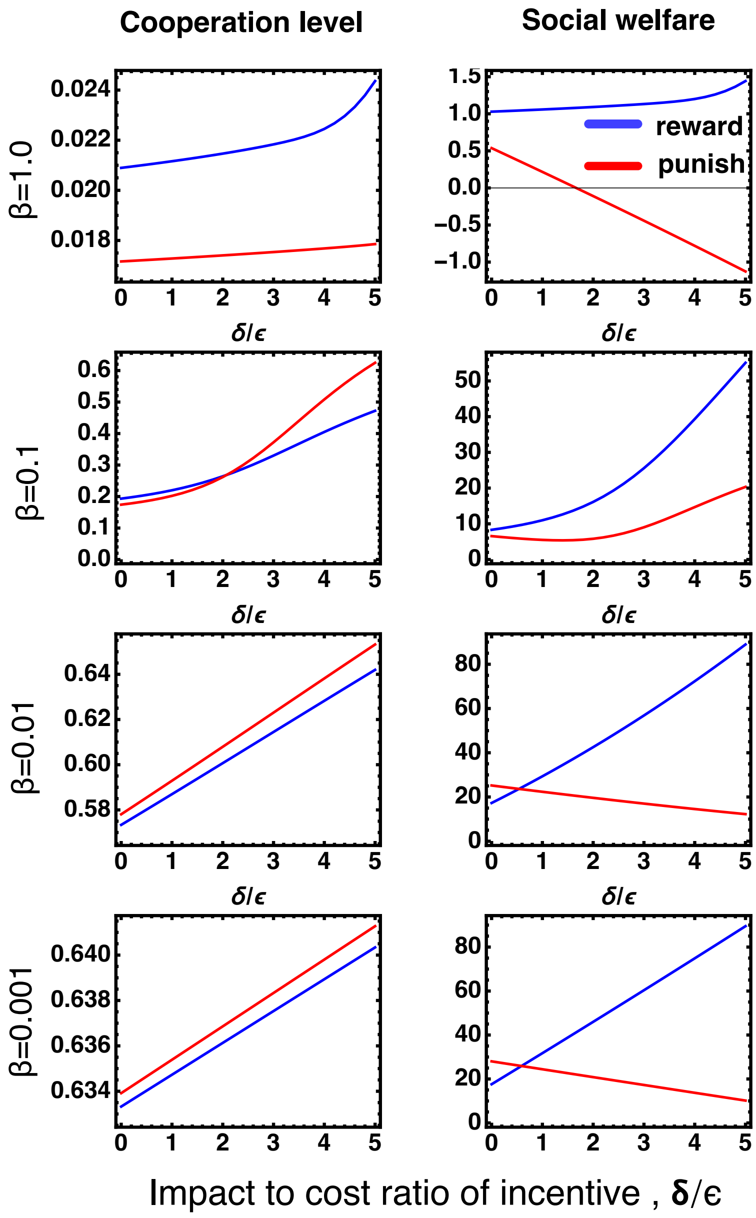

Figure 1. Impact of peer reward vs peer punishment for the long-term level of cooperation (f_C, see Equation 3) and population social welfare (SW, see Equation 4), for varying the efficiency of incentive δ/ϵ, for different values of the intensities of selection β. We observe that punishment is better than reward for promoting cooperation in most cases, especially for weaker selection and when the impact to cost ratio of incentive is sufficiently high. However, reward leads to higher social welfare than punishment in most cases. Parameters: Population size, $N = 50$, mutation rate $\mu = 0.01$, cost of peer incentive $\epsilon = 1$, Prisoner's Dilemma payoff matrix $R = 1, S = -1, T = 2, P = 0$.

welfare. It would be interesting to examine extended models of peer incentives for example when antisocial incentives (i.e. punishing cooperators and rewarding defectors) are included (Han, 2016, Herrmann et al., 2008, Rand et al., 2010), or when incentives are provided in a conditional or stochastic approach (Chen et al., 2014, Cimpeanu and Han, 2020, Xiao et al., 2023).

