INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Han TA, Duong MH, Perc M. 2024 Evolutionary mechanisms that promote cooperation may not promote social welfare. *J. R. Soc. Interface* **21**: 20240547.
https://doi.org/10.1098/rsif.2024.0547

Received: 10 August 2024
Accepted: 8 October 2024

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
evolution, biocomplexity, biomathematics

Keywords:
social welfare, cost efficiency, reward, punishment, evolution of cooperation, social dilemma

Author for correspondence:
The Anh Han
e-mail: t.han@tees.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.7539188.

THE ROYAL SOCIETY PUBLISHING

Evolutionary mechanisms that promote cooperation may not promote social welfare

The Anh Han[1], Manh Hong Duong[2] and 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

TAH, 0000-0002-3095-7714; MHD, 0000-0002-4361-0795; MP, 0000-0002-3087-541X

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 pay-offs, it is, however, possible that aiming for highest levels of cooperation might be detrimental for social welfare—the latter broadly defined as the total population pay-off, taking into account all costs involved for inducing increased prosocial behaviours. Herein, by comparing stochastic evolutionary models of two well-established mechanisms of prosocial behaviour—namely, peer and institutional incentives—we demonstrate that the objectives of maximizing cooperation and of maximizing social welfare are often misaligned. First, while peer punishment is often more effective than peer reward in promoting cooperation—especially with a higher impact-to-cost ratio—the opposite is true for social welfare. In fact, welfare typically decreases (increases) with this ratio for punishment (reward). Second, for institutional incentives, while maintaining similar levels of cooperation, rewards result in positive social welfare across a much broader range of parameters. Furthermore, both types of incentives often achieve optimal social welfare when their impact is moderate rather than maximal, indicating that careful planning is essential for costly institutional mechanisms to optimize social outcomes. These findings are consistent across varying mutation rates, selection intensities and game configurations. Overall, we argue for the need of adopting social welfare as the main optimization objective when designing and implementing evolutionary mechanisms for social and collective goods.

1. Introduction

Since Darwin, the challenge of explaining the evolution of cooperative behaviour has been actively explored across various fields, including evolutionary biology, ecology, economics and multi-agent systems [1–6]. 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 [3,7]. 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 pay-offs, 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 pay-off of the population [8], including all costs associated with inducing behavioural 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 pay-off of the incentive recipient [9–12]. 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 pay-offs when playing with rewarders, compared with no increase when playing with punishers, the overall population pay-offs 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 (§4).

In this article, we demonstrate that it might be more important to optimize 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 [9,13–15].

We adopt evolutionary game theory (EGT) [2,16,17], a well-established mathematical framework for modelling and analysing cooperative behaviours and their emergence and stability [1,3]. 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 [7]. Our analysis is carried out using the one-shot Prisoner's Dilemma, a well-adopted game for modelling a social dilemma of cooperation [2,18,19].

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 electronic supplementary material.

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 pay-offs given by the following pay-off 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 pay-off: R for mutual cooperation and P for mutual defection. If the agents play different strategies, the cooperator gets the sucker's pay-off S, and the defector gets the temptation to defect T. The pay-off matrix corresponds to the preferences associated with the PD when the parameters satisfy the ordering $T > R > P > S$ [18].

The strength of the dilemma in the PD game can be varied adopting a simplified scaling approach from [6,19,20]. Indeed, by fixing $T - R = P - S = 1$, the dilemma strength decreases when $R - P$ increases.

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

$$\Delta_N^m := \{\mathbf{n} = (n_1, \ldots, n_m) | 0 \le n_i \le N, \sum_{i=1}^{m} n_i = N\}, \tag{2.1}$$

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 [21,22],

$$P_{A,B} = \left(1 + e^{\beta(f_A - f_B)}\right)^{-1}, \tag{2.2}$$

where f_A (f_B) denotes the fitness of individual A (B). The parameter β represents the intensity of selection, i.e. how strongly the individuals base their decision to imitate on fitness comparison. For $\beta = 0$, we obtain the limit of neutral drift—the imitation decision is random. For large β, imitation becomes increasingly deterministic. Overall, this approach leads to a unified framework for evolutionary dynamics at all intensities of selection, from random drift to imitation dynamics [16,22].

