Resolving Social Dilemmas with Minimal Reward Transfer - Extended Abstract

JAAMAS Track

Richard Willis
King's College London
London, United Kingdom
richard.willis@kcl.ac.uk

Yali Du
King's College London
London, United Kingdom
yali.du@kcl.ac.uk

Joel Z Leibo
Google DeepMind
London, United Kingdom
jzl@deepmind.com

Michael Luck
University of Sussex
Brighton, United Kingdom
michael.luck@sussex.ac.uk

ABSTRACT

In this paper we introduce a novel metric, the *general self-interest level*, to quantify the disparity between individual and group rationality in social dilemma games. This metric represents the maximum proportion of their individual rewards that agents can retain while guaranteeing that a social welfare optimum is achieved.

This work provides both a tool for describing social dilemmas and a prescriptive solution for resolving them via reward transfer contracts. In contrast to existing metrics, the general self-interest level can enable more efficient solutions to be found. Applications include mechanism design, where we can assess the impact on collective behaviour of modifications to models of environments.

KEYWORDS

Game Theory; Social Dilemma; Reward Transfer; Cooperation

ACM Reference Format:
Richard Willis, Yali Du, Joel Z Leibo, and Michael Luck. 2025. Resolving Social Dilemmas with Minimal Reward Transfer - Extended Abstract: JAAMAS Track. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Social dilemmas represent a tension between collective and individual rationality, and are characterised by agents engaging in selfish behaviours outperforming those engaging in collective behaviours within a group, while prosocial groups outperform selfish groups.

In our paper [5], we introduce a metric to measure the agents' willingness to cooperate by quantifying the disparity between individual and group incentives. In contrast to prior approaches [1–4], we relax the requirement that all agents receive the same additional incentives, and allow different agents to receive different incentives. This can allow us to find more efficient solutions by exploiting the game structure, reducing the barriers to cooperation. Furthermore, we use reward transfer as our mechanism, which allows us to resolve social dilemmas via a redistribution of the extrinsic game rewards, without appealing to notions of intrinsic motivations such as altruism.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 NORMAL-FORM SOCIAL DILEMMAS

Social dilemmas are situations in which individuals face the choice between acting selfishly (to defect) for personal gain or acting in a prosocial manner (to cooperate) which yields greater overall benefits to the collective. A social dilemma is characterised by, for all agents: (i) the collective does better when an agent chooses to cooperate than when the agent chooses to defect; (ii) each agent may be better off individually when it defects; and, (iii) all agents prefer mutual cooperation over mutual defection.

Consider a normal-form game $(N, A, \vec{R})$, where each agent faces a choice to either cooperate, C, or defect, D:

- N is a finite set of n agents, indexed by i.
- $A = \{C, D\} \times ... \times \{C, D\}$
- $\vec{R} = (R_1, ..., R_n)$ where $R_i : A \to \mathbb{R}$ is a real-valued reward function for agent i.

We use the sum of rewards obtained by all agents as our notion of group good. A normal-form game is a social dilemma if, for any action profile $\vec{a} \in A$:

$$\text{(i)} \quad \forall i \quad \sum_j R_j(C^\frown \overrightarrow{a_{-i}}) > \sum_j R_j(D^\frown \overrightarrow{a_{-i}})$$
$$\text{(ii)} \quad \forall i \quad \exists \overrightarrow{a_{-i}} : \; R_i(D^\frown \overrightarrow{a_{-i}}) > R_i(C^\frown \overrightarrow{a_{-i}})$$
$$\text{(iii)} \quad \forall i \quad R_i((C, C, \cdots, C)) > R_i((D, D, \cdots, D))$$

Where $\overrightarrow{a_{-i}}$ represents the tuple of actions of all players other than player i, and $\frown$ is a coupling operator that inserts a_i into $\overrightarrow{a_{-i}}$ such that $\vec{a} = a_i^\frown \overrightarrow{a_{-i}}$. Prisoner's Dilemma (Table 1a) is a classical example of a social dilemmas. We say that a social dilemma is *resolved* if an action profile that maximises social welfare is dominant.

