The Selfishness Level of Social Dilemmas

Extended Abstract

Stefan Roesch
King's College London
London, United Kingdom
stefan.roesch@kcl.ac.uk

Stefanos Leonardos
King's College London
London, United Kingdom
stefanos.leonardos@kcl.ac.uk

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

ABSTRACT

A key contributor to the success of modern societies is humanity's innate ability to meaningfully cooperate. Game-theoretic reasoning shows however, that an individual's amenity to cooperation is directly linked with the mechanics of the scenario at hand. Social dilemmas constitute a subset of such scenarios where players are caught in a dichotomy between the decision to cooperate, prioritising collective welfare, or defect, prioritising their own welfare. In this work, we study such games through the lens of 'the selfishness level', a standard game-theoretic metric which quantifies the extent to which a game's payoffs incentivize self-directed behaviours. Using this framework, we derive the conditions under which SDs can be resolved and, additionally, produce a first-step towards extending this metric to Markov games. Finally, we present an empirical analysis indicating the positive effects of selfishness-level-directed mechanisms in such environments.

KEYWORDS

Social Dilemma; Game Theory; Markov Game; Reinforcement Learning; Multi-agent Reinforcement Learning

ACM Reference Format:
Stefan Roesch, Stefanos Leonardos, and Yali Du. 2024. The Selfishness Level of Social Dilemmas: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Social dilemmas [9] (SDs) are well studied, and have been the subject of much work in fields such as psychology [3] and sociology [6]. SDs are particularly interesting as they are known to model many real-world coordination problems. A striking example is the case of nuclear weapons proliferation. Here, it is individually rational for a state to maintain a stockpile of nuclear warheads as it serves to deter international conflict. However, when multiple states engage in nuclear arms production, the global community becomes endangered by arms races, geopolitical tensions and, accidental use. Ideally, all states should agree to dismantle their stockpiles, but if any one state were to do so then any opposing, nuclear-armed, states would gain a critical military advantage. The many years of discourse surrounding this problem illustrates that finding solutions to SDs is hard, often requiring external mechanisms to align individual incentives with broader societal goals. Sequential social dilemmas (SeqSDs) [8], extend SDs to the Markov game setting and are well known to more accurately represent the complexities of real-world dilemmas. As such they are used as the standard test-bed for mechanisms such as formal contracting [2], social value orientation [10, 11], inequity aversion [5, 16], and conformity to emergent social norms [15]. In this work we take an interdependence perspective [4], which has recently gained attention in the AI community [10, 11, 16]. We claim that, in simulated scenarios, extrinsic payoffs can be framed as a miss-specification of objective, requiring some external intervention to align with human values. In this light, we investigate the use of the selfishness level [1] as such an intervention mechanism, studying its effects on SDs and extending the notion to the SeqSD setting, empirically verifying its ability to induce agent cooperation.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 SELFISHNESS LEVEL & SOCIAL DILEMMAS

The selfishness level [1] is a scalar metric on the pure Nash equilibria of a normal-form game. Intuitively, a game's selfishness level indicates how much an egotistical player values their own payoff over the collective welfare.

Definition 2.1 (Selfishness level of a normal-form game [1]). Given any normal-form game $G \doteq \{N, \{S_i\}_{i\in N}, \{p_i\}_{i\in N}\}$, we can induce an altruistic game $G(\alpha) \doteq \{N, \{S_i\}_{i\in N}, \{r_i\}_{i\in N}\}$ where, $r_i(s) \doteq p_i(s) + \alpha SW(s)$. The selfishness level of a strategic game G is:

$$\alpha_G = \inf_\alpha\{\alpha \in \mathbb{R}_+ | G \text{ is } \alpha\text{-selfish}\},$$

where, G is α-selfish if, for some $\alpha \geq 0$, a pure Nash equilibrium of $G(\alpha)$ is a social optimum of G.

SDs [9] are a class of normal-form game which emphasise a dichotomy between individual preferences and the collective good:

	C	D
C	R, R	S, T
D	T, S	P, P

Table 1: Outcome categories in the SD payoff matrix

where R denotes the payoff for mutual cooperation, P mutual defection, S cooperation when an opponent defects and T defection when an opponent cooperates. SDs are further defined by a set of inequalities which prescribe the tensions between individual and collective preferences:

$$R > P,\ R > S,\ 2R > T + S, \tag{1}$$

$$T > R \text{ (greed) or } P > S \text{ (fear).} \tag{2}$$

The inequalities in 1 work to establish mutual cooperation as the unique, stable, social optimum and the inequalities in 2 dictate the modality of the SD (e.g., when both inequalities in 2 are satisfied, the resultant game is a prisoner's dilemma).

3 RESOLVING SOCIAL DILEMMAS

Examining SDs through the lens of selfishness level highlights some interesting properties. We delegate proofs for all theorems in this section to the full version of this paper at: kclpure.kcl.ac.uk [12].