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

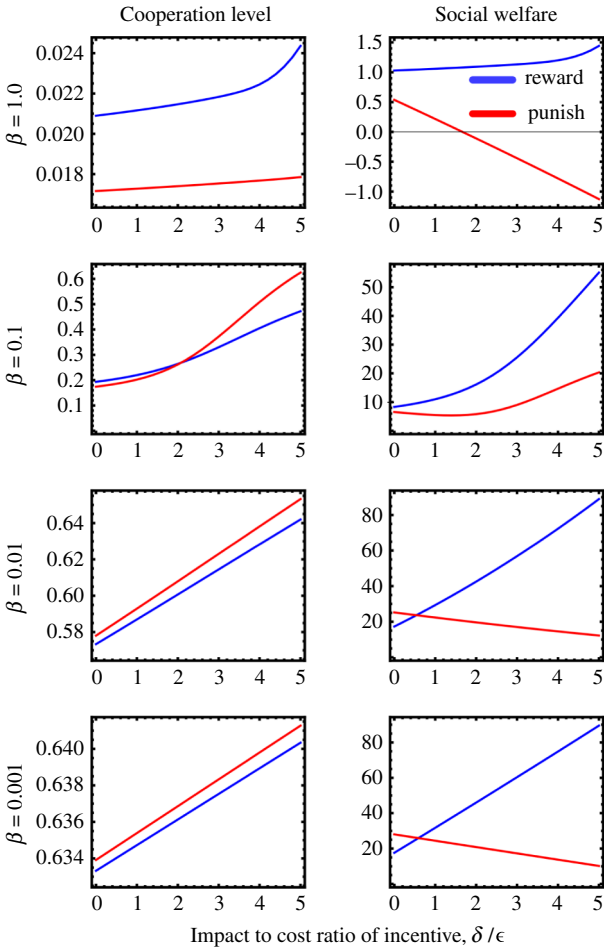

Figure 1. Impact of peer reward versus peer punishment for the long-term level of cooperation (f_C, see equation (2.5)) and population social welfare (SW, see equation (2.6)), for varying the efficiency of incentive δ/ϵ, for different values of the intensities of selection β. We observe that punishment 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 pay-off matrix $R = 1, S = -1, T = 2, P = 0$.

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

$$S = |\Delta_N^m| = \binom{N+m-1}{m-1}. \tag{2.3}$$

For example, in the peer incentive models we will analyse below (§3.1), there are three strategies, i.e. $m = 3$. Thus, in that case, the Markov chain has $S = (N+2)(N+1)/2$ states. In the institutional incentive models (§3.2), there are two strategies ($m = 2$) and the Markov chain has $S = N+1$ states.

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

$$\omega_{\mathbf{n},\mathbf{n}'} = \begin{cases} \frac{n_i}{N}\left(\frac{\mu}{m-1} + (1-\mu)\frac{n_j}{N}P_{i,j}\right) & \text{if } n_i' = n_i - 1,\ n_j' = n_j + 1, n_l' = n_l \text{ for } l \notin \{i,j\}, \\ 1 - \sum_{j \neq i} \frac{n_i}{N}\left(\frac{\mu}{m-1} + (1-\mu)\frac{n_j}{N}P_{i,j}\right) & \text{if } \mathbf{n} = \mathbf{n}', \\ 0 \ \text{ otherwise.} \end{cases} \tag{2.4}$$

The first equation represents the probability that a player who chooses strategy i will adopt strategy j. It comprises the probability when it is due to mutation, i.e. $\mu\,(n_i/N)\,(1/(m-1))$ (the second and the third fractions represent the probability of choosing a strategist i from the population and that of selecting strategy j from the set of $m-1$ strategies other than i), and the probability when it is due to imitation, i.e. $(1-\mu)(n_i/N)\,(n_j/N)P_{i,j}$ (the second and the third fractions represent the probabilities of choosing a strategist i and j from the population, respectively).

Finally, by computing the normalized 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 $\bar{p}_{\mathbf{n}}$

$$f_i = \sum_{\mathbf{n}} \frac{\mathbf{n}_i \bar{p}_{\mathbf{n}}}{N}, \tag{2.5}$$

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 pay-off (social welfare), SW, is given as follows

$$\mathrm{SW} = \sum_{\mathbf{n}} \frac{SW(\mathbf{n})\bar{p}_{\mathbf{n}}}{N}, \tag{2.6}$$

where $SW(\mathbf{n})$ is the population total pay-off 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 [15,23,24]. Now, when optimizing 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. maximizing 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

$$E(\Theta) = \sum_{\mathbf{n}} \theta_{\mathbf{n}} \bar{p}_{\mathbf{n}}. \tag{2.7}$$

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

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

Note that the population pay-off depends on incentive policy Θ. Namely, it alters the transition probabilities given in equation (2.4) (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 pay-off.