Table 1: Prisoner's Dilemma

(a) Before transfers

	C	D
C	3, 3	0, 4
D	4, 0	1, 1

(b) After transfers

	C	D
C	3, 3	1, 3
D	3, 1	1, 1

3 GENERAL SELF-INTEREST LEVEL

We introduce a mechanism by which agents can commit to transferring proportions of their rewards to one another. By engaging in reward transfer, an agent provides incentives for the recipients to help it prosper which, paradoxically, can lead to a net profit for the transferring agent if it causes a beneficial behavioural change in the recipients. Before playing a game, the agents can commit to a transfer scheme specified by a transfer matrix, $\mathbf{T}$. The tuple of post-transfer rewards, $\overrightarrow{r'}$, is given by the game rewards, $\vec{r}$, multiplied by the transfer matrix:

$$\overrightarrow{r'} = \mathbf{T}\,\vec{r} = \mathbf{T} = \begin{vmatrix} t_{1,1} & \cdots & t_{1,n} \\ \vdots & \ddots & \vdots \\ t_{n,1} & \cdots & t_{n,n} \end{vmatrix} \begin{pmatrix} r_1 \\ \vdots \\ r_2 \end{pmatrix}$$

Where $\forall i\ \forall j\ t_{ij} \in [0, 1]$, ensuring an agent can transfer at most all its reward to another agent, and cannot transfer a negative proportion. Furthermore, we require that the rows sum one, $\forall i\ \sum_j t_{ij} = 1$, so that the total game reward is conserved. We refer to the diagonal values of the transfer matrix, t_{ii}, as the self-interest of the agents, because these coefficients represent the proportion of their own game rewards that they retain.

Prisoner's Dilemma using a transfer matrix $\mathbf{T} = \begin{vmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{vmatrix}$ is displayed in Table 1b. Here, both agents are ambivalent between cooperating or defecting. This occurs because each agent stands to gain only 3/4 as much from a possible defect action, and cooperating increases the reward of its opponent, of which the agent is entitled to a proportion 1/4. For any smaller self-interest, the social dilemma is resolved, as cooperation is dominant for both players.

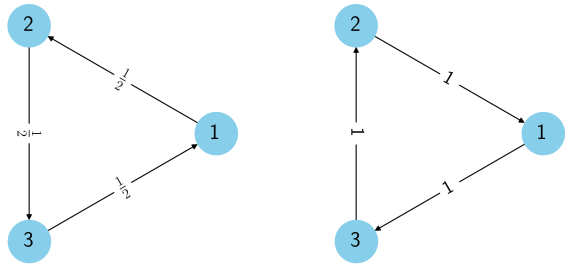
We can always find a transfer matrix that resolves a social dilemma, because a matrix with all elements equal to $1/n$ makes the post-transfer reward for all agents equal to $\frac{1}{n}\sum_i r_i$. Consequently, due to inequality (i), cooperation becomes dominant for all agents. Out of all possible transfer matrices that make mutual cooperation dominant, we find the matrices with the largest minimum of diagonal elements. These are the matrices with the greatest amount of self-interest that the agent(s) with the least self-interest retain, and we call the value of its minimum diagonal element the *general self-interest level* of the game, denoted by g^*. Formally, writing $diag(\mathbf{T})$ as a function returning the tuple of diagonal values of $\mathbf{T}$, and the reward function returning the post-transfer reward to agent i as $R'_i(\vec{r}, \mathbf{T})$, we have:

$$g^* = \max\{\min(diag(\mathbf{T})) \mid \forall i\ R'_i(C^\frown \overrightarrow{a_{-i}}, \mathbf{T}) > R'_i(D^\frown \overrightarrow{a_{-i}}, \mathbf{T})\}$$

We refer to a transfer matrix that achieves the general self-interest level as a *minimal transfer matrix*, and we denote it as $\mathbf{T}^*$.

4 RESULTS

For illustration, we introduce two multi-player variants of Prisoner's Dilemma, created using a weighted, directed graph with n nodes, each representing an agent. The agents play Prisoner's Dilemma with every co-player they share an edge with, receiving the weighted rewards where they have an inbound edge. Each agent simultaneously selects a single action that they play in all their games.



