Selfishness Level Induces Cooperation in Sequential Social Dilemmas

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. Modern 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 particularly thorny such scenarios, typically modelled as normal-form or sequential games, where players are caught in a dichotomy between the decision to cooperate with teammates or to defect, and further their own goals. In this work, we study such social dilemmas through the lens of 'selfishness level', a standard game-theoretic metric which quantifies the extent to which a game's payoffs incentivize defective behaviours.

The selfishness level is significant in this context as it doubles as a prescriptive notion, describing the exact payoff modifications necessary to induce players with prosocial preferences. Using this framework, we are able to derive conditions, and means, under which normal-form social dilemmas can be resolved. We also produce a first-step towards extending this metric to Markov-game or sequential social dilemmas with the aim of quantitatively measuring the magnitude to which such environments incentivize selfish behaviours. Finally, we present an exploratory empirical analysis showing the positive effects of using a selfishness level directed reward shaping scheme 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. Selfishness Level Induces Cooperation in Sequential Social Dilemmas. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 8 pages.

1 INTRODUCTION

Social dilemmas [10] are normal-form game theoretic models that capture collective action problems—situations where individuals face a trade-off between their own self-interest and the welfare of the group. In these scenarios, what appears to be the most rational decision from an individual's perspective often leads to negative outcomes for the collective. Social dilemmas are well studied, and have been the subject of much work in fields such as psychology [3] and sociology [6]. The prominence of these games is due to their relevance to many real-world coordination problems. An enlightening example of this is the case of nuclear weapons proliferation between opposing nation states. It is individually rational for a state to maintain a stockpile of nuclear warheads as it serves as a deterrent against potential conflicts. However, when multiple states engage in nuclear arms production, the global community faces increased dangers due to the potential for accidental use, arms races, and geopolitical tensions. The ideal situation here would be for all states to agree to destroy their nuclear stockpiles, but the problem is if any one state were to disarm then any opposing, nuclear-armed, states would gain a threatening military advantage. This illustrates the fact that finding solutions to such dilemmas are difficult and often require external mechanisms that align individual incentives with broader societal goals.

Sequential social dilemmas [9] serve as an extension of social dilemmas which allow for more complex simulations of real-world cooperation problems. They aim to introduce both time and space to normal-form social dilemmas, which model only simple, one-state-one-shot, interactions. Use of sequential social dilemmas has been growing in the multi-agent reinforcement learning community with many researchers adopting them as the theoretical framework for complex environment settings. In this context, they are used as the test-bed for mechanisms, such as formal contracting [2], social value orientation [11][12], inequity aversion [5][19], and conformity to emergent social norms [18], that aim to incentivize agents to act in favour of collective rationality.

Where the simplicity of social dilemmas makes their analytical study tractable, unfortunately sequential social dilemmas introduce a level of complexity that makes a similar study of their properties difficult. As such there is currently a poor understanding of these environments, with most characterisations being reliant on the surface level mechanics of each game paired with the researcher's intuition. The main means of taxonomy comes from the well known distinctions of public good dilemmas/tragedy of the commons dilemmas [6], a method of distinguishing between N-player social dilemmas (where $N > 2$), and through the use of Schelling diagrams [14], the construction of which is reliant on the researcher's assumptions towards what constitutes cooperative and defective behaviour in the given environment. While these categorisation methods are insightful, we propose a shift of focus to a more quantitative understanding of sequential social dilemmas.

The selfishness level [1] provides a quantitative measure on the extent to which games incentivize individualistic (or 'selfish') behaviours. It is defined as the smallest multiple of the social welfare that must be given to each player in order to make a socially optimal strategy profile a Nash equilibrium [13]. This also makes it a prescriptive notion as it details *how* to change the payoff structure of a game in a way that would alter players' payoffs to be more

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.

conducive to cooperation. As such, the selfishness level can be said to share the perspective of interdependence theory [4], which has recently gained attention from AI researchers [11][12][19]. We concur with, and affirm, the following notion:

> humans deviate from game theoretic predictions in economic games because each player acts not on the given matrix of a game - which reflects the extrinsic payoffs set forth by the game rules - but on an "effective matrix", which represents the set of outcomes as subjectively evaluated by that player [12].

This "account" from interdependence theory suggests that humans exhibit strong cooperative ability because, in an economic game, players' preferences are influenced by the *extrinsic* payoffs (the payoffs given by the game rules), but not determined by them. Therefore, in order to improve the cooperative ability of artificial agents, we should aim to design a mechanism that can shape the players' individual preferences in such a way that mutual cooperation becomes desirable. The selfishness level does this by prescribing a set of *intrinsic* payoffs which represent the subjectively experienced payoffs of players. These new payoffs are constructed in a way that shapes players' preferences to be more conducive to cooperation and, under some conditions, can completely resolve the social dilemma.

Our contributions in this work are as follows:

- We find the exact selfishness level of general normal-form social dilemmas and derive the conditions under which social dilemmas can be resolved through the selfishness level,
- We provide a first step towards extending these ideas to sequential social dilemmas, a much more complex setting,
- We perform some exploratory empirical analysis in two, well-known, sequential social dilemmas (harvest and cleanup), illustrating the benefits of our methodology to agent cooperation in a multi-agent reinforcement learning setting.

The following sections are organised as follows: In section 2 we discuss preliminary material, covering the formal definitions of social dilemmas, the selfishness level and sequential social dilemmas. Section 3 covers our analytical contributions with the selfishness level of social dilemmas, the conditions under which the selfishness level can be used to resolve social dilemmas and our preliminary extension of the selfishness level to sequential social dilemmas. Section 4 presents our empirical analysis of our methodology in two sequential social dilemmas, including our experimental setup, hyperparameter settings and interpretations of our results. We then conclude the paper in section 5 with some discussions of the limitations of this work alongside some possible further research directions.

2 PRELIMINARY

2.1 Selfishness Level

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, of that game, values their own payoff over the collective welfare. The higher the selfishness level, the more compensation an egotistical player needs to align their preferences with the welfare of the population. In other words, it is the smallest scalar multiple of the social welfare that needs to be gifted to players in order to turn a socially optimal strategy into a Nash equilibrium.

Definition 2.1 (Selfishness level of a Normal-form Game)**.** 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 then defined as:

$$\alpha_G = \inf_{\alpha}\{\alpha \in \mathbb{R}_+ | G \text{ is } \alpha\text{-selfish}\} \tag{1}$$

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

2.2 Social Dilemmas

Social dilemmas are a class of normal-form game which emphasise a dichotomy between individual preferences and the collective good. Macy et al. [10] describe social dilemmas as:

> mixed-motive, two-person games with two action choices - cooperate (be honest, truthful, helpful etc.) or defect (lie, cheat, steal, etc.).

More formally, social dilemmas are symmetrical normal-form games with payoff matrices similar to Table 1.

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

Table 1: Outcome categories in the payoff matrix

where the payoff categories are described as follows:

- R or 'reward' which denotes the payoff for mutual cooperation,
- P or 'punishment' which denotes the payoff for mutual defection,
- S or 'sucker' which denotes the payoff for a cooperating player when their opponent defects,
- T or 'temptation' which denotes the payoff for a defecting player when their opponent cooperates.

Social dilemmas are defined by a set of four inequalities which prescribe the tensions between individual and group preferences:

Inequality		Preference
$R > P$	(2)	the individual prefers mutual cooperation (C, C) to mutual defection (D, D).
$R > S$	(3)	the individual prefers mutual cooperation to unilateral cooperation (C, D).
$2R > T + S$	(4)	the group prefers mutual cooperation to unilateral cooperation/defection.
$T > R$ or $P > S$	(5)	the individual prefers unilateral defection (D, C) to mutual cooperation (aka, greed) OR the individual prefers mutual defection to unilateral cooperation (aka fear).

Table 2: Set of inequalities that define social dilemmas. They prescribe tensions between individual and group preferences.

Equations 2-4 work to establish mutual cooperation as the unique, stable, social optimum. The final two inequalities in equation 5 dictate the modality of the social dilemma, of which there exist three:

(1) $T > R$ and $P > S$: which we denote as 'prisoner's dilemmas',
(2) $T > R$ and $P \leq S$: which we denote as 'chicken dilemmas',
(3) $T \leq R$ and $P > S$: which we denote as 'stag hunt dilemmas'.

In prisoner's and chicken dilemmas, the dichotomy between individual preferences and collective good is, most plainly, expressed through the fact that the socially optimal strategy profile (C, C) is not a Nash equilibrium (I am better off defecting when you cooperate). In stag hunt however, mutual cooperation is, in fact, already a Nash equilibrium solution. This illustrates the notion that, under a population of individually rational players, the absence of a Nash equilibrium is not the only factor influencing players to stray from mutually cooperative behaviours. [16]. Stag hunts, for instance, are dichotomised by the *risk* associated with trusting partner players' adherence to cooperative strategies (if I cooperate but you defect, I am worse off than if I were to defect) as is dictated by the 'fear' condition in equation 5.

For our experiments, we investigate how an altruistic reward affects players' collective performance in two, well known, sequential social dilemmas [9]. First we shall describe the notion of Markov games and then move to sequential social dilemmas, which are a special case of Markov games.

2.3 Markov Games

Markov games (or stochastic games [8]) have been adopted as the standard formalisation for problem settings in the multi-agent reinforcement learning literature. They are a temporal and spacial extension of normal-form games, where each sub-game, or state s (not to be confused with the, similar, use of s to denote a strategy profile), can be seen as normal-form. The rewards $R_i(s, \vec{a})$ for each joint action $\vec{a} = \{a_1, ..., a_N\}$ available in state s are equivalent to payoffs for a particular strategy in that state.

Definition 2.2 (Markov Game [20])**.** A Markov game is defined as a tuple $\{N, S, \{A^i\}_{i\in\{1,...,N\}}, P, \{R^i\}_{\{i\in\{1,...,N\}}, \gamma\}$ where,

- N is the number of agents
- S is the set of, Markovian, environment states
- A^i is the set of actions availiable to agent i
- $P : S \times A^1 \times ... \times A^N \rightarrow \Delta(S)$, where $\Delta(S)$ denotes a distribution over the state space S, is the state transition function, providing the probability of transitioning from a state $s \in S$ to the next state $s' \in S$ given joint
- $R^i : S \times A^1 \times ... \times A^N \times S \rightarrow \mathbb{R}$ is the reward function, returning a scalar value to the ith agent that describes the quality of a state transition
- $\gamma \in [0, 1]$ is the discount factor which determines how much the agent values future rewards.

Sequential social dilemmas [9] are a special case of Markov games. Similarly to social dilemmas, sequential social dilemmas emphasise a dichotomy between individual preferences and the collective good but express this dichotomy through the more spatially and temporally complex setting of Markov games.

Definition 2.3 (Sequential Social Dilemma)**.** A sequential social dilemma is defined as a tuple $\langle \mathcal{M}, \Pi = \Pi^C \cup \Pi^D \rangle$ where $\mathcal{M}$ is a Markov game as defined in 2.2 and Π is a policy space which is constituted by the union of a set of cooperative policies $\pi^c \in \Pi^c$ and defecting policies $\pi^D \in \Pi^D$. A key assumption underpinning the notion of sequential social dilemmas is the existence of a set of critical states $S_c \subseteq S$ where each state $s_c \in S_c$ induces a normal-form sub-game where, players' preferences can be expressed as a social dilemma. The similarities between critical state sub-games and the normal-form social dilemma can be seen in Table 3.

	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 3: Empirical payoff matrix for the sub-game induced by a critical state of a sequential social dilemma

where,

$$R(s_c) \doteq V_i^{\pi_i^C, \pi_{-i}^C}(s_c)$$
$$P(s_c) \doteq V_i^{\pi_i^D, \pi_{-i}^D}(s_c)$$
$$S(s_c) \doteq V_i^{\pi_i^C, \pi_{-i}^D}(s_c)$$
$$T(s_c) \doteq V_i^{\pi_i^D, \pi_{-i}^C}(s_c).$$

3 METHODOLOGY

To simplify our following analysis we re-state the definition of social dilemma. We construct a new game, without loss of generality, by applying the positive affine transformation $p_i(s) - S, \ \forall s \in \{S_i\}_{i\in N}$:

	C	D
C	$R - S, R - S$	$S - S, T - S$
D	$T - S, S - S$	$P - S, P - S$

and simplify notation:

	C	D
C	R, R	$0, T$
D	$T, 0$	P, P

and finally, re-write the social dilemma inequalities as $R > P, R > 0$, $2R > T$, and either: $T > R$ or, $P > 0$.

3.1 Selfishness Level of Social Dilemmas

Examining social dilemmas through the lens of selfishness level, highlights interesting, and intuitive, properties of social dilemmas. To find the selfishness level of a normal-form game, it suffices to, first, derive the conditions under which the social optima of the altruistic modification of that game also become Nash equilibria and, second, to identify the smallest value of α under which those conditions are satisfied.

Theorem 1. The selfishness level of a social dilemma is

$$\alpha_G = \begin{cases} 0 & \text{if } T \leq R, \\ \frac{T-R}{2R-T} & \text{if } T > R. \end{cases} \tag{6}$$

Proof. Recall that the unique, stable, social optimum of a social dilemma is obtained through mutual cooperation (equations 2 - 4), easing this process to simply finding the exact α under which (C, C) becomes Nash. Also recall that there exist three, distinct, modalities of social dilemma:

(1) $T > R$ and $P > S$: *Prisoner's Dilemmas*

(2) $T > R$ and $P \leq S$: *Chicken Dilemmas* and,
(3) $T \leq R$ and $P > S$: *Stag Hunt Dilemmas*.

It is straightforward to see that, in stag hunt dilemmas, (C, C) is a Nash equilibrium. This means that, if the social dilemma is a stag hunt, the selfishness level is $\alpha_G = 0$. To see the selfishness level in prisoner's and chicken dilemmas, we introduce notation for the payoffs of the altruistic modification of G, $G(\alpha)$ (see Table 4).

	C	D
C	R', R'	S', T'
D	T', S'	P', P'

Table 4: Payoff matrix for $G(\alpha)$

where,

$$R' = R + \alpha 2R,$$
$$T' = T + \alpha T,$$
$$S' = \alpha T,$$
$$P' = P + \alpha 2P.$$

For both prisoner's dilemmas and chicken dilemmas, (C, C) is not a Nash equilibrium in G as $T > R$. For (C, C) to be a Nash equilibrium in $G(\alpha)$, the following must hold:

$$\begin{aligned} R' &\geq T' \\ R' - T' &\geq 0 \\ (R + \alpha 2R) - (T + \alpha T) &\geq 0 \end{aligned} \tag{7}$$

Changing the inequality to an equality and solving for α gives us the lowest bound on α which satisfies the condition:

$$\alpha = \frac{T - R}{2R - T} \tag{8}$$

□

Equation 6 gives some insight into the information presented by the selfishness level under the context of social dilemmas. Namely, when $\alpha_G = 0$, i.e. when G is a stag hunt dilemma, the game is already conducive to cooperation among individually rational agents (players are troubled only by an equilibrium selection problem). However, when $\alpha_G > 0$, i.e. prisoner's or chicken dilemmas, the game is not conducive to cooperation among individually rational agents (mutual cooperation is dominated by unilateral deviation). As such, we require some external intervention on player's preferences to make cooperation possible. For these kinds of dilemmas, the selfishness level can give us insight into the magnitude of the intervention required. Equation 8 formally quantifies the magnitude of the incentive for agents to deviate and how that incentive relates to the payoffs of the game. The higher the temptation payoff T or the lower the reward payoff R, the higher the selfishness level α and, hence, the more incentive to deviate from the cooperative outcome. Conversely, when α is low, the game is more favourable towards cooperation, as players have less incentives to deviate.

3.2 Resolving social dilemmas with the selfishness level

The selfishness level is a measure of how much an egotistical player values their own payoff over the collective good in a game. It can be used to analyse how the game's characteristics influence the players' willingness to cooperate or deviate from the cooperative outcome. In the following section, we also investigate how the selfishness level can help us design mechanisms or incentives that can align the players' preferences with the social welfare and foster cooperation among individually rational agents.

As has been established, the selfishness level α_G provides information on how to modify player's payoffs in such a way that cooperation becomes possible. In stag hunt games, $\alpha_G = 0$, meaning that cooperation is already possible among individually rational players (it is an equilibrium selection problem). However, when $\alpha_G > 0$ (as is the case in prisoner's and chicken dilemmas), mutual cooperation cannot be Nash. If we inspect the properties of the resulting altruistic game we find that for chicken dilemmas and a subset of prisoner's dilemmas, we are able to resolve the social dilemma entirely.

Theorem 2. Given a social dilemma G, let $T > R$ and $P \leq 0$ (a chicken dilemma). If $G(\alpha_G)$ is the altruistic game of G where $\alpha = \alpha_G$, then $G(\alpha_G)$ is resolved.

Proof. Given a chicken dilemma, we have the following payoffs in $G(\alpha_G)$

$$R' = R + \frac{T - R}{2R - T} 2R, \tag{9}$$
$$T' = T + \frac{T - R}{2R - T} T, \tag{10}$$
$$S' = \frac{T - R}{2R - T} T, \tag{11}$$
$$P' = P + \frac{T - R}{2R - T} 2P. \tag{12}$$

For $G(\alpha_G)$ to be resolved, we need to have $T' \leq R'$ and $P' \leq S'$. Given equations 7 and 8, we already have that $T' = R'$. The second inequality follows from the following claims

- Claim 1: $S' > 0$.Recall that in a chicken dilemma $T > R$. As $T > R > 0$ and $2R > T$, $\frac{T-R}{2R-T} > 0$ hence $S' > 0$.
- Claim 2: $P' \leq 0$. Recall that in a chicken dilemma, $P \leq 0$. If $P = 0$, $P' = 0$. If $P < 0$, $P' < 0$

Combining the above claims, we get that $P' < S'$. Hence, $G(\alpha_G)$ is resolved. □

Theorem 2 illustrates the fact that the selfishness level provides a measure of the influence of greed in social dilemmas. As chicken dilemmas are troubled only by greed it is natural that an altruistic game with $\alpha = \alpha_G$ completely resolves the dilemma. Our next result shows that, under certain conditions, the selfishness level modified altruistic game of prisoner's dilemma is also completely resolved.

Theorem 3. Given some social dilemma, let $T > R$ and $P > 0$ (a prisoner's dilemma). If $G(\alpha_G)$ is the altruistic game of G where $\alpha = \alpha_G$, then $G(\alpha_G)$ is resolved when $P \leq T - R$.

Proof. Given a prisoner's dilemma, we have payoffs consistent with equations 9 - 12 in $G(\alpha_G)$. For $G(\alpha_G)$ to be resolved, $T' \leq R'$ and $P' \leq S'$. Given equations 7 and 8, $T' = R'$. We can set $P' \leq S'$ and simplify to find the appropriate bound:

$$P' \leq S' \implies P + \frac{T - R}{2R - T} 2P \leq \frac{T - R}{2R - T} T \tag{13}$$

which after some simple algebra leads to $P \leq T - R$ as claimed. □

Theorem 3 shows that, if $P \leq T - R$, then the prisoner's dilemma will be completely resolved. Conversely, if $P > T - R$, $G(\alpha_G)$ is a stag hunt (a social dilemma devoid of greed).

3.3 First Steps Towards an Extension to Markov Games

While interesting, remaining in the domain of normal-form games has limited applicability to real-world problems. In this light, we look towards an exploration of the extension of the ideas presented in previous sections to the, much more complex and real-world adjacent, Markov game setting. A perfect extension of selfishness level to the Markov game setting is a highly non-trivial task as the additional spacio-temporal complexities are vast and difficult, even, to quantify. We present here, what we believe to be, a 'first step' towards this goal by theorising an idea of selfishness level in two-player sequential social dilemmas (a special case of Markov games).

We start by introducing the *altruistic Markov game*, which is analogous to the altruistic normal-form game presented in definition 2.1 but in a Markov game setting.

Definition 3.1 (Altruistic Markov Game)**.** Given a Markov game

$$\mathcal{M} \coloneqq \{N, S, \{A^i\}_{i\in\{1,\dots,N\}}, P, \{R^i\}_{\{i\in\{1,\dots,N\}}, \gamma\}$$

using the language of [1] we can induce an *altruistic* Markov game

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

where $\lambda_i(s, a, s') \coloneqq R_i(s, a, s') + \alpha(\sum_{j\in N} R_j(s, a, s'))$. Note that the notion of social welfare is replaced by a simple summation of immediate rewards over each agent for that state.

Consider the special case of altrustic Markov games where the host Markov game is a two-player sequential social dilemma as defined in definition 2.3. In this case, for each critical-state s_c, we have empirical payoffs:

$$\begin{aligned}
R'(s_c) &\doteq V_i^{\pi_i^C, \pi_{-i}^C}(s_c) + \alpha(2V_i^{\pi_i^C, \pi_{-i}^C}(s_c)) \\
&\doteq R(s_c) + \alpha(2R(s_c)) \\
P'(s_c) &\doteq V_i^{\pi_i^D, \pi_{-i}^D}(s_c) + \alpha(2V_j^{\pi_j^D, \pi_{-j}^D}(s_c)) \\
&\doteq P(s_c) + \alpha(2P(s_c)) \\
S'(s_c) &\doteq V_i^{\pi_i^C, \pi_{-i}^D}(s_c) + \alpha(V_j^{\pi_j^C, \pi_{-j}^D}(s_c) + V_j^{\pi_j^D, \pi_{-j}^C}(s_c)) \\
&\doteq S(s_c) + \alpha(S(s_c) + T(s_c)) \\
T'(s_c) &\doteq V_i^{\pi_i^D, \pi_{-i}^C}(s_c) + \alpha(V_j^{\pi_j^C, \pi_{-j}^D}(s_c) + V_j^{\pi_j^D, \pi_{-j}^C}(s_c)) \\
&\doteq T(s_c) + \alpha(S(s_c) + T(s_c)).
\end{aligned}$$

As s_c can be considered a normal-form sub-game of the overall sequential social dilemma, it is safe to say that the notion of selfishness level easily extends to such a characterisation of state. As such, the selfishness level of s_c can be found via equation 6. Given this formalisation we propose the selfishness level of the sequential social dilemma by constructing the set of selfishness levels for all critical states $\vec{\alpha} \doteq \{\alpha_{s_c} | \alpha_{s_c} = \frac{T(s_c)-R(s_c)}{2R(s_c)-T(s_c)} \forall s_c \in S_c\}$. This set is constructed such that $|\vec{\alpha}| = |S_c|$ ($\vec{\alpha}$ contains the selfishness level of every critical state) and we assume that $\forall \alpha_{s_c} \in \vec{\alpha},\ \alpha_{s_c} < \infty$ (which is reasonable as each critical state sub-game is considered a social dilemma). We then say that the selfishness level of the sequential social dilemma is

$$\Gamma = \max_{s_c} \vec{\alpha} \tag{14}$$

It is important to note the following property of altruistic games: if for some $\alpha \geq 0$ a social optimum of $G(\alpha)$ is a Nash equilibrium, then it is also the case for every $\beta \geq \alpha$ [1]. This means that, for a given s_c with selfishness level α_{s_c}, even in the altruistic game $s_c(\beta)$, where $\beta >> \alpha_{s_c}$, we can still retain the property that the social optima of $s_c(\beta)$ are Nash equilibria. Under this formalism, we have a single, scalar, value Γ which describes the selfishness level for the whole Markov game. More intuitively, it can be said that this formalism takes a conservative view with respect to rating a Markov game's cooperativeness. In other words, given our definition, the selfishness level of a Markov game is equal to, and as such completely dictated by, that of it's most selfish state. This means that, even if there is only a single state under which players are able to grossly exploit their peers, then the selfishness level of the whole game is defined by that interaction alone.

3.4 Rewards propagating to values

Sequential social dilemmas are defined based on player preferences over long-term value functions instead of immediate rewards. It is well known that value functions are, at best, expensive to obtain making a direct translation of selfishness level to the value function impossible for most complex applications. In this light, we avoid directly manipulating values and, instead, adopt the reward shaping method shown by λ_i in definition 3.1. We show that such a reward shaping technique propagates intrinsic rewards to the value function indirectly. As shown in definition 3.1 we shape the reward of each agent i according to $\lambda_i(s, a, s')$ where,

$$\lambda_i(s, a, s') \coloneqq R_i(s, a, s') + \alpha \sum_{j\in N} R_j(s, a, s')$$

with α consistent with equation 6. Given agent i's value function under the *extrinsic rewards alone* $V_i^r(s)$ where,

$$V_i^r(s) \coloneqq \mathbb{E}_{\vec{\pi}}\left[\sum_{t=0}^{T} \gamma^t R_t\right]$$

we want our *intrinsic* value function to be of the following form:

$$V_i^\lambda(s) = V_i^r(s) + \alpha\left(\sum_{j\in N} V_j^r(s)\right). \tag{15}$$

We verify that rewards are adequately propagated to the values as follows:

$$\begin{aligned}
V_i^\lambda(s) &= \mathbb{E}_{\vec{\pi}}\left[\sum_t \gamma^t \left(R_t^i + \alpha \sum_{j\in N} R_t^j\right) \mid s_0 = s\right] \\
&= \mathbb{E}_{\vec{\pi}}\left[\sum_t \gamma^t R_t^i\right] + \mathbb{E}_{\hat{\pi}}\left[\sum_t \alpha\gamma^t \sum_{j\in N} R_t^j\right] \\
&= \mathbb{E}_{\vec{\pi}}\left[\sum_t \gamma^t R_t^i\right] + \alpha\left(\sum_{j\in N} \mathbb{E}_{\vec{\pi}}\left[\sum_t \gamma^t R_t^j\right]\right) \\
&= V_i^r(s) + \alpha\left(\sum_{j\in N} V_j^r(s)\right).
\end{aligned} \tag{16}$$

The goal of each agent becomes the maximisation of the expected payoff in 16.

Figure 1: Illustration of cleanup. To the left is a river, represented by blue pixels with the brown pixels representing pollution in the river. To the right is an orchard with green pixels representing apples. Black pixels represent empty space and bright coloured pixels represent agents.

4 EXPERIMENTS

Here, we present experiments in which we study the effects on the cooperative performance of agents under various values of α (in equation 16) in two well-known mixed-motive sequential social dilemmas [17]. This is motivated by the model of selfishness level in sequential social dilemmas presented in section 3.3.

4.1 Environments

In 'cleanup' (a *public goods dilemma* [7]) agents are rewarded by collecting apples which spawn randomly in an orchard. The spawn rate of these apples however is tied to the cleanliness of a nearby river. As time goes on the river is progressively polluted, lowering the apple spawn rate, until a *saturation point* is met at which apples no longer spawn. The dilemma here is characterised by the fact that, in order to maintain an abundance of apples, agents must sacrifice some reward in the short-term to clean the river (for which they receive no reward). An effective policy in cleanup is one which effectively balances time spent cleaning pollution with time spent collecting apples.

In 'harvest' (a *commons dilemma* [7]) agents are, again, rewarded by collecting apples but there is no river or pollution which controls their spawn rate. Instead, the spawn rate is tied to the amount of apples available in the local area. The more apples in the area, the more likely it is that a new apple will spawn nearby. If no apples are available then new apples are spawned with a very low probability. Here, if agents are considerate only of their short-term gains then they will quickly collect all of the apples, reducing the spawn rate of new apples which will stifle their returns. If agents show restraint however, then a high apple spawn rate can be maintained and more apples collected overall.

4.2 Setup

Both environments are run under partial observability (as is consistent with current literature) with each agent's observation consisting of a 15×15 pixel view window, centred on the respective agent's current location.

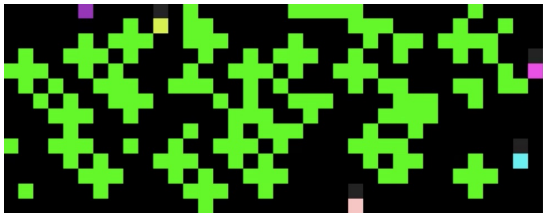

Figure 2: Illustration of harvest. Green pixels represent apples, black pixels represent empty space and bright coloured pixels represent agents.

Our codebase is derived from an open source repository [17]. We use proximal policy optimisation (PPO) [15] as the base learning algorithm for our policies with agents sharing network parameters. PPO utilises a pair of deep neural networks, the actor and the critic respectively. Both networks take, as input, the agents' observations which are fed to a convolutional layer, followed by 3 dense, linear, layers and finally the output layers outputting actions and advantage estimations for the actor and critic networks respectively. The actor network represents a parameterised policy which is updated according to the policy gradient loss $\mathbb{E}_t[\log \pi_\theta(a_t|s_t)A_t]$, where A_t is the advantage function estimated by the critic network. We ran our experiments over 12 parallel environments with 5 agents per environment over $5e7$ total timesteps. Training was split into episodes, where $T = 1000$. We maintained a batch size of $12,000$ and learning rate of 0.0001 for all experiments.

For each environment, we ran experiments with values for $\alpha \in \{0, 1, 10, 100, 1000\}$ to explore how a selfishness level of the type introduced in equation 14 would affect the social welfare of groups of agents in sequential social dilemmas. We set boolean variables a_t^i, z_t^i and p_t^i to 'true' when agent i picks up an apple, is hit by another agent's zapper beam or cleans up a tile of pollution at time t respectively and 'false' otherwise. We additionally use the indicator function $\mathbb{I}(.)$ which returns a value of 1 when it's input is 'true' and 0 otherwise.

The primary metric used to judge the population of agents' tendency to cooperate is the social welfare (SW)

$$SW_t = \sum_{i \in N} R_t^i$$

where, R_i is the extrinsic reward given by the environment to agent i at time t. We also plot the number of apples consumed (AC) and the number of times agents are hit with a zapper (Z) where,

$$AC_t = \sum_{i \in N} \mathbb{I}(a_t^i)$$

$$Z_t = \sum_{i \in N} \mathbb{I}(z_t^i)$$

Finally, we plot (exclusively for cleanup) the Gini coefficient over apples consumed ($Gini$) and the number of pollution tiles cleaned (P) where,

$$Gini_t = \frac{\sum_{i \in N} \sum_{j \in N} |\sum_T \mathbb{I}(a_t^i) - \sum_T \mathbb{I}(a_t^j)|}{2N \sum_i \mathbb{I}(a_t^i)}$$

$$P_t = \sum_{i \in N} \mathbb{I}(p_t^i).$$

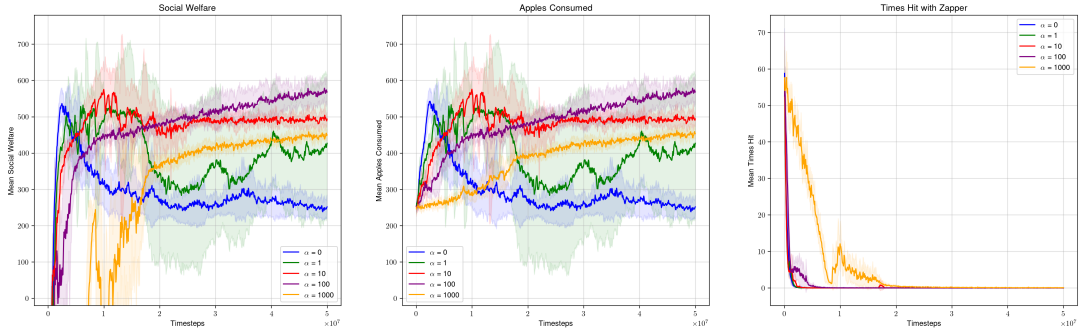

Figure 3: Performance of varying α values under the harvest environment. Bold lines represent the rolling average of the respective metric over 5 runs with the shaded areas surrounding representing the standard deviation

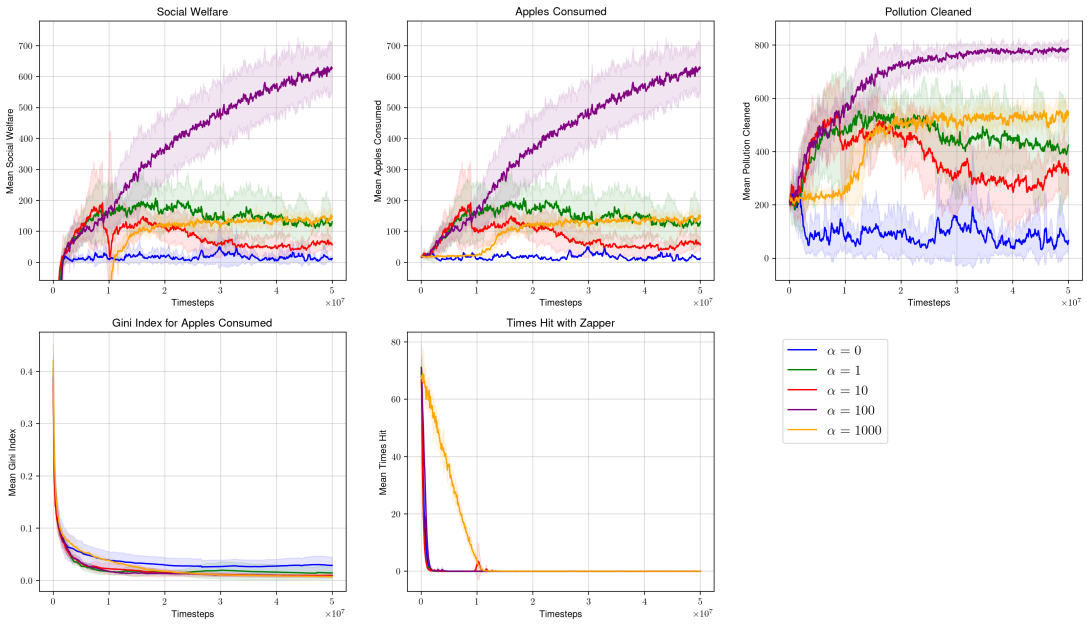

Figure 4: Performance of varying α values under the cleanup environment. Bold lines represent the rolling average of the respective metric over 5 runs with the shaded areas surrounding representing the standard deviation

4.3 Results

Figure 3 shows our results in the harvest environment. We observe that, in harvest, agent learning is generally noisy. Agents with $\alpha > 0$ tend to outperform agents with $\alpha = 0$ but for agents with $\alpha = 1$ or $\alpha = 10$, there is a large variance in performance between runs. Similarly to cleanup, we observe that $\alpha = 100$ results in the best performing agents overall.

Figure 4 shows our results in the cleanup environment. Here, we observe relatively strong performance from agents with $\alpha = 100$. We hypothesise that this is because, under other values of α, agents find themselves stuck in local optima (α is not high enough to for agents to converge to collectively optimal policies) or, in the case of $\alpha = 1000$, the inflation of the rewards given cause the perception of good or bad behaviours to become muddied. We further observe that, even though there is no reward for doing so, our methodology instils agents with the desire to remove pollution from the river. We hypothesise that, similarly to the normal form case, this is because our intrinsic reward method bridges the gap between individually rational behaviours and collectively rational behaviours - agents

value functions are shaped such that a direct link is made between agents' own rewards and the social welfare.

In both environments, those agents with $\alpha > 0$ (agents that are not purely independent learners) consistently produce higher social welfare and apple consumption than those without. We also find that, in both environments, agents quickly learn to avoid zapping each other. This is likely due to the strong negative consequences to agent's rewards, and hence the social welfare, associated with being zapped. In the case of $\alpha = 0$ however, this is not the case as agent's individual rewards are independent of others rewards. In this case, and as a compounding affect to agents with $\alpha > 0$, we hypothesise that agents quickly learn to avoid zapping others due to the lack of positive feedback associated with the action of zapping. i.e. it wastes time that could be better spent acquiring apples. This drop-off is not so quickly realised with $\alpha = 1000$ - we hypothesise that this is due to the large amount of reward inflation in that setting. We also hypothesised that, due to the simplicity of the method of reward shaping, team performance would increase monotonically with α. Interestingly, and especially in cleanup, our results show that this is not the case suggesting that the value of α is indeed meaningful. We observe that agents learning under $\alpha = 1000$ perform strictly worse than those with $\alpha = 100$ and, in cleanup, agents with $\alpha = 10$ perform worse on average than those with $\alpha = 1$. Here, we conjecture that the best performing values of α are those that are closest to the true selfishness level (as described in equation 14) of the respective environment.

Note that, for our social welfare plots, we have anchored the y-axis to just below zero. This is to improve legibility of the plots as at the start of learning, agents exhibit a high frequency of zapping causing the social welfare to start extremely low.

5 CONCLUSION

In this work, we explored the effectiveness of analysing social dilemmas through the lens of their selfishness levels. We have derived some interesting properties of social dilemmas in the normal-form case, finding exact conditions under which a selfishness level modification of the game's payoffs can result in complete resolution of the dilemma. We have extended this work by providing a first-step to bringing a similar analysis to the more complex setting of sequential social dilemmas. Going further, our empirical results in this setting, suggest that our method can indeed provide a benefit to the cooperative performance of learning agents in sequential social dilemmas. The overall impact of our method is to add additional (socially optimal) equilibria to the strategy (or policy) space, not to prescribe any particular solution. While we hypothesise an increase in the probability of convergence to socially optimal joint policies, we suspect that the problem of *equilibrium selection* is still present within our method.

6 ACKNOWLEDGEMENTS

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 G. Phillips, Karl Tuyls, Edgar A. Duéñez-Guzmán, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin R. McKee, Raphael Koster, Heather Roff, and Thore Graepel. 2018. *Inequity Aversion Improves Cooperation in Intertemporal Social Dilemmas.* Deepmind. arXiv:1803.08884 [cs, q-bio] http://arxiv.org/abs/1803.08884

[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 arXiv:https://doi.org/10.1146/annurev.soc.24.1.183

[8] L. S. Shapley. 1953. Stochastic Games. *Proceedings of the National Academy of Sciences* 39, 10 (1953), 1095–1100. https://doi.org/10.1073/pnas.39.10.1095 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.39.10.1095

[9] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. *Multi-Agent Reinforcement Learning in Sequential Social Dilemmas.* Deepmind. arXiv:1702.03037 [cs] http://arxiv.org/abs/1702.03037

[10] 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

[11] 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.

[12] 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.

[13] John Nash. 1951. Non-Cooperative Games. *Annals of Mathematics* 54, 2 (1951), 286–295. http://www.jstor.org/stable/1969529

[14] Thomas C Schelling. 1973. Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities. *Journal of Conflict resolution* 17, 3 (1973), 381–428.

[15] 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

[16] Brian Skyrms. 2012. *The stag hunt and the evolution of social structure.* Cambridge University Press, Cambridge, England.

[17] 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.

[18] 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.

[19] 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.

[20] Yaodong Yang and Jun Wang. 2021. *An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective.* University College London. arXiv:2011.00583 [cs] http://arxiv.org/abs/2011.00583