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 pay-off matrices for peer punishment and peer reward cases are given as follows, respectively

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

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. Note that when β is very low, representing weak selection scenarios (e.g. the last row of figure 1, with $\beta = 0.001$), the transitions between states of the Markov chain are increasingly close to random. That results in all strategies in the population having a similar frequency in stationary distribution. Thus, even when the impact-to-cost ratio is 0, the total level of cooperation (i.e. sum of frequencies of C and SP or SR) is close to 2/3. On the other hand, when β is sufficiently high, representing strong selection scenarios (see the first row of figure 1, with $\beta = 1.0$), the level of cooperation is very low even for the most efficient punishment/reward scenarios that are considered in our analysis (i.e. up to five). The reason is that strong selection intensifies the transitions that replace a C-player with a D-player among the states in the Markov chain.

Note, however, that if we further increase the incentive efficiency, high levels of cooperation can be achieved (see electronic supplementary material, figure S3).

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 the 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 electronic supplementary material, figure S1). In fact, our observations indicate that a higher mutation rate further intensifies the inefficiency of peer punishment concerning social welfare. Moreover, we examine the robustness of the observations for varying the dilemma strength of the PD game, adopting a simplified scaling approach from [6,19,20] (see electronic supplementary material, figure S2). Again, similar observations are achieved, with a slightly improved performance of peer punishment for a weaker dilemma strength, but only when the intensity of selection is moderate (see $\beta = 0.1$).

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 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 [12,25,26] or when incentives are provided in a conditional or stochastic approach [27–29].

3.2. Institutional incentives

We now analyse the alignment between promoting cooperation and social welfare in minimal models of institutional reward and punishment [14,15,30]. 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 δ/a (δ/b, respectively) so that the cooperator's (defector's) pay-off increases (decreases) by δ, where $a, b > 0$ are constants representing the efficiency ratios of providing this type of incentive.

Figure 2 shows that, as expected [15,30,31], both types of incentives lead to the same level of cooperation, assuming they 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 is robust for different intensities of selection, see electronic supplementary material, figure S4.

For cost-efficient institutional reward (i.e. $a > 1$), it is best for social welfare to maximize the impact (i.e. strong reward with a high per capita 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 [14,30,32,33] 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.

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 $a = 0.7$, 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 optimizing cooperation and social welfare might not be always aligned.

The above-mentioned observations are also valid when considering varying dilemma strengths of the Prisoner's Dilemma (see electronic supplementary material, figure S5). Notably, we also find that for weaker dilemmas (i.e. for larger $R - P$), both institutional reward and punishment result in higher levels of social welfare, while as before, reward performs better than punishment for similar effect-to-cost ratios (i.e. $a = b$). This further reinforces the finding that the optimal institutional incentives for cooperation and social welfare are often distinct, and that increasing the costs and impacts of incentives may be counterproductive for social welfare, regardless of whether they are rewards or punishments.

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 [2,3,7]. The emphasis is often placed on the extent of cooperation that a given