(a) Symmetrical-3PD (b) Cyclical-3PD

Figure 1: Representations of the graphical dilemmas

In both variants, each agent plays against both its co-players. In Symmetrical-3PD (Figure 1a) an agent receives the reward from both games, halved, whereas in Cymmetrical-3PD (Figure 1b), an agent only receives the reward from its game with the agent with index $i+1 \mod n$. Their minimal transfer matrices are, respectively:

$$\mathbf{T}^* = \begin{vmatrix} 3/5 & 1/5 & 1/5 \\ 1/5 & 3/5 & 1/5 \\ 1/5 & 1/5 & 3/5 \end{vmatrix} \text{ and } \mathbf{T}^* = \begin{vmatrix} 3/4 & 1/4 & 0 \\ 0 & 3/4 & 1/4 \\ 1/4 & 0 & 3/4 \end{vmatrix}$$

In Cyclical-3PD, the minimal transfer matrix permits the agents to retain a larger proportion ($g^* = 3/4$) of their own rewards compared to Symmetrical-3PD ($g^* = 3/5$). This is because the rewards for each agent depend only on its own action and the action of one other agent. Consequently, each agent only needs to offer a proportion of its rewards to the agent who impacts its game reward. The situation is different for Symmetrical-3PD, where each agent must incentivise both co-players to cooperate, resulting in a lower general self-interest level. Prior approaches [1–4] are unable to capture this greater general self-interest level for Cyclical-3PD, as they use a single parameter to govern the distribution of incentives.

In our full paper [5], we introduce an algorithm to compute the minimal transfer matrix for normal-form social dilemmas. We provide results for several normal-form social dilemmas applied to different graphs structures. We observe that if the agents are most strongly connected only to a neighbourhood of agents, then the general self-interest level of the game remains stable as the number of agents increases. Conversely, if the connectivity of the agents increases with the number of agents, representing a mixed community, the general self-interest level of the game tends to zero as the number of agents increases. Furthermore, the most connected agent in the network may be the limiting factor.

The structure of the minimal transfer matrix is informative, as agents typically provide incentives only to those who impact their outcomes the most. Games with a sparse minimal transfer matrix inform which agents influence others, and tend to permit agents to retain a greater self-interest level as the number of agents increases.

ACKNOWLEDGMENTS

This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence and a BT/EPSRC funded iCASE Studentship [grant number EP/T517380/1].

REFERENCES

[1] Krzysztof R. Apt and Guido Schaefer. 2014. Selfishness Level of Strategic Games. *Journal of Artificial Intelligence Research* 49 (Feb. 2014), 207–240. https://doi.org/10.1613/jair.4164

[2] Ioannis Caragiannis, Christos Kaklamanis, Panagiotis Kanellopoulos, Maria Kyropoulou, and Evi Papaioannou. 2010. The Impact of Altruism on the Efficiency of Atomic Congestion Games. In *Trustworthly Global Computing*, Martin Wirsing, Martin Hofmann, and Axel Rauschmayer (Eds.). Vol. 6084. Springer Berlin Heidelberg, Berlin, Heidelberg, 172–188. https://doi.org/10.1007/978-3-642-15640-3_12

[3] Po-An Chen, Bart De Keijzer, David Kempe, and Guido Schäfer. 2011. The Robust Price of Anarchy of Altruistic Games. In *Internet and Network Economics*, Ning Chen, Edith Elkind, and Elias Koutsoupias (Eds.). Vol. 7090. Springer Berlin Heidelberg, Berlin, Heidelberg, 383–390. https://doi.org/10.1007/978-3-642-25510-6_33

[4] Po-An Chen and David Kempe. 2008. Altruism, Selfishness, and Spite in Traffic Routing. In *Proceedings of the 9th ACM Conference on Electronic Commerce*. ACM, Chicago Il USA, 140–149. https://doi.org/10.1145/1386790.1386816

[5] Richard Willis, Yali Du, Joel Z. Leibo, and Michael Luck. 2024. Resolving Social Dilemmas with Minimal Reward Transfer. *Autonomous Agents and Multi-Agent Systems* 38, 2 (Oct. 2024), 49. https://doi.org/10.1007/s10458-024-09675-4