Simulating Human Cooperation through Inequity Aversion: An Agent-Based Approach to Dynamic Social Dilemmas

Candidate Number: 31326

Capstone Project

Submitted to the Department of Methodology

at the London School of Economics and Political Science

in partial fulfilment of

MSc in Applied Social Data Science

Supervisor:

Dr Friedrich Geiecke

Date of Submission: August 2024

Word Count: 9670 words

Abstract

The emergence and persistence of cooperation in complex social dilemmas have long puzzled researchers across disciplines, from economics and psychology to political and social sciences. While human societies often demonstrate remarkable abilities to cooperate in the face of conflicting individual and collective interests, classical economic models have struggled to fully explain this phenomenon, particularly in realistic, temporally extended scenarios. This paper addresses this challenge by employing Multi-Agent Reinforcement Learning (MARL) to model and analyse social dilemmas in dynamic contexts. By incorporating inequity aversion into the reward functions of individual agents, we extend insights from behavioural economics beyond simple matrix games to more realistic environments. In this research we build 2 realistic, videogame like environment, and integrated inequity aversion into agent-based modelling to explore how large-scale cooperation can emerge and persist. Our approach demonstrates that inequity aversion, when modelled in a dynamic MARL setting, promotes cooperation across various types of intertemporal social dilemmas, primarily through improved temporal credit assignment through a direct and indirect mechanism that is analogous to guilt or envy. Notably, our models successfully reproduce turn-taking behaviour among agents without explicitly training them for such interactions, suggesting an emergent property driven by the inherent dynamics of inequity aversion. This finding is particularly significant for intertemporal social dilemmas, where the benefits of cooperation are delayed or distributed over time. By elucidating the role of inequity aversion in dynamic contexts, this research contributes to our understanding of how large-scale cooperation may emerge and persist in human societies, offering implications for sustainable resource management, public goods provision, and collective action problems in an increasingly interconnected world.

Keywords: Social Dilemma, Inequity Aversion, Agent Based Modelling, Multi-Agent Reinforcement Learning, Social Science Theory

Introduction

The study of human cooperation in temporally extended social dilemmas is crucial for addressing some of the most significant challenges of our time. From the governance of natural resources to the global efforts to combat climate change (Acheson, 1981), cooperation serves as the essential glue binding collective action. In temporally extended social dilemmas, individuals often face a trade-off between immediate personal gain and the broader, long-term collective interest (Janssen, 2010). Classic models of human behaviour, grounded in rational choice theory, often predict that cooperation in such situations is unlikely or impossible (Olson, 1965). However, this presents a paradox, as humans routinely find ways to cooperate in everyday intertemporal social dilemmas, a fact well-documented through decades of fieldwork and laboratory experiments (Besley & Coate, 2003). Understanding and explaining how individual behaviours coalesce into societal cooperation remains a fundamental goal across disciplines such as economics, psychology, sociology, and behavioural science.

Previous studies have explored the problem of cooperation in social dilemmas through extensive behavioural experiments, such as investigating the effects of social preferences on maintaining stable cooperation in repeated matrix games (Charness & Rabin, 2002), the role reputation management in experiments, surveys, and longitudinal studies (Wu et al., 2016), identification of individuals through scoring in matrix games (Wedekind & Milinski, 2000), and indirect reciprocity through large-scale field experiments (Yoelia et al., 2013). However, these studies have often employed a reductionist approach to explain human cooperation in social dilemmas, typically treating cooperation and defection as isolated, atomic decisions. While they have provided valuable frameworks for modelling, the limitations in their methodologies make the modelling of human cooperation in realistic environments challenging.

Inspiration and Limitation of Behavioural Game Theory Studies

A central paradigm in the research of cooperation within social dilemmas is the use of repeated matrix games, a fundamental tool in behavioural game theory used to analyse strategic interactions between two or more players (Auer et al., 2003). Each player in a matrix game chooses a strategy from a set of possible actions, and the outcomes (or payoffs) of these choices are represented in a payoff matrix (Fehr & Schmidt, 1999). This matrix shows the payoffs for each player based on the combination of strategies chosen by all players involved. For example, in a simple two-player game like the Prisoner's Dilemma, each player can choose to either cooperate or defect. The resulting payoffs, depending on the combination of these choices, are displayed in the matrix. Each player's goal is to choose a strategy that maximizes their own payoff, given the strategies they expect the other player to choose. These games have been effectively utilized in the various studies of social dilemmas, exploring mechanisms through which the socially preferred outcome of mutual cooperation can be maintained (Peysakhovich et al., 2014) Examples include modelling using on reinforcement learning models (Macy & Flache, 2002) and MARL (Bloembergen et al., 2015; Sandholm & Crites, 1996; Wunder et al., 2010). Frameworks based on emotions (Yu et al., 2015), direct and indirect reciprocity (Nowak & Sigmund, 1998; Nowak & Sigmund, 1993; Nowak & Sigmund, 1992) norm enforcement (Axelrod, 1986), spatial structures (Nowak & May, 1992), and the effects of social networks (Ohtsuki et al., 2006). While these studies have provided valuable insights and frameworks for understanding cooperation, the repeated matrix game paradigm has inherent limitations. These limitations of the repeated matrix game paradigm arise from its highly abstractive representation of strategic interaction and assume that the incentive structure of cooperation and defection can be faithfully mapped onto atomic choices—discrete, isolated decisions to cooperate or defect (De Cote et al., 2006; Sandholm & Crites, 1996). The standard matrix form games are limited in their ability to capture several key aspects of real-world dynamics:

1. **Temporal Extension**: Real-world social dilemmas unfold over time, unlike the instantaneous decisions in matrix games.
2. **Behavioural Complexity**: Cooperation and defection are labels applied to behaviours that develop over time, not simple, one-time choices in matrix games.
3. **Gradation of Cooperativeness**: In reality, cooperativeness is a spectrum, we can choose much we cooperate, not a binary state as often presented in matrix games.
4. **Quasi-simultaneous Decision-Making**: In reality, decision-making occurs only quasi-simultaneously, with players adjusting their strategies based on evolving information about others' actions, whereas matrix games typically assume decisions are made simultaneously without such interaction.
5. **Partial Observability**: Unlike in matrix games, real-world decisions are often made with only partial knowledge of the state of the world and other players' activities.

The Need for a New Approach

These limitations reveal the first significant gap in current research. While previous research has provided a range highly simplistic yet sophisticated theories of human behaviour in intertemporal social dilemmas, extending them to the complex dynamics of the real world is challenging due to the limitations of repeated matrix games. This highlights the need for more realistic environments that can capture the true nature of cooperative behaviour. There has been a trend of using realistic, videogame like, intertemporal social dilemmas in behavioural experiments to study intertemporal social dilemma (Leibo et al., 2017; Ostrom, 2009). Several studies on the governance of common-pool resources have constructed these environments to reflect the complexities of real-world social-ecological systems (Janssen, 2010, 2013; Janssen et al., 2010). These studies build more realistic environments by incorporating elements such

as spatial dynamics, temporal extensions, and resource regeneration, which more accurately represent the challenges faced in actual common-pool resource management. In these environments human subjects do not simply choose basic, discrete actions—such as "cooperate" or "defect"—that are atomic in matrix games, rather they must develop a cooperative or defective policy to carry out their strategic decisions, while simultaneously adapting to the changing environment created by other agents who are also learning and adjusting their strategies (Ostrom, 2009; Schlager et al., 1994; Wilson et al., 1994). In this research, we attempt model human cooperation with this novel approach to address the first gap in research by building two realistic, videogame-like environments based on existing libraries for training MARL agents.

This shift toward using these more realistic environments in behavioural experiments presents new challenges for modelling, revealing a second gap in current approaches: Existing agent-based models of cooperation in intertemporal social dilemmas often rely on techniques such as parameter sharing or reward decomposition to achieve stable cooperation (Yu et al., 2022). However, these methods can be somewhat artificial and fail to realistically reflect the complexity of human decision-making in such dilemmas. As a result theories from past behavioural experiment are often not directly applicable. This underscores the necessity for modelling techniques that can emulate human-like decision-making processes in these dynamic, temporally extended environments. Our approach seeks to integrate insights from behavioural game theory into agent-based modelling to bridge this second gap, addresses some of difficulties in agent-based modelling of human cooperation and hence, better understand the mechanisms of cooperation in complex social dilemmas.

Main Challenges in Agent Based Modelling of Human Group Behaviours

While reinforcement learning algorithms have been extensively used to model individual human decision-making in game-like environment (Botvinick et al., 2020; Rushworth & Walton, 2009), the adoption of MARL in modelling of human group behaviour has been relatively limited. Several previous studies have applied MARL and planning techniques to foster cooperation in these more realistic settings (Foerster et al., 2018; Leibo et al., 2017; Lerer & Peysakhovich, 2017; Pérolat et al., 2017). However, this approach has yet to achieve robust cooperation in games involving more than two players, a level of cooperation commonly observed in human behavioural experiments.