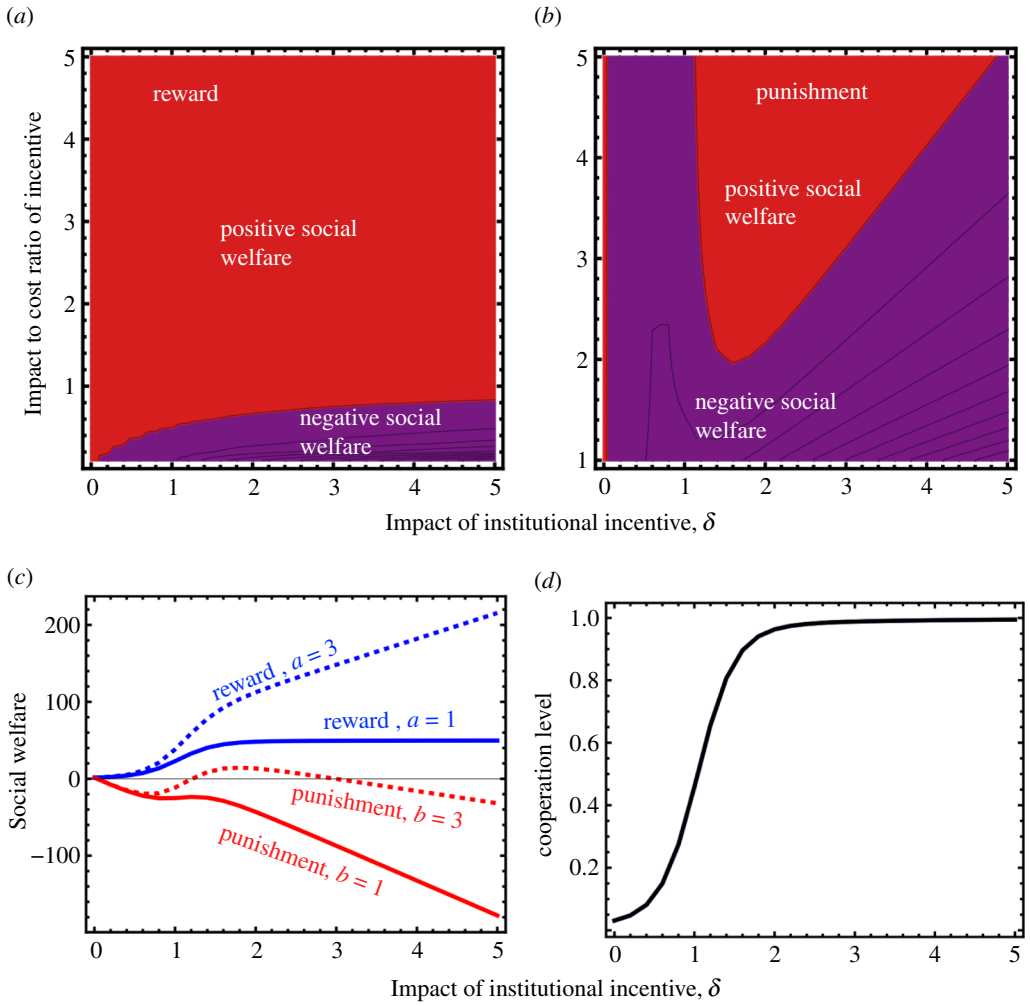

Figure 2. Impact of institutional reward versus institutional punishment for the long-term level of cooperation and population social welfare. We observe that although both types of incentives lead to the same level of cooperation given the same incentive impact on the incentive recipient δ (assuming their impact-to-cost ratios are the same, i.e. $a = b$, see panel *d*), reward leads to positive social welfare (panel *c*) 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 $R = 1, S = -1, T = 2, P = 0$.

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 pay-offs, it is possible that aiming for highest levels of cooperation might be detrimental for social welfare. In this article, using numerical simulations for stochastic evolutionary models for two important incentive mechanisms, peer and institutional incentives, we have demonstrated exactly that. For peer incentives, our finding showed that while peer punishment is often more effective than peer reward in promoting cooperation, in particular for a higher impact-to-cost ratio of the two incentives, the opposite is true for social welfare. In fact, the welfare typically decreases with this ratio for peer punishment, while it increases in the case of peer reward. This is an important finding given that previous works usually focus more on peer punishment than peer reward, given the former advantage for enhancing cooperation [9,12,25,32,34,35]. For institutional incentives, we showed that reward fosters positive social welfare over a much wider range of parameters while maintaining similar levels of cooperation. Furthermore, both types of institutional incentives often achieve optimal social welfare when their impact is moderate rather than at the maximal level. This observation is not possible in previous models of institutional incentives as they consider only the objective of maximizing cooperation [2,30,33]. These results indicate that careful planning is essential for costly institutional mechanisms to optimize social goods.

Closely related to the present work is a recent body of literature on cost optimization of institutional incentives for promoting cooperation and fairness [15,23,24,34,36–41]. These works usually consider a bi-objective optimization 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 into the optimal incentive policy, they still do not guarantee optimal social welfare. In fact, the two objectives are often misaligned (see electronic supplementary material, figure S6). Moreover, we argue that focusing on the social welfare provides a more convenient, single-objective optimization problem.

We analysed here two incentive mechanisms. It might be interesting to study whether maximizing social welfare is aligned with maximizing 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) [42]. For indirect reciprocity to work [43], one might need to enhance transparency of reputation, e.g. via implementing institution broadcasting reputation scores [44] or other costly communication mechanisms [45,46]. Thus one might need to balance between promoting cooperation and the cost of enabling it. For direct reciprocity to work [47], one might need to reduce

noise, increase cognitive capacity [48,49] and improve trust [50], which are costly and thus need to be taken into account for balancing the population social welfare. For pre-commitment to work [51], it usually requires a third party such as an institution that provides incentives [31] or broadcasts commitment-based reputation [52] for ensuring commitment compliance, which are costly and need 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 [39,53–59]. 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 optimization 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. Using 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 prioritize STEM education, arts and humanities, vocational training or critical thinking skills. Optimizing 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 prioritizing 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.