Theorem 3.1. *The selfishness level of a SD is*

$$\alpha_G = \begin{cases} 0 & \text{if } T \leq R, \\ \frac{T-R}{2R-T} & \text{if } T > R. \end{cases} \tag{3}$$

Equation 3 shows that when players are troubled only by an equilibrium selection problem (i.e., when G is a stag hunt dilemma), $\alpha_G = 0$. Conversely, when $\alpha_G > 0$ (i.e. a prisoner's or chicken dilemma), the game is not naturally conducive to cooperation. Intuitively, the selfishness level is directly linked with T and R - the greater the value of α_G, the higher the incentive to deviate from mutual cooperation, and vice-versa. As such, the selfishness level formally quantifies the magnitude of the intervention required to realise cooperation.

Here, we investigate how the selfishness level can be used in the design of intrinsic payoff mechanisms to align the players' preferences towards mutual cooperation. Our analysis shows that, in chicken and prisoner's dilemmas, the resultant selfishness level modified payoffs can be relieved of any individual-group tensions, that is, neither inequality in 2 holds.

Theorem 3.2. *Given a SD G, let $T > R$ and $P \leq 0$ (a chicken dilemma). $G(\alpha)$ is always resolved when $\alpha = \alpha_G$.*

The result of Theorem 3.2 reflects the fact that the selfishness level works only to alleviate the burden of greed. As chicken dilemmas are troubled only by greed it is natural that the altruistic game induced by α_G, $G(\alpha_G)$ is free of any dilemma.

Theorem 3.3. *Given a SD G, let $T > R$ and $P > 0$ (a prisoner's dilemma). $G(\alpha)$ is resolved when $\alpha = \alpha_G$ and $P \leq T - R$*

Theorem 3.3 mirrors Theorem 3.2. If the personal benefit of exploitation is the driving force behind a player's willingness to defect then a selfishness level modification of payoffs is able to completely resolve the dilemma. Conversely, if $P > T - R$, $G(\alpha_G)$ is a stag hunt.

3.1 Extending to Markov Games

We present here a 'first step' towards the highly non-trivial goal of theorising the selfishness level in the Markov game setting starting with two-player *sequential social dilemmas* (SeqSDs).

Definition 3.4 (Sequential Social Dilemma [8]). SeqSDs are characterised by the presence of *critical states* $S_c \subseteq S$. Each $s_c \in S_c$ induces a sub-game such that players' preferences can be expressed as a social dilemma. This can be more easily intuited through table 2.

	π^C	π^D
π^C	$R(s_c), R(s_c)$	$S(s_c), T(s_c)$
π^D	$T(s_c), S(s_c)$	$P(s_c), P(s_c)$

Table 2: Empirical payoff matrix for $s_c \in S_c \subseteq S$.

where, $R(s_c) \doteq V_i^{\pi_i^C, \pi_{-i}^C}(s_c)$, and $T(s_c), S(s_c)$ and, $P(s_c)$ are defined analogously.

Our extension is defined via the *altruistic Markov game*, which is analogous to the normal-form game presented in definition 2.1.

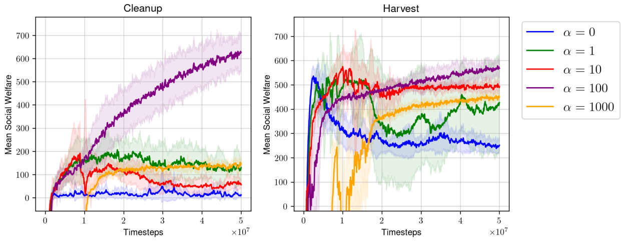

Figure 1: Results in cleanup and harvest [5] averaged over five runs (using PPO [13] with parameter sharing). α is set manually for each experiment, due to the complexity of Γ, with $\alpha = 100$ producing noticeably improved social welfare over both environments.

Definition 3.5 (Altruistic Markov Game). Given a Markov game, $\mathcal{M}$, we can induce an *altruistic* Markov game (cf. 2.1)

$$\mathcal{M}(\alpha) := \{N, S, \{A^i\}_{i \in \{1,\dots,N\}}, P, \{\lambda^i\}_{\{i \in \{1,\dots,N\}}, \gamma\}$$

where $\lambda_i(s, a, s') := R_i(s, a, s') + \alpha(\sum_{j \in N} R_j(s, a, s'))$.

We now define a scalar-valued selfishness level for the SeqSD.