3.2 Institutional incentives

We now analyse the alignment between promoting cooperation and social welfare in minimal models of institutional reward and punishment (Duong and Han, 2021, Góis et al., 2019, Sasaki et al., 2015). Namely, we consider a well-mixed population consisting of individuals playing the one-shot PD game who can adopt either C or D in the interactions.

The social welfare now needs to take into account the cost for providing incentives spent by the external institution for promoting cooperation. The institution might have different levels of efficiency using the budget, which can also be different for implementing reward or punishment.

To reward a cooperator (to punish a defector), the institution has to pay an amount θ/a (θ/b, respectively) so that the cooperator's (defector's) payoff increases (decreases) by θ, where $a, b > 0$ are constants representing the efficiency ratios of providing this type of incentive.

Figure 2 shows that, as expected (Duong and Han, 2021, Góis et al., 2019, Han, 2022b), both types of incentives lead to the same level of cooperation assuming the are equally effective (i.e. $a = b$) and costly. However, institutional reward leads to positive social welfare (red areas in panels a and b) for a much wider range of incentive impact and cost. For reward, even when using incentive is rather cost-inefficient, e.g. when $a = 0.7$ (panel c), positive social welfare can be achieved for intermediate values of δ. While for punishment, it needs to be highly efficient ($b > 2$) for positive social welfare. Our observation are robust for different intensities of selection, see Supporting Information Figure S1.

For cost-efficient institutional reward (i.e., $a > 1$), it is best for social welfare to maximise the impact (i.e. strong reward with a high per capital budget); for $a = 1$, the impact δ needs to reach a certain threshold and then the performance levels up. For lower a, there is an optimal threshold of δ for highest social welfare.

Sufficiently strong institutional punishment (see panel c, with $b = 3$) can lead to positive social welfare, but even in this case it can be detrimental for social welfare when imposing a larger impact on defectors. There is an optimal intermediate δ for highest social welfare.

This is a notable observation as previous models of institutional incentives (Góis et al., 2019, Sasaki et al., 2012, 2015, Sigmund et al., 2010) do not and are unable to provide insights on how strong punishment is strong enough, focusing on promoting high levels of cooperation. In fact, previous works only consider that strong punishment is needed to ensure cooperation.

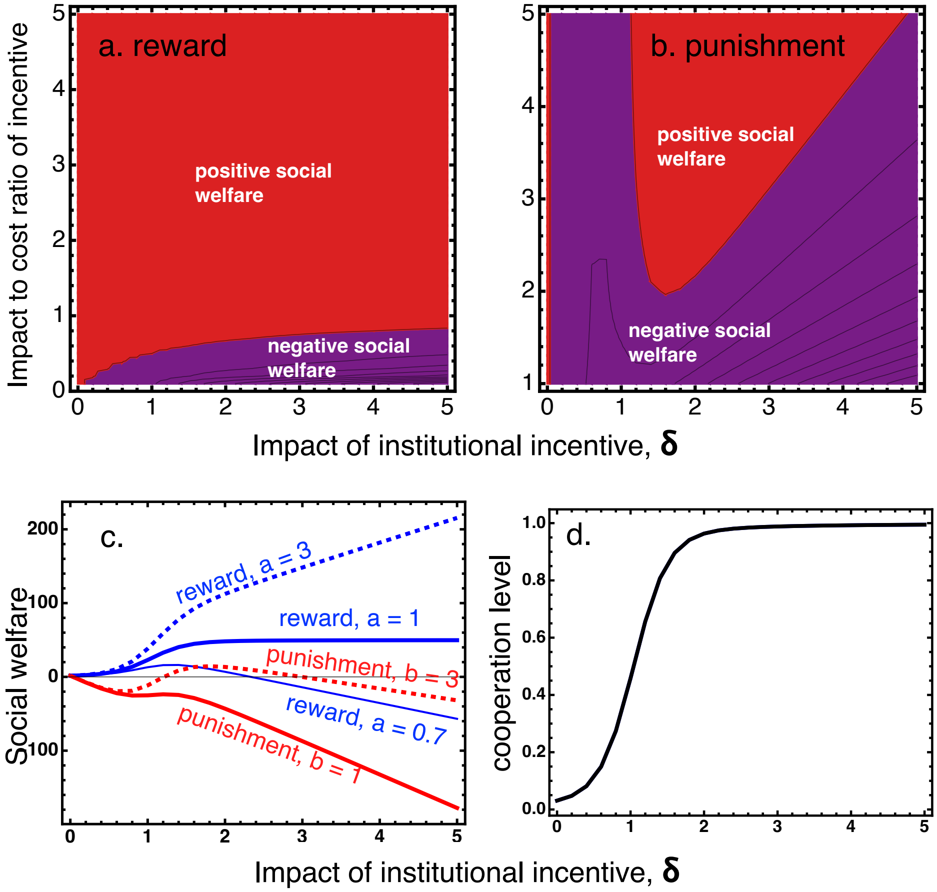

Figure 2. Impact of institutional reward vs institutional punishment for the long-term level of cooperation and population social welfare. We observe that although both types of incentives lead to the same level of cooperation given the same incentive impact on the incentive recipient δ (assuming their impact to cost ratios are the same, i.e. $a = b$, see panel d), reward leads to positive social welfare for a much wider range of parameters (compare red areas in panels a and b). Parameters: Population size, $N = 50$, mutation rate $\mu = 0.001$, intensity of selection $\beta = 0.1$, Prisoner's Dilemma with $R = 1, S = -1, T = 2, P = 0$.

Furthermore, we can observe that, for some value of a and b, optimal social welfare is achieved when cooperation is not at its highest possible level (i.e. 100% in this case). For example, for a $a = 0.7$, optimal social welfare is achieved when $\delta \approx 1.2$, and for $b = 3$, when $\delta \approx 1.6$. The corresponding levels of cooperation for those values are approximately 0.5 and 0.8. These clearly demonstrate that optimising cooperation and social welfare might not be always aligned.

4 Discussion

Over the past decades, significant attention has been given to studying effective incentive mechanisms that promote the evolution of cooperation in social dilemmas (Novak, 2006, Perc et al., 2017, Sigmund, 2010). The emphasis is often placed on the extent of cooperation that a given mechanism can induce, and the conditions regarding the parameters involved. Since mutual cooperation is collectively more desirable than mutation defection, ensuring high levels of cooperation usually also leads to high population welfare.

However, as these mechanisms usually involve costs that alter individual payoffs, it is possible that aiming for highest levels of cooperation might be detrimental for social welfare. In this paper, using numerical simulations for stochastic evolutionary models for two important incentive mechanisms, peer and institutional incentives, we have demonstrated exactly that.

Closely related to the present work are a recent body of literature on cost optimisation of institutional incentives for promoting cooperation and fairness (Cimpeanu et al., 2023, Cimpeanu and Han, 2024, Cimpeanu et al., 2021, Duong et al., 2023, Duong and Han, 2021, Han and Tran-Thanh, 2018, Sun et al., 2021, Wang et al., 2019, 2022). These works usually consider a bi-objective optimisation problem where one aims to ensure a certain minimal level of cooperation at a smallest cost to the institution. While these works can provide insights on the optimal incentive policy, they still do not guarantee optimal social welfare. In fact, the two objectives are often misaligned (see Supporting Information, Figure S2). Moreover, we argue that focusing on the social welfare provides a more convenient, single objective optimisation problem.

We analysed here two incentive mechanisms. It might be interesting to study whether maximising social welfare is aligned with maximising cooperation for other mechanisms of cooperation. For example, with kin selection, favouring related players might lead to reduction of social welfare due to unfair use of power to favour or patronage one's relatives or friends (aka nepotism) (Wilson and Dugatkin, 1991). For indirect reciprocity to work (Okada, 2020), one might need to enhance transparency of reputation, e.g. via implementing institution broadcasting reputation scores (Radzvilavicius et al., 2021) or other costly communication mechanisms (Krellner et al., 2021, Santos et al., 2018). Thus one might need to balance between promoting cooperation and the cost of enabling it. For direct reciprocity to work (Xia et al., 2023), one might need to reduce noise, increase cognitive capacity (Han et al., 2012, Lenaerts et al., 2024), and improve trust (Han et al., 2021), which are costly and thus need to be taken into account for balancing the population social welfare. For pre-commitment to work (Nesse, 2001), it usually requires a third party such as an institution that provides incentives (Han, 2022b) or broadcast commitment-based reputation (Krellner and Han, 2023) for ensuring commitment compliance, which are costly and needed to be taken into account for enhanced social welfare.