We focused here on the maximization of the population social welfare as a whole. In many domains, it can be useful to look deeper into the distribution of the social welfare across the population, as for the same total welfare, a fairer distribution of it might be more desirable. (It is noteworthy that this is not a problem in our model as, in a well-mixed population, all cooperators would receive the same pay-off.) For example, in the resource allocation scenarios such as those modelled by the Ultimatum and Dictator games [39,60–62], a fair distribution between resource providers and recipients are preferred, even for a lower total welfare. In addition, a critical ongoing issue in advanced artificial intelligence (AI) developments is to avoid power concentration among fewer big techs [63], in order to ensure equitable redistribution of the vast profits from advanced AI. In this domain, it would be interesting to explore mechanisms such as windfall clause [64], open-source developments [65], regulatory markets [66,67] and voluntary safety commitments [68] as potential approaches for ensuring fair and optimal social welfare. In all these examples, one might need to balance between efficiency (i.e. the total welfare) and equity (fair distribution of the welfare). This issue has been studied extensively in operation research literature (e.g. a survey in [69]), providing suitable objective functions that combine both factors.

An important ongoing initiative in AI research is the design and implementation of AI for social goods [70], 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 [4,5,71–76]. 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, optimizing the overall system pay-off or social welfare is crucial for the effective use of AI for social goods.

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. The code for replicating the results is available at [77].

Supplementary material is available online [78].

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

Authors' contributions. T.A.H.: conceptualization, formal analysis, funding acquisition, investigation, methodology, software, validation, visualization, writing—original draft, writing—review and editing; M.H.D.: conceptualization, formal analysis, investigation, methodology, validation, writing—original draft, writing—review and editing; M.P.: investigation, validation, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. 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).

References

1. Nowak MA. 2006 Five rules for the evolution of cooperation. *Science* **314**, 1560–1563. (doi:10.1126/science.1133755)
2. Sigmund K. 2010 *The calculus of selfishness*. Princeton, NJ: Princeton University Press.
3. Perc M, Jordan JJ, Rand DG, Wang Z, Boccaletti S, Szolnoki A. 2017 Statistical physics of human cooperation. *Phys. Rep.* **687**, 1–51. (doi:10.1016/j.physrep.2017.05.004)
4. Han TA. 2022 Emergent behaviours in multi-agent systems with evolutionary game theory. *AI Commun.* **35**, 327–337. (doi:10.3233/AIC-220104)
5. Paiva A, Santos F, Santos F. 2018 Engineering pro-sociality with autonomous agents. In *Proc. of the AAAI Conf. on Artificial Intelligence*. vol. **32**. (doi:10.1609/aaai.v32i1.12215)
6. Arefin MR, Kabir KA, Jusup M, Ito H, Tanimoto J. 2020 Social efficiency deficit deciphers social dilemmas. *Sci. Rep.* **10**, 16092. (doi:10.1038/s41598-020-72971-y)
7. Novak MA. 2006 *Evolutionary dynamics: exploring the equations of life*. Cambridge, MA: Harvard University Press.
8. Kaneko M, Nakamura K. 1979 The Nash social welfare function. *Econometrica* **47**, 423–435. (doi:10.2307/1914191)
9. Sigmund K, Hauert C, Nowak MA. 2001 Reward and punishment. *Proc. Natl Acad. Sci. USA* **98**, 10 757–10 762. (doi:10.1073/pnas.161155698)