Definition 3.6 (Selfishness Level of Two-Player SeqSDs). Consider the special case of altruistic Markov games where the host game is a two-player SeqSD. Each $s_c \in S_c$ can be considered as a normal-form sub-game. Given this, we construct the set

$$\vec{\alpha} \doteq \{\alpha_{s_c} | \alpha_{s_c} = \frac{T(s_c) - R(s_c)}{2R(s_c) - T(s_c)} \forall s_c \in S_c\},$$

and define the selfishness level of the SeqSD as

$$\Gamma = \max_{s_c} \vec{\alpha}.$$

It is known that, if for some $\alpha \geq 0$ a social optimum of $G(\alpha)$ is Nash, then it remains as such for every $\beta \geq \alpha$ [1]. I.e., for an s_c with selfishness level α_{s_c}, even in the altruistic game $s_c(\beta)$, where $\beta >> \alpha_{s_c}$, the social optima of $s_c(\beta)$ remains Nash. Under this formalism, we have a single, scalar, value Γ describing the selfishness level for the whole Markov game, taking a conservative view with respect to rating a Markov game's cooperativeness. If there is only a single state under which players are able to grossly exploit their peers then the selfishness level of the game becomes, principally, defined by that interaction alone.

4 EXPERIMENTS

We present our empirical analysis (see Fig. 1) studying the effect of a selfishness level inspired reward shaping mechanism in two well-known SeqSDs, 'cleanup' and 'harvest' [5] (*public goods* and *commons* dilemmas [7], respectively) with code adapted from [14].

In both cleanup and harvest, agents are tasked with collecting apples that lie in an orchard. For both scenarios, rewards are acquired exclusively through the collection of apples, with the respective dilemmas arising from the means through which the pool of available apples is replenished.

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 (www.safeandtrustedai.org).

REFERENCES

[1] Krzysztof R. Apt and Guido Schäfer. 2012. Selfishness Level of Strategic Games. In *Algorithmic Game Theory* (Berlin, Heidelberg), Maria Serna (Ed.). Springer Berlin Heidelberg, Amsterdam, NL, 13–24.

[2] Phillip J.K. Christoffersen, Andreas A. Haupt, and Dylan Hadfield-Menell. 2023. Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom) *(AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 448–456.

[3] Robyn M Dawes. 1980. Social dilemmas. *Annual review of psychology* 31, 1 (1980), 169–193.

[4] Harold H. Kelley and John W. Thibaut. 1978. *Interpersonal Relations: A Theory of Interdependence.* John Wiley & Sons, NY, United States.

[5] Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Heather Roff, and Thore Graepel. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 3330–3340.

[6] Peter Kollock. 1998. Social dilemmas: The anatomy of cooperation. *Annual review of sociology* 24, 1 (1998), 183–214.

[7] Peter Kollock. 1998. Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology* 24, 1 (1998), 183–214. https://doi.org/10.1146/annurev.soc.24.1.183

[8] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-Agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (São Paulo, Brazil) *(AAMAS '17)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 464–473.

[9] Michael W. Macy and Andreas Flache. 2002. Learning Dynamics in Social Dilemmas. *Proceedings of the National Academy of Sciences of the United States of America* 99, 10 (2002), 7229–7236. http://www.jstor.org/stable/3057846

[10] Udari Madhushani, Kevin R McKee, John P Agapiou, Joel Z Leibo, Richard Everett, Thomas Anthony, Edward Hughes, Karl Tuyls, and Edgar A Duéñez-Guzmán. 2023. Heterogeneous Social Value Orientation Leads to Meaningful Diversity in Sequential Social Dilemmas. *arXiv preprint arXiv:2305.00768* 0 (2023), 9.

[11] Kevin R. McKee, Ian Gemp, Brian McWilliams, Edgar A. Duèñez Guzmán, Edward Hughes, and Joel Z. Leibo. 2020. Social Diversity and Social Preferences in Mixed-Motive Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 869–877.

[12] Stefan Roesch, Stefanos Leonardos, and Yali Du. 2024. Selfishness Level Induces Cooperation in Sequential Social Dilemmas. In *The 23rd International Conference on Autonomous Agents and Multi-Agent Systems (Extended Abstract)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, (forthcoming). https://kclpure.kcl.ac.uk/ws/portalfiles/portal/245808244/Stefan_AAMAS_Selfishness_paper-3.pdf

[13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017), 12. arXiv:1707.06347 http://arxiv.org/abs/1707.06347

[14] Eugene [Vinitsky, Natasha Jaques, Joel Leibo, Antonio Castenada, and Edward] Hughes. 2019. An Open Source Implementation of Sequential Social Dilemma Games. https://github.com/eugenevinitsky/sequential_social_dilemma_games/issues/182. GitHub repository.

[15] Eugene Vinitsky, Raphael Köster, John P Agapiou, Edgar A Duéñez-Guzmán, Alexander S Vezhnevets, and Joel Z Leibo. 2023. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence* 2, 2 (2023), 26339137231162025.

[16] Jane X. Wang, Edward Hughes, Chrisantha Fernando, Wojciech M. Czarnecki, Edgar A. Duéñez Guzmán, and Joel Z. Leibo. 2019. Evolving Intrinsic Motivations for Altruistic Behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 683–692.