Stable multi-agent cooperation in social dilemmas is a challenging task due to non-stationarity and scalability issues (Leibo, et al., 2019). The presence of multiple agents interacting in a shared environment means that any change in one agent's policy can affect the experiences and learning trajectories of others, leading to a constantly shifting environment. For example, exploration in multi-agent settings requires a delicate balance, as agents must coordinate their strategies within a shared action space rather than independently adding noise to their actions as in single-agent scenarios (Leibo et al., 2019). This non-stationarity makes it hard for agents to learn stable strategies (Bard et al., 2020). Moreover, as the number of agents increases, the joint action space expands rapidly, leading to longer training times and computational complexities (Claus & Boutilier, 1998). While using independent agents without parameter sharing can partially address non-stationarity and are a closer analogy of human interation in real world, this approach often fails to capture crucial interdependencies between agents' actions, resulting in suboptimal performance (Foerster et al., 2017).

Another challenge with independent agents is related to credit assignment between agents: simply optimizing for group reward for tends to be ineffective due to the "lazy agent" problem, where some agents rely on others to contribute, leading to imbalanced efforts, ineffective coordination, and reduced overall cooperation (Sunehag et al., 2017). The partial observability inherent in realistic environments exacerbates this issue, making it difficult for agents to accurately interpret co-players' actions and their impact on shared rewards (Ndousse et al., 2021). Although recent advances in reward decomposition and the use of transformers have shown promise in addressing these challenges, much of this work remains limited to small-scale scenarios and lacks interpretability, particularly in context of modelling human cooperation.

Finding cooperative solutions to intertemporal social dilemmas is challenging for both natural and artificial agents because they must navigate the complexities of collective action and temporal credit assignment (Kelly, 2019). Temporal credit assignment is the key process of linking immediate actions with their long-term outcomes, this process is inherently variable and prone to errors, posing challenges for both human and reinforcement learning algorithms (Kearns & Singh, 2000). To avoid falling into socially deficient outcomes, individuals need to learn and coordinate group-level strategies that promote cooperation. Additionally, it's crucial for agents to recognise and link short-term defection with long-term negative consequences, ensuring that immediate self-interest does not jeopardize future cooperative efforts.

Insight from Inequity Aversion

To address these limitations in social science research and in engineering, we integrate insights from behavioural experiments. Studies in behavioural economics, psychology, sociology, and

political science have all highlighted the importance of fairness norms in resolving social dilemmas (Falk & Fischbacher, 2006; Frey & Bohnet, 1995; Hart, 1955; Johanson et al., 2022; Kelly, 2019; Klosko, 1987; Zheng et al., 2022). One key concept in understanding how cooperation emerges and persists, particularly in repeated matrix games, is inequity aversion (Fehr & Schmidt, 1999). This refers to individuals' preference to avoid unequal outcomes in their interactions, where they experience discomfort if they perceive inequity—whether they are at a disadvantage (disadvantageous inequity) or have an advantage (advantageous inequity) over others. These concepts can be seen as simplified representations of common social emotions: guilt, which corresponds to advantageous inequity aversion, and envy, which aligns with disadvantageous inequity aversion. (Camerer, 2003; Kollock, 1998). Inequity aversion has been extensively explored in behavioural game theory, notably by Fehr and Schmidt (1999), who developed a model that integrates fairness concerns into economic decision-making, and by Falk and Fischbacher (2006), who introduced reciprocity as a critical factor driving cooperative behaviour. The inequity aversion model has been effectively utilised to understand human behaviour in various laboratory economic games. These include the dictator game, the ultimatum game, the gift exchange game, public goods games, the trust game, and the market games (Eckel & Gintis, 2010; Greif, 2010). While these models have been foundational in the field, they are often simplified into binary choices, which oversimplify the complexities of human interactions (Leibo et al., 2017). Although they offer valuable insights, their applicability is limited, as they primarily generate predictions when problems are framed as matrix games. The simplicity of these paradigms, though useful for analysis, often fails to capture the full richness and variability of human cooperative behaviour (Dörner, 1990).

Our Approach

This limitation highlights the third significant gap in current literature: while inequity aversion models have successfully captured cooperative behaviour in social dilemmas using repeated matrix games, they have not yet been effectively integrated into models that address more realistic intertemporal social dilemmas. Our work bridges this gap by combining inequity aversion with independent multi-agent reinforcement learning, hence each agent is motivated not only by their own outcomes but also by concerns about fairness and equity (Camerer, 2003). By modelling cooperation and defection as dynamic, context-dependent behaviours, our approach seeks to capture the complexity of real-world social dilemmas and provide deeper insights into the mechanisms that foster sustainable cooperation in society (Gotts et al., 2003). This method allows us to explore how varying levels of inequity aversion influence the emergence of cooperative behaviour and how these behaviours evolve within different social and environmental contexts. We also propose inequity aversion as a solution to both the lazy agent problem and temporal credit assignment problem. By introducing fairness considerations, we address the lazy agent problem by discouraging free-riding and promoting balanced contributions among agents. Simultaneously, inequity aversion provides immediate feedback on perceived fairness, serving as a proxy for long-term outcomes and thus mitigating the temporal credit assignment problem. Integration of inequity aversion enable us to model more realistic and robust cooperative behaviours in intertemporal social dilemmas, potentially bridging the gap between theoretical models and observed human cooperation in complex, real-world scenarios.