10. Fehr E, Gächter S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)
11. Boyd R, Gintis H, Bowles S, Richerson PJ. 2003 The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535. (doi:10.1073/pnas.0630443100)
12. Han TA. 2016 Emergence of social punishment and cooperation through prior commitments. In *Proc. of the 30th AAAI Conf. on Artificial Intelligence*, Phoenix, AZ, pp. 2494–2500.
13. VanPA, Rockenbach B, Yamagishi T. 2014 *Reward and punishment in social dilemmas*. Oxford, UK: Oxford University Press.
14. Chen X, Sasaki T, Brännström Å, Dieckmann U. 2015 First carrot, then stick: how the adaptive hybridization of incentives promotes cooperation. *J. R. Soc. Interface* **12**, 20140935. (doi:10.1098/rsif.2014.0935)
15. Duong MH, Han TA. 2021 Cost efficiency of institutional incentives for promoting cooperation in finite populations. *Proc. R. Soc. A* **477**, 20210568. (doi:10.1098/rspa.2021.0568)
16. Imhof LA, Fudenberg D, Nowak MA. 2005 Evolutionary cycles of cooperation and defection. *Proc. Natl Acad. Sci. USA* **102**, 10 797–10 800. (doi:10.1073/pnas.0502589102)
17. Smith JM. 1974 The theory of games and the evolution of animal conflicts. *J. Theor. Biol*. **47**, 209–221. (doi:10.1016/0022-5193(74)90110-6)
18. Coombs CH. 1973 A reparameterization of the Prisoner's Dilemma game. *Behav. Sci.* **18**, 424–428. (doi:10.1002/bs.3830180605)
19. Wang Z, Kokubo S, Jusup M, Tanimoto J. 2015 Universal scaling for the dilemma strength in evolutionary games. *Phys. Life Rev*. **14**, 1–30. (doi:10.1016/j.plrev.2015.04.033)
20. Ito H, Tanimoto J. 2018 Scaling the phase-planes of social dilemma strengths shows game-class changes in the five rules governing the evolution of cooperation. *R. Soc. Open Sci*. **5**, 181085. (doi:10.1098/rsos.181085)
21. Szabó G, Tőke C. 1998 Evolutionary Prisoner's Dilemma game on a square lattice. *Phys. Rev. E* **58**, 69–73. (doi:10.1103/PhysRevE.58.69)
22. Traulsen A, Nowak MA. 2006 Evolution of cooperation by multilevel selection. *Proc. Natl Acad. Sci. USA* **103**, 10 952–10 955. (doi:10.1073/pnas.0602530103)
23. Han TA, Tran-Thanh L. 2018 Cost-effective external interference for promoting the evolution of cooperation. *Sci. Rep.* **8**, 1–9. (doi:10.1038/s41598-018-34435-2)
24. Wang S, Chen X, Szolnoki A. 2019 Exploring optimal institutional incentives for public cooperation. *Commun. Nonlinear Sci. Numer. Simul*. **79**, 104914. (doi:10.1016/j.cnsns.2019.104914)
25. Herrmann B, Thoni C, Gachter S. 2008 Antisocial punishment across societies. *Science* **319**, 1362–1367. (doi:10.1126/science.1153808)
26. Rand DG, Armao IV JJ, Nakamaru M, Ohtsuki H. 2010 Anti-social punishment can prevent the co-evolution of punishment and cooperation. *J. Theor. Biol*. **265**, 624–632. (doi:10.1016/j.jtbi.2010.06.010)
27. Xiao J, Liu L, Chen X, Szolnoki A. 2023 Evolution of cooperation driven by sampling punishment. *Phys. Lett. A* **475**, 128879. (doi:10.1016/j.physleta.2023.128879)
28. Chen X, Szolnoki A, Perc M. 2014 Probabilistic sharing solves the problem of costly punishment. *New J. Phys*. **16**, 083016. (doi:10.1088/1367-2630/16/8/083016)
29. Cimpeanu T, Han TA. 2020 Making an example: signalling threat in the evolution of cooperation. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, Glasgow, UK, pp. 1–8. IEEE. (doi:10.1109/CEC48606.2020.9185749)
30. Góis AR, Santos FP, Pacheco JM, Santos FC. 2019 Reward and punishment in climate change dilemmas. *Sci. Rep.* **9**, 1–9. (doi:10.1038/s41598-019-52524-8)
31. Han TA. 2022 Institutional incentives for the evolution of committed cooperation: ensuring participation is as important as enhancing compliance. *J. R. Soc. Interface* **19**, 20220036. (doi:10.1098/rsif.2022.0036)
32. Sigmund K, De Silva H, Traulsen A, Hauert C. 2010 Social learning promotes institutions for governing the commons. *Nature* **466**, 861–863. (doi:10.1038/nature09203)
33. Sasaki T, Brännström Å, Dieckmann U, Sigmund K. 2012 The take-it-or-leave-it option allows small penalties to overcome social dilemmas. *Proc. Natl Acad. Sci. USA* **109**, 1165–1169. (doi:10.1073/pnas.1115219109)