It is noteworthy that our study focused on cooperation, but the same argument would be applicable for other prosocial behaviours such as coordination, trust, fairness, moral behaviour, technology safety development, collective risk avoidance and pandemic intervention compliance (Andras et al., 2018, Capraro and Perc, 2021, Cimpeanu et al., 2021, Han et al., 2020, Santos and Pacheco, 2011, Santos et al., 2006, Traulsen et al., 2023). It would be interesting to re-examine existing evolutionary mechanisms for such prosocial behaviours to see whether they promote social welfare.

Beyond prosocial behaviours such as cooperation, it is often unclear or debatable which behaviour or social norm should be promoted. In these cases, using social welfare as the optimisation objective can be particularly convenient, facilitating integrated decision-making that aims for the overall social good, especially in complex scenarios with multiple and sometimes conflicting priorities.

Such scenarios are common. For example, when designing public health programmes, determining if vaccination should be the top priority can be challenging. Utilising social welfare as the objective allows policymakers to focus on the overall health and well-being of the population. Similarly, in education, opinions may differ on whether to prioritise STEM education, arts and humanities, vocational training, or critical thinking skills. Optimising for social welfare ensures the education system offers a balanced curriculum that supports the diverse talents and interests of students, fostering their overall development. Another example in environment domains with debates about prioritising the reduction of carbon emissions, the preservation of biodiversity, the promotion of renewable energy, or ensuring clean water access. By focusing on social welfare, environmental policies can be created to balance these various goals, resulting in comprehensive strategies that benefit both the environment and the population.

An important ongoing initiative in Artificial Intelligence (AI) research is the design and implementation of AI for social goods (Tomašev et al., 2020), using AI-based solutions for effectively addressing social problems. To this end, there is an emerging body of evolutionary modelling studies that address prosocial behaviours in hybrid systems of human and AI agents in co-presence (Capraro et al., 2024, Fernández Domingos et al., 2022, Guo et al., 2023, Han, 2022a, Paiva et al., 2018, Santos, 2024, Shen et al., 2024, Zimmaro et al., ress). As developing AI is costly, it is crucial to understand what kind of AI are most conducive for prosocial behaviours, in a cost effective way. Thus, optimising the overall system payoff or social welfare is crucial for effective use of AI for social goods.

Data Accessibility

This work does not contain any data.

Declaration of AI use

We have not used AI-assisted technologies in creating this article.

Author Contributions

T.A.H: conceptualization, formal analysis, investigation, methodology, software, visualisation, validation, writing—original draft; M.H.D: conceptualization, formal analysis, investigation, methodology, validation and writing—original draft; M.P: investigation, validation, and writing - review & editing; All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Competing Interests

Authors declare no competing interests.

Acknowledgements

T.A.H. is supported by EPSRC (grant EP/Y00857X/1) and the Future of Life Institute. M.H.D is supported by EPSRC (grant EP/Y008561/1) and a Royal International Exchange Grant IES-R3-223047. M.P. is supported by the Slovenian Research and Innovation Agency (Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije) (grants P1-0403 and N1-0232).

Supporting Information

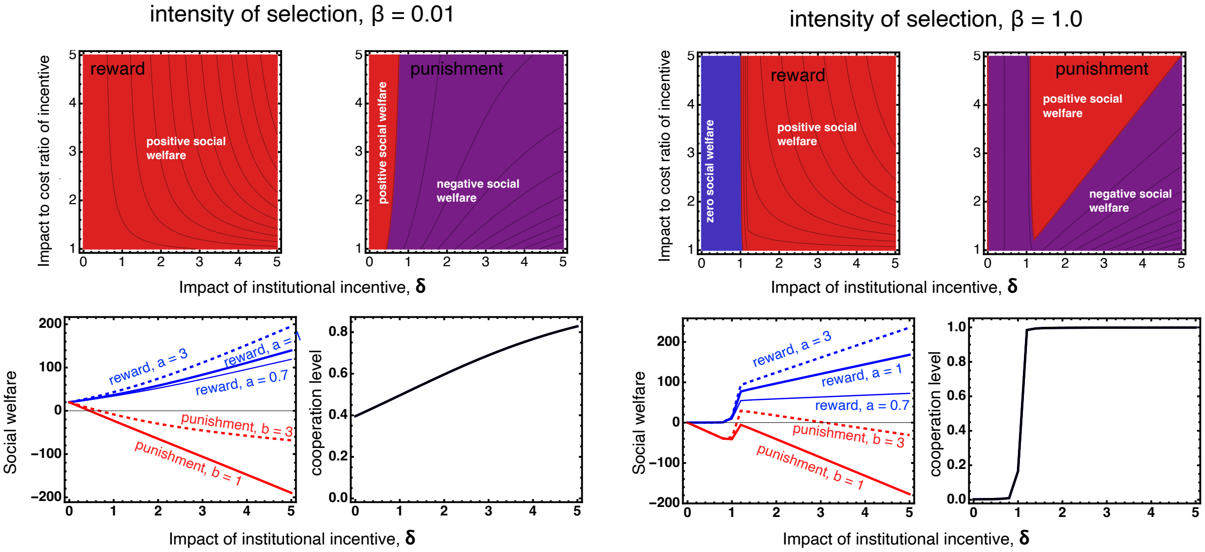

Figure S1. Impact of institutional reward vs institutional punishment for the long-term level of cooperation and population social welfare, for different intensities of selection. Similar observations as in Figure 2 in the main text. Other parameters as in Figure 2.

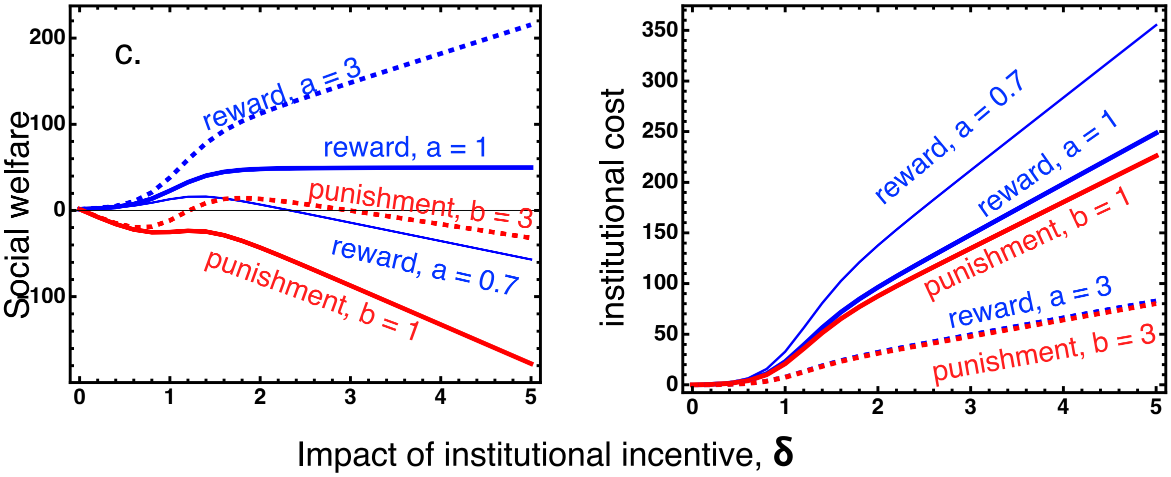

Figure S2. Maximise social welfare vs minimise the cost of institutional incentive. We observe that the these two objectives are often not aligned. For example, for highly efficient reward ($a = 3$), increasing δ leads to better social welfare. For minimising the cost while ensuring a certain desired level of cooperation (see Figure 2 in the main text), the optimal value of δ would be exactly the one that achieves the desired level of cooperation (because the institutional cost E increases with δ). While for maximising social welfare, one should aim for the highest possible δ as it is an increasing function. Parameters similar to Figure 2 in the main text.

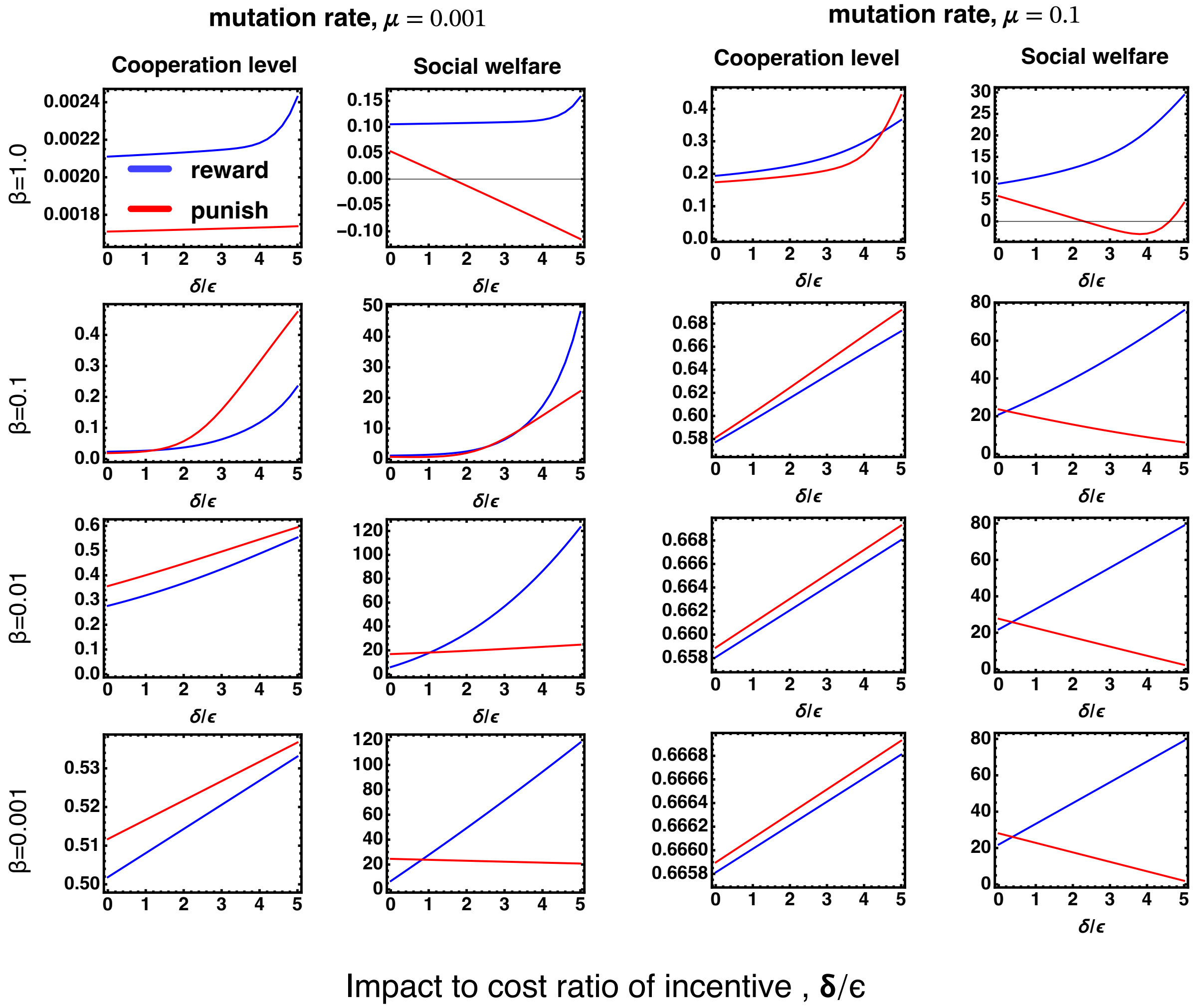

Figure S3. Impact of peer reward vs peer punishment for the long-term level of cooperation and population social welfare, for varying the efficiency of incentive δ/ϵ, for different values of the intensities of selection β and μ. Other parameters as in Figure 1 in the main text.

References

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.

Boyd, R., Gintis, H., Bowles, S., and Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6):3531–3535.

Capraro, V., Di Paolo, R., Perc, M., and Pizziol, V. (2024). Language-based game theory in the age of artificial intelligence. *Journal of the Royal Society Interface*, 21(212):20230720.

Capraro, V. and Perc, M. (2021). Mathematical foundations of moral preferences. *Journal of the Royal Society interface*, 18(175):20200880.

Chen, X., Szolnoki, A., and Perc, M. (2014). Probabilistic sharing solves the problem of costly punishment. *New Journal of Physics*, 16(8):083016.

Cimpeanu, T., Di Stefano, A., Perret, C., and Han, T. A. (2023). Social diversity reduces the complexity and cost of fostering fairness. *Chaos, Solitons & Fractals*, 167:113051.

Cimpeanu, T. and Han, T. A. (2020). Making an example: signalling threat in the evolution of cooperation. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE.

Cimpeanu, T. and Han, T. A. (2024). The digital gallows: Threat of institutional punishment fosters the emergence of cooperation. In *ALIFE 2024: Proceedings of the 2024 Artificial Life Conference*. MIT Press.

Cimpeanu, T., Perret, C., and Han, T. A. (2021). Cost-efficient interventions for promoting fairness in the ultimatum game. *Knowledge-Based Systems*, 233:107545.

Coombs, C. H. (1973). A reparameterization of the prisoner's dilemma game. *Behavioral Science*, 18(6):424–428.

Duong, M. H., Durbac, C. M., and Han, T. A. (2023). Cost optimisation of hybrid institutional incentives for promoting cooperation in finite populations. *J. Math. Biol.*, 87(77).

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.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Fernández Domingos, E., Terrucha, I., Suchon, R., Grujić, J., Burguillo, J. C., Santos, F. C., and Lenaerts, T. (2022). Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma. *Scientific reports*, 12(1):8492.

Góis, A. R., Santos, F. P., Pacheco, J. M., and Santos, F. C. (2019). Reward and punishment in climate change dilemmas. *Sci. Rep.*, 9(1):1–9.

Guo, H., Shen, C., Hu, S., Xing, J., Tao, P., Shi, Y., and Wang, Z. (2023). Facilitating cooperation in human-agent hybrid populations through autonomous agents. *Iscience*, 26(11).

Han, T. A. (2016). Emergence of social punishment and cooperation through prior commitments. In *Proceedings of the thirtieth aaai conference on artificial intelligence*, pages 2494–2500.

Han, T. A. (2022a). Emergent behaviours in multi-agent systems with evolutionary game theory. *AI Communications*, 35(4):327 – 337.

Han, T. A. (2022b). 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., 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. (2020). To Regulate or Not: A Social Dynamics Analysis of an Idealised AI Race. *Journal of Artificial Intelligence Research*, 69:881–921.

Han, T. A., Perret, C., and Powers, S. T. (2021). When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games. *Cognitive Systems Research*, 68:111–124.

Han, T. A. and Tran-Thanh, L. (2018). Cost-effective external interference for promoting the evolution of cooperation. *Scientific reports*, 8(1):1–9.

Herrmann, B., Thoni, C., and Gachter, S. (2008). Antisocial punishment across societies. *Science*, 319(5868):1362–1367.

Imhof, L. A., Fudenberg, D., and Nowak, M. A. (2005). Evolutionary cycles of cooperation and defection. *Proceedings of the National Academy of Sciences*, 102(31):10797–10800.

Kaneko, M. and Nakamura, K. (1979). The nash social welfare function. *Econometrica: Journal of the Econometric Society*, pages 423–435.

Krellner, M. et al. (2021). Pleasing enhances indirect reciprocity-based cooperation under private assessment. *Artificial Life*, 27(3–4):246–276.

Krellner, M. and Han, T. A. (2023). Words are not wind–how joint commitment and reputation solve social dilemmas, without repeated interactions or enforcement by third parties. *arXiv preprint arXiv:2307.06898*.

Lenaerts, T., Saponara, M., Pacheco, J. M., and Santos, F. C. (2024). Evolution of a theory of mind. *Iscience*, 27(2).

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

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

Nowak, M. A. (2006a). *Evolutionary dynamics: exploring the equations of life*. Harvard university press.

Nowak, M. A. (2006b). Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563.

Okada, I. (2020). A review of theoretical studies on indirect reciprocity. *Games*, 11(3):27.

Paiva, A., Santos, F., and Santos, F. (2018). Engineering pro-sociality with autonomous agents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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

Radzvilavicius, A. L., Kessinger, T. A., and Plotkin, J. B. (2021). Adherence to public institutions that foster cooperation. *Nature communications*, 12(1):3567.

Rand, D. G., Armao IV, J. J., Nakamaru, M., and Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of theoretical biology*, 265(4):624–632.

Santos, F. and Pacheco, J. M. (2011). Risk of collective failure provides an escape from the tragedy of the commons. *Proceedings of the National Academy of Sciences*, 108(26):10421–10425.

Santos, F. C., Pacheco, J. M., and Lenaerts, T. (2006). Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proceedings of the National Academy of Sciences*, 103(9):3490–3494.

Santos, F. P. (2024). Prosocial dynamics in multiagent systems. *AI Magazine*, 45(1):131–138.

Santos, F. P., Santos, F. C., and Pacheco, J. M. (2018). Social norm complexity and past reputations in the evolution of cooperation. *Nature*, 555(7695):242–245.

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., Chen, X., Brännström, Å., and Dieckmann, U. (2015). First carrot, then stick: How the adaptive hybridization of incentives promotes cooperation. *Journal of the Royal Society Interface*, 12:20140935.

Shen, C., He, Z., Shi, L., Wang, Z., and Tanimoto, J. (2024). Prosocial punishment bots breed social punishment in human players. *Journal of The Royal Society Interface*, 21(212):20240019.

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

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

Sigmund, K., Hauert, C., and Nowak, M. A. (2001). Reward and punishment. *Proceedings of the National Academy of Sciences*, 98(19):10757–10762.

Smith, J. M. (1974). The theory of games and the evolution of animal conflicts. *Journal of theoretical biology*, 47(1):209–221.

Sun, W., Liu, L., Chen, X., Szolnoki, A., and Vasconcelos, V. V. (2021). Combination of institutional incentives for cooperative governance of risky commons. *Iscience*, 24(8).

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D. C., Ezer, D., Haert, F. C. v. d., Mugisha, F., et al. (2020). Ai for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11(1):2468.

Traulsen, A., Levin, S. A., and Saad-Roy, C. M. (2023). Individual costs and societal benefits of interventions during the covid-19 pandemic. *Proceedings of the National Academy of Sciences*, 120(24):e2303546120.

Traulsen, A. and Nowak, M. A. (2006). Evolution of cooperation by multilevel selection. *Proceedings of the National Academy of Sciences*, 103(29):10952–10955.

Van Lange, P. A., Rockenbach, B., and Yamagishi, T. (2014). *Reward and punishment in social dilemmas*. Oxford University Press.

Wang, S., Chen, X., and Szolnoki, A. (2019). Exploring optimal institutional incentives for public cooperation. *Communications in Nonlinear Science and Numerical Simulation*, 79:104914.

Wang, S., Liu, L., and Chen, X. (2022). Incentive strategies for the evolution of cooperation: Analysis and optimization. *Europhysics Letters*, 136(6):68002.

Wilson, D. S. and Dugatkin, L. A. (1991). Nepotism vs tit-for-tat, or, why should you be nice to your rotten brother? *Evolutionary ecology*, 5:291–299.

Xia, C., Wang, J., Perc, M., and Wang, Z. (2023). Reputation and reciprocity. *Physics of Life Reviews*.

Xiao, J., Liu, L., Chen, X., and Szolnoki, A. (2023). Evolution of cooperation driven by sampling punishment. *Physics Letters A*, 475:128879.

Zimmaro, F., Miranda, M., Fernández, J. M. R., A. Moreno López, J., Reddel, M., Widler, V., Antonioni, A., and Han, T. A. (2024 (In press)). Emergence of cooperation in the one-shot prisoner's dilemma through discriminatory and samaritan ais. *Journal of the Royal Society Interface*.