Research Questions

In this paper we build two realistic intertemporal social dilemma and incorporate inequity aversion into agent-based modelling in intertemporal social dilemmas to explore how this concept influences cooperation over time. Our analysis focuses on two key aspects: advantageous and disadvantageous inequity aversion. We formulate 3 research questions: 1. How does inequity aversion influence collective action in intertemporal social dilemmas? 2. Can inequity aversion effectively address the temporal credit assignment problem in scenarios where short-term actions have long-term consequences? 3. What are the specific mechanisms through which inequity aversion alters the payoff structure and promotes cooperative behaviour among agents?

Hypothesis

We hypothesize that these mechanisms address the challenges of collective action and temporal credit assignment by influencing perceived payoffs and promoting cooperative behaviour through both direct and indirect pathways: Hypothesis 1 (direct pathway): Inequity aversion directly addresses the challenges of collective action by modifying the perceived payoffs associated with defection and cooperation. Specifically, agents who experience advantageous inequity aversion will perceive reduced marginal benefits from defection, thereby diminishing their incentive to defect. Hypothesis 2 (indirect pathway): Cooperating agents, motivated by disadvantageous inequity aversion, will be more likely to sanction defectors, reducing the overall incentives for free riding. This indirect mechanism is expected to discourage cooperative agents from switching to defection, even when exploring new strategies. Hypothesis 3: inequity aversion mitigates the temporal credit assignment problem by serving as the “early warning system” for agents. This system provides immediate negative feedback

for defectors through both direct and indirect pathways, aligning short-term rewards with long-term cooperative outcomes. In the direct pathway defect agents would receive an immediate negative reward due to advantageous inequity aversion. In the indirect pathway, defector would be immediately penalised by others due to disadvantageous inequity aversion. The combination of direct and indirect mechanisms will condition agents to prioritize group interests, even in the face of long-term uncertainties.

Potential Contributions

This research could potentially contribute to deepen our understanding of the theories of cooperative behaviour and potentially can be applied to build cooperative AI that understand human cooperation better through the integration of psychological concepts like inequity aversion into agent-based models (Zhang et al., 2021). Additionally, our work could potentially advance the study of cooperation theories by providing an agent-based modelling framework that can incorporates social theories such as inequity aversion (Kollock, 1998). This integration not only improves the realism of simulations but also bridges the gap between abstract theoretical models and complex, real-world social dynamics. Our framework also provided a platform for future research to model a wide range of social science theories beyond social dilemmas and inequity aversion. The potential applications of these findings are vast, spanning from policymaking in resource management to the development of more human-like AI systems. Ultimately, this work contributes to a deeper understanding of social science theories behind cooperation, paving the way for more effective strategies to address the collective challenges faced by both humans and AI in complex, interdependent settings.

Materials

Environment

According to Kollock (1998), multi-person social dilemmas can be broadly classified into two categories: 1. Public Good Dilemmas: Individuals incur personal cost to provide a resource that benefit everyone. 2. Common Dilemmas: Individuals are tempted by personal gain, which depletes a resource that is shared by all.

In this paper, we explore two specific dilemmas, one of each type. Both dilemma environments are implemented as partially observable Markov games on a 2D grid and are intertemporal in nature, meaning that selfish actions yield immediate rewards but have longer-term negative effects on the collective. Both environments are build based on the Melting Pot, a grid world game engine and the PettingZoo API, an library for MARL training and testing (Agapiou et al., 2023; Terry et al., 2021).

Public Goods Dilemma: The Cleanup Game

The Cleanup game is an example of a public goods dilemma. In this scenario, the main objective is for agents to collect apples from a field, where each collected apple grants a reward of 1 point. Apple spawning depends on an aquifer, located in a separate part of the map, which supplies water and nutrients to the apple field. However, this aquifer becomes increasingly polluted with waste over time, which in turn reduces the apple respawn rate. When waste levels are too high, apple spawning ceases entirely. Each game begins with waste levels just above this saturation point. To encourage apple spawning, agents must clean the waste from the

aquifer (Hughes et al., 2018). A screen shot from the cleanup environment is shown in Figure 1A.