34. Cimpeanu T, Han TA. 2024 The digital gallows: threat of institutional punishment fosters the emergence of cooperation. In *ALIFE 2024: Proc. of the 2024 Artificial Life Conf*. MIT Press. (doi:10.1162/isal_a_00822)
35. Hauert C, Traulsen A, Brandt H, Nowak MA, Sigmund K. 2007 Via freedom to coercion: the emergence of costly punishment. *Science* **316**, 1905–1907. (doi:10.1126/science.1141588)
36. Duong MH, Durbac CM, Han TA. 2023 Cost optimisation of hybrid institutional incentives for promoting cooperation in finite populations. *J. Math. Biol*. **87**, 77. (doi:10.1007/s00285-023-02011-6)
37. Wang S, Liu L, Chen X. 2021 Incentive strategies for the evolution of cooperation: analysis and optimization. *Europhys. Lett*. **136**, 68002. (doi:10.1209/0295-5075/ac3c8a)
38. Cimpeanu T, Di Stefano A, Perret C, Han TA. 2023 Social diversity reduces the complexity and cost of fostering fairness. *Chaos Solitons Fractals* **167**, 113051. (doi:10.1016/j.chaos.2022.113051)
39. Cimpeanu T, Perret C, Han TA. 2021 Cost-efficient interventions for promoting fairness in the ultimatum game. *Knowl. Based Syst*. **233**, 107545. (doi:10.1016/j.knosys.2021.107545)
40. Sun W, Liu L, Chen X, Szolnoki A, Vasconcelos VV. 2021 Combination of institutional incentives for cooperative governance of risky commons. *iScience* **24**, 102844. (doi:10.1016/j.isci.2021.102844)
41. Sun Z, Chen X, Szolnoki A. 2023 State-dependent optimal incentive allocation protocols for cooperation in public goods games on regular networks. *IEEE Trans. Netw. Sci. Eng*. **10**, 3975–3988. (doi:10.1109/TNSE.2023.3279094)
42. Wilson DS, Dugatkin LA. 1991 Nepotism vs tit-for-tat, or, why should you be nice to your rotten brother? *Evol. Ecol*. **5**, 291–299. (doi:10.1007/BF02214233)
43. Okada I. 2020 A review of theoretical studies on indirect reciprocity. *Games* **11**, 27. (doi:10.3390/g11030027)
44. Radzvilavicius AL, Kessinger TA, Plotkin JB. 2021 Adherence to public institutions that foster cooperation. *Nat. Commun*. **12**, 3567. (doi:10.1038/s41467-021-23783-9)
45. Krellner M, Han TA. 2022 Pleasing enhances indirect reciprocity-based cooperation under private assessment. *Artif. Life* **27**, 246–276. (doi:10.1162/artl_a_00344)
46. Santos FP, Santos FC, Pacheco JM. 2018 Social norm complexity and past reputations in the evolution of cooperation. *Nature* **555**, 242–245. (doi:10.1038/nature25763)
47. Xia C, Wang J, Perc M, Wang Z. 2023 Reputation and reciprocity. *Phys. Life Rev.* **46**, 8–45. (doi:10.1016/j.plrev.2023.05.002)
48. Lenaerts T, Saponara M, Pacheco JM, Santos FC. 2024 Evolution of a theory of mind. *iScience* **27**, 108862. (doi:10.1016/j.isci.2024.108862)
49. Han TA, Pereira LM, Santos FC. 2012 Corpus-based intention recognition in cooperation dilemmas. *Artif. Life* **18**, 365–383. (doi:10.1162/ARTL_a_00072)
50. Han TA, Perret C, Powers ST. 2021 When to (or not to) trust intelligent machines: insights from an evolutionary game theory analysis of trust in repeated games. *Cogn. Syst. Res*. **68**, 111–124. (doi:10.1016/j.cogsys.2021.02.003)
51. Nesse RM. 2001 *Evolution and the capacity for commitment.* Foundation Series on Trust. New York, NY: Russell Sage.
52. Krellner M, Han TA. 2023 Words are not wind – how joint commitment and reputation solve social dilemmas, without repeated interactions or enforcement by third parties. *arXiv*. (doi:10.48550/arXiv.2307.06898)
53. Santos FC, Pacheco JM. 2011 Risk of collective failure provides an escape from the tragedy of the commons. *Proc. Natl Acad. Sci. USA* **108**, 10 421–10 425. (doi:10.1073/pnas.1015648108)
54. Capraro V, Perc M. 2021 Mathematical foundations of moral preferences. *J. R. Soc. Interface* **18**, 20200880. (doi:10.1098/rsif.2020.0880)
55. Santos FC, Pacheco JM, Lenaerts T. 2006 Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proc. Natl Acad. Sci. USA* **103**, 3490–3494. (doi:10.1073/pnas.0508201103)

56. Han TA, Pereira LM, Lenaerts T, Santos FC. 2021 Mediating artificial intelligence developments through negative and positive incentives. *PLoS One* **16**, e0244592. (doi:10.1371/journal.pone.0244592)
57. Andras P *et al*. 2018 Trusting intelligent machines: deepening trust within socio-technical systems. *IEEE Technol. Soc. Mag*. **37**, 76–83. (doi:10.1109/MTS.2018.2876107)
58. Han TA, Moniz Pereira L, Santos FC, Lenaerts T. 2020 To regulate or not: a social dynamics analysis of an idealised AI race. *J. Artif. Intell. Res.* **69**, 881–921. (doi:10.1613/jair.1.12225)
59. Traulsen A, Levin SA, Saad-Roy CM. 2023 Individual costs and societal benefits of interventions during the COVID-19 pandemic. *Proc. Natl Acad. Sci. USA* **120**, e2303546120. (doi:10.1073/pnas.2303546120)
60. Hoffman E, McCabe K, Shachat K, Smith V. 1994 Preferences, property rights, and anonymity in bargaining games. *Games Econ. Behav*. **7**, 346–380. (doi:10.1006/game.1994.1056)
61. Zisis I, Di Guida S, Han TA, Kirchsteiger G, Lenaerts T. 2015 Generosity motivated by acceptance-evolutionary analysis of an anticipation game. *Sci. Rep.* **5**, 1–11. (doi:10.1038/srep18076)
62. Rand DG, Tarnita CE, Ohtsuki H, Nowak MA. 2013 Evolution of fairness in the one-shot anonymous ultimatum game. *Proc. Natl Acad. Sci. USA* **110**, 2581–2586. (doi:10.1073/pnas.1214167110)
63. Verdegem P. 2022 Dismantling AI capitalism: the commons as an alternative to the power concentration of big tech. *AI Soc*. **39**, 727–737. (doi:10.1007/s00146-022-01437-8)
64. O'Keefe C, Cihon P, Garfinkel B, Flynn C, Leung J, Dafoe A. 2020 The windfall clause: distributing the benefits of AI for the common good. In *Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society*, New York, NY, pp. 327–331. New York, NY: Association for Computing Machinery. (doi:10.1145/3375627.3375842)
65. Widder DG, West S, Whittaker M. 2023 Open (for business): big tech, concentrated power, and the political economy of open ai. In *Concentrated power, and the political economy of open ai*. (doi:10.2139/ssrn.4543807)
66. Clark J, Hadfield GK. 2019 Regulatory markets for AI safety. *arXiv*. (doi:10.48550/arXiv.2001.00078)
67. Bova P, Di Stefano A, Han TA. 2024 Both eyes open: vigilant Incentives help auditors improve AI safety. *J. Phys. Complex* **5**, 025009. (doi:10.1088/2632-072X/ad424c)
68. Han TA, Lenaerts T, Santos FC, Pereira LM. 2022 Voluntary safety commitments provide an escape from over-regulation in AI development. *Technol. Soc*. **68**, 101843. (doi:10.1016/j.techsoc.2021.101843)
69. Karsu Ö, Morton A. 2015 Inequity averse optimization in operational research. *Eur. J. Oper. Res*. **245**, 343–359. (doi:10.1016/j.ejor.2015.02.035)
70. Tomašev N *et al*. 2020 AI for social good: unlocking the opportunity for positive impact. *Nat. Commun*. **11**, 2468. (doi:10.1038/s41467-020-15871-z)
71. Santos FP. 2024 Prosocial dynamics in multiagent systems. *AI Mag*. **45**, 131–138. (doi:10.1002/aaai.12143)
72. Guo H, Shen C, Hu S, Xing J, Tao P, Shi Y, Wang Z. 2023 Facilitating cooperation in human-agent hybrid populations through autonomous agents. *iScience* **26**, 108179. (doi:10.1016/j.isci.2023.108179)
73. Fernández Domingos E, Terrucha I, Suchon R, Grujić J, Burguillo JC, Santos FC, Lenaerts T. 2022 Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma. *Sci. Rep*. **12**, 8492. (doi:10.1038/s41598-022-11518-9)
74. Zimmaro F, Miranda M, Fernández JMR, Moreno López JA, Reddel M, Widler V, Antonioni A, Han TA. 2024 Emergence of cooperation in the one-shot Prisoner's Dilemma through Discriminatory and Samaritan AIs. *J. R. Soc. Interface* **21**, 20240212. (doi:10.1098/rsif.2024.0212)
75. Capraro V, Di Paolo R, Perc M, Pizziol V. 2024 Language-based game theory in the age of artificial intelligence. *J. R. Soc. Interface* **21**, 20230720. (doi:10.1098/rsif.2023.0720)
76. Shen C, He Z, Shi L, Wang Z, Tanimoto J. 2024 Prosocial punishment bots breed social punishment in human players. *J. R. Soc. Interface* **21**, 20240019. (doi:10.1098/rsif.2024.0019)
77. Han TA. 2024 Mathematica code for: Evolutionary mechanisms that promote cooperation may not promote social welfare. *Zenodo*. (doi:10.5281/zenodo.13897564)
78. Han,TA, Duong H, Perc M. 2024 Supplementary material from: Evolutionary mechanisms that promote cooperation may not promote social welfare. Figshare. (doi:10.6084/m9.figshare.c.7539188)