The dilemma arises because while it is more rewarding for individual agents to stay in the apple field and collect apples, yet the long-term availability of apples hinges on some agents taking the less rewarding action of cleaning the aquifer. If all agents choose to defect and avoid cleaning, the collective reward diminishes to zero as no apples can respawn. Thus, a successful group need to navigate the tension between the temptation to free-ride and the need to contribute to the public good. Cooperative agents must make a commitment to the well-being of the group to ensure long-term success.

Commons Dilemma: The Harvest Game

The Harvest game exemplifies a commons dilemma. Here, the goal is also to collect apples, which is worth a reward of 1 each. The rate at which apples regrow varies across the map as they are affected by the spatial distribution of uncollected apples; regions with a higher concentration of uncollected apples experience faster regrowth. Conversely, if all apples in a particular region are harvested, none will grow back, leading to permanent resource depletion. The game runs for 1000 steps, after which it resets (Köster et al., 2020). A screen shot from the harvest environment is shown in Figure 1C.

The dilemma in the Harvest game lies in the conflict between individual and collective interests. Individually, players are incentivized to harvest apples as quickly as possible to maximize their short-term gain. However, if too many players harvest in the same area, the resource can be permanently depleted, reducing the overall long-term reward for the group. Cooperators must

resist the immediate temptation to harvest, sacrificing personal gain for the long-term benefit of the entire group.

Importance of Punishment

The availability of costly punishment plays a crucial role in human sequential social dilemmas. In both the Cleanup and Harvest games, the ability to punish defectors is included as an action available to the agents, reinforcing cooperative behaviour and deterring free-riding (Janssen et al., 2010).

Game Mechanics and Agent Capabilities

We setup both “clean up” and “harvest” environment in a way that addresses the limitation of repeated matrix games:

Agent Capabilities

In both the Harvest and Cleanup games, agents can move around and collect apples, agents have the ability to use the "fining beam," a tool that allows them to penalize other agents. When an agent uses the fining beam, they incur a cost of -1 reward, while the fined agent loses -50 rewards. Importantly, there is no penalty for using the fining beam unsuccessfully. Additionally, in the cleanup game, each agent has access to a "cleaning beam," which they can use to remove waste from the aquifer. collecting apples in both cleanup and harvest grants a reward of 1, and there are no other extrinsic rewards provided to the agents.

Game Dynamics

Cleanup Game: In the Cleanup game, waste is generated uniformly within the river at the probability of 0.5 at each timestep until the river reaches saturation, which occurs when waste covers 40% of the river. The probability of apples spawning in the field is dependent on the saturation level of the river; specifically, apples spawn with a probability of 0.125 times the current saturation level, x. Since the game starts with the river already saturated with waste, agents must contribute to cleaning the aquifer if they wish to receive any rewards from apple collection.

Harvest Game: In the Harvest game, the spawning of apples is influenced by the number of other apples within a Manhattan distance (l1 radius) of 2 units. The probabilities for apple spawning are 0, 0.005, 0.02, and 0.05 for areas with 0, 1, 2, and 3 or more apples within this radius, respectively. The initial arrangement of distribution of apples produces regions with varying degrees of connectivity and density. To implement sustainable harvesting policies, agents must focus on harvesting in denser regions while avoiding the removal of key apples that maintain the connectivity of these regions.

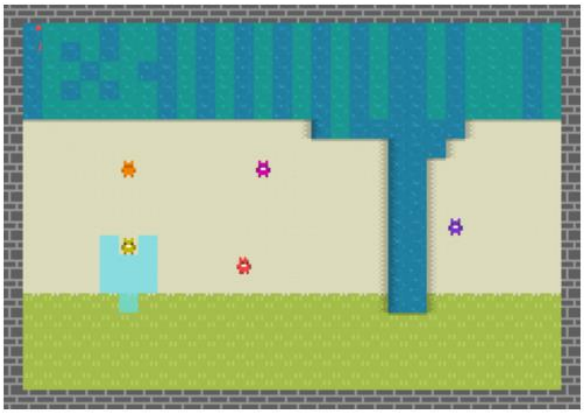
Strategic Implications

In both games, agents face the challenge of balancing immediate rewards from apple collection with the longer-term benefits of maintaining a sustainable environment. In the Cleanup game, this involves contributing to the public good by cleaning waste from the aquifer to ensure ongoing apple production. In the Harvest game, this requires careful harvesting to avoid depleting apple patches, ensuring that apple resources are not permanently exhausted. The use

of the fining beam adds an additional layer of strategy, as agents can deter selfish behaviour in others, but must weigh the cost of fining against potential long-term benefits for the group.

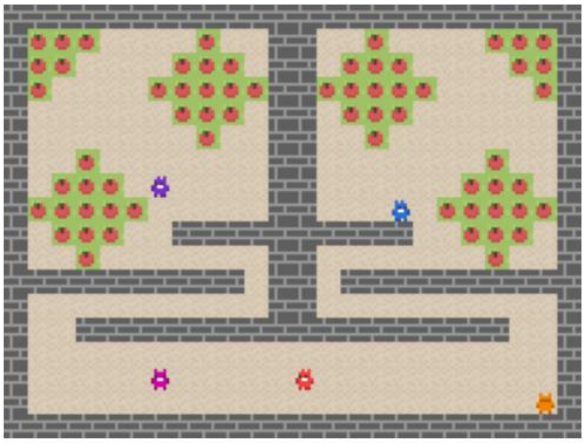
Figure 1

A

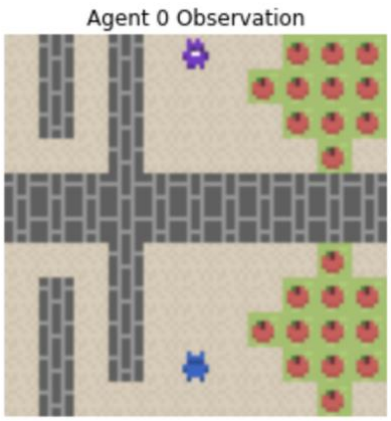


B

C



D



Note. A) a screen shot from the initial state of the cleanup game, the blue water body in the cleanup game indicates the aquifer, with the dark blue areas showing pollution, B) a screen shot from the initial state of the harvest game. In both games, the blue beam represents the cleaning beam, the yellow beam is used for fining, and the apples are shown in red. C. Agents observation in cleanup game; D. Agents observation in harvest game

The settings used in this study address the first four limitations commonly found in repeated matrix games. The remaining limitation of incomplete information is tackled by giving agents only a partially observable environment, specifically a 15x15 grid. From each agent's

perspective, this partial observability can lead to a "local decision process" that is non-Markovian. This is intentional and reflects a key feature of real-world environments that our model aims to capture, making the study more descriptively accurate. Examples of each agents' observation in both cleanup and harvest game is shown in Figure 1B and 1D respectively.

Formally define the environment

We aim to test how can MARL agents can model stable cooperation as humans in partially observable general-sum Markov games, as discussed in previous works. In these games, each agent operates based on a limited observation of the environment's state and receives individual rewards. The agents must independently develop strategies through ongoing interactions and learning from their experiences within the environment. The problem is structured as follows:

Let M denote an N-player partially observable Markov game defined on a finite state space S (N = 5 in our setup). $O: S \times \{1, \dots, N\} \to \mathbb{R}^d$ is the observation function that provides each player with a d-dimensional observation of the state (d = 15 × 15 × 3, observation space × RGB). From any state, agents select actions from their respective action spaces $A_1, \dots, A_N$. The state transitions based on the collective actions $\vec{a} = (a_1, \dots, a_n)$ according to a stochastic transition function $T: S \times A_1 \times \cdots \times A_N \to \Delta(S)$, where $\Delta(S)$ represents the set of probability distributions over S.

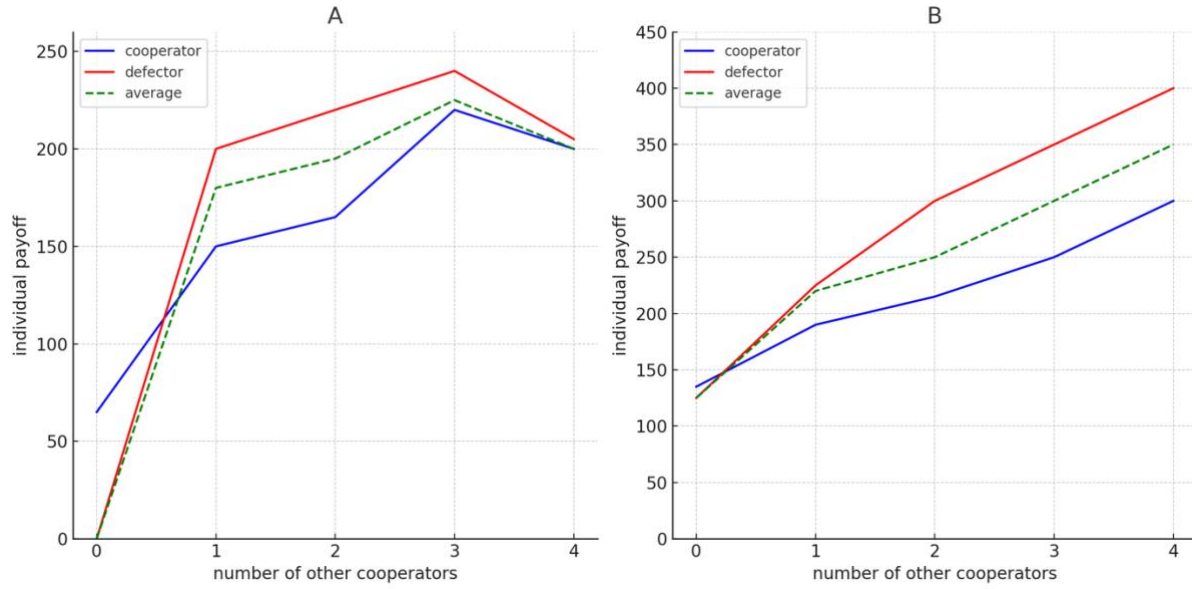
The observation space for each agent i is defined as $O_i = \{o_i \mid s \in S, o_i = O(s, i)\}$. Each agent receives an extrinsic reward determined by the function $r_i: S \times A_i \times \cdots \times A_N \to \mathbb{R}$. Each agent's objective is to independently learn a policy $\pi_i: O_i \to \Delta(A_i)$, where $\pi(a_i \mid o_i)$,

represents the probability of selecting action a_i given the observation, $o_i = O(s, i)$. For simplicity, joint actions, observations, and policies are denoted as $\vec{a} = (a_1, \ldots, a_N), \vec{o} = (o_1, \ldots, o_N)$ and $\vec{\pi}(\cdot|\, \vec{o}) = (\pi_1(\cdot|\, o_1), \ldots, \pi_N(\cdot|\, o_N))$, respectively. The objective for each agent is to maximize the expected long-term reward, expressed as:

$$V_i(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_i(s, \vec{a}) \mid \vec{a} \sim \vec{\pi}, s \sim T(s, \vec{a})],$$

Where γ being the discount factor that balances immediate and future rewards. We formally define the dynamic of intertemporal social dilemma in the appendix.

Figure 2



Note. The Cleanup and the Harvest games represent public good and common social dilemmas. A and B shows the Schelling diagram for Cleanup and Harvest game respectively. The dashed line in each diagram represents the average return if the individual opts to defect.

Validating the Environment

To demonstrate that these environments indeed function as social dilemmas, we plot Schelling diagrams as empirical evidence. Given the complexity and the spatial and temporal dimensions of these Markov games, it would be unfeasible to derive clear cooperating and defecting policies analytically. Therefore, we adopt an empirical approach to explore these dynamics. We use reinforcement learning to train agents on both cooperative and defecting strategies. To enforce these behaviours, we introduce specific modifications to the environment: In the Harvest game, cooperation is enforced by altering the environment so that certain agents are restricted from collecting apples in low-density areas, thereby encouraging more sustainable harvesting practices. In the Cleanup game, free-riding is encouraged by disabling the waste-cleaning ability for certain agents, which compels the remaining agents to undertake the cleanup tasks. Additionally, a small group reward signal is introduced to motivate cooperation among the remaining agents. The empirical Schelling diagrams derived from these experiments, as shown in Figure 2, confirm that these environments exhibit the characteristics of social dilemmas, validating the design and objectives of the study.

Learning Agents

In this study, we utilize the Proximal Policy Optimization (PPO) algorithm as the base model to train our agents(Yu et al., 2022). PPO is designed to optimize policies by adjusting them in a constrained manner, ensuring stable learning without significant oscillations. In our implementation, each agent operates independently, with no shared parameters between them, allowing for diverse strategies and behaviours to evolve naturally. PPO updates the policy using

a clipped surrogate objective, which prevents drastic changes during policy updates, maintaining stability and consistency. The objective function of PPO is defined as:

$$L^{CLIP}(\theta) = \mathbb{E}_t\left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right]$$

Where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)}$ is the probability ratio, the advantage function is defined as $\hat{A}_t$, which is calculated with k-step returns:

$$\hat{A}_t = A(s_t, a_t;\ \theta, \theta_v) = \sum_{i=0}^{k-1} \gamma^i u_{t+i} + \gamma^k V(s_{t+k}; \theta_v) - V(s_v; \theta_v)$$

Here, u_{t+i} represents the reward perceived by the agent which we will break down in detail in the following section. The policy is updated by maximizing the clipped objective, using the policy gradient:

$$\nabla_\theta L^{CLIP}(\theta) = \mathbb{E}_t\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\hat{A}_t\right]$$

This guides the direction and magnitude of policy parameter adjustments to ensure stable learning.

Integrating Inequity Aversion

We created advantageous inequity aversion agents and disadvantageous inequity agents by integrating inequity aversion into utility function of PPO agents. We begin with the utility function used in Fehr and Schmidt (1999), which is originally designed for static stateless games. We then extend this model to address sequential or multi-state problems, utilizing deep reinforcement learning to do so. The original form of the inequity aversion function is applicable in scenarios where $r_1, \ldots, r_N$, represent the extrinsic payoffs earned by each of the N players. The utility received by each agent i is defined as:

$$U_i(r_i, \ldots, r_N) = r_i - \alpha_i \sum_{j \neq i} max\,(r_j - r_i, 0) - \beta_j \sum_{j \neq i} max\,(r_i - r_j, 0)$$

Here, the terms involving α_i and β_i can be interpreted as intrinsic payoffs, The parameter α_i regulates the aversion to disadvantageous inequity of an agent, meaning a greater α_i leads to a larger utility loss when others achieve higher rewards. Conversely, β_i regulates the agent's aversion to advantageous inequity, representing the utility lost when the agent performs better than others. According to Fehr and Schmidt (1999) when $\alpha > \beta$ suggesting that individuals tend to be more sensitive to losses in social comparisons. In our experiments, we observed the strongest outcomes with $\alpha = 5$ and $\beta = 0.05$.

The concept of inequity aversion has primarily been modelled within the context of matrix games. The utility function in Fehr and Schmidt (1999) is directly applicable only in stateless game settings we extend the inequity aversion model to be suitable for temporally extended Markov games. The challenge in adapting the social preference model from Fehr and Schmidt (1999) to Markov games lies in the fact that different players may receive rewards at different timesteps. To address this, we introduce a method for smoothing the reward traces over time for each player. Let $r_i^t := r_i(s^t, a^t)$ represent the reward received by player *i* when taking action *a* in state *s* at time *t.* The subjective reward $u_i(s, a)$ that player *i* receives at state *s* from taking action *a* is defined as:

$$u_i(s_i^t, a_i^t) = r_i(s_i^t, a_i^t) - \frac{\alpha_i}{N-1} \sum_{j \neq i} max\,(e_j^t(s_j^t, a_j^t) - e_i^t(s_i^t, a_i^t), 0)$$

$$- \frac{\beta_i}{N-1} \sum_{j \neq i} max\,(e_j^t(s_i^t, a_i^t) - e_j^t(s_j^t, a_j^t), 0)$$

At each timestep *t*, the temporally smoothed rewards e_j^t for each agent *j* = 1, …, *N* are updated using the following equation:

$$e_j^t(s_j^t, a_j^t) = \gamma\lambda e_j^{t-1}(s_j^{t-1}, a_j^{t-1}) + r_j^t(s_j^t, a_j^t)$$

Here, γ represents the discount factor, and hyperparameter λ controls the smoothing. This method is conceptually similar to the use of eligibility traces (Sutton & Barto, 2018). Additionally, in our model, agents have access to the smoothed reward information of all players at each timestep. This reward function, which can be conceptualized as a combination of environmental and social rewards, opens up avenues for future research to explore the integration of other social theories beyond inequity aversion.

Metrics for Social Outcomes

In contrast to reinforcement learning in single agents, where the value function typically serves as the standard measure of agent performance, multi-agent systems with mixed incentives lack a single scalar metric that can fully capture the system's state (Hughes et al., 2018). Consequently, we employ and extend multiple social outcome metrics to summarize group behaviour and support comprehensive analysis.

Consider a system of *N* independent agents. For the *i*-th agent, let$\{r_t^i \mid t = 1, \dots, T\}$, represent the sequence of rewards accumulated over an episode of length *T*, and let $\{o_t^i \mid t = 1, \dots, T\}$ denote the corresponding observation sequence. The return for agent *i* is defined as $R^i = \sum_{t=1}^{T} r_t^i$.

We define three metrics, first the collective return (U), also referred as the utilitarian matric, as the average sum of rewards across all agents (Peysakhovich & Lerer, 2018). We use this metric to determine whether our agents have learned to cooperative in intertemporal social dilemma:

$$U = \mathbb{E}\left[\frac{1}{T}\sum_{i=1}^{N} R^i\right]$$

The distribution of rewards among agents is evaluated using an equality metric (E), quantified by an inverted Gini coefficient (Ceriani & Verme, 2012). A value closer to one indicates that rewards are more evenly distributed among agents.:

$$E = 1 - \frac{1}{2N}\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left|R^i - R^j\right|}{\sum_{i=1}^{N} R^i}$$

The Sustainability metric (S) reflects the average timing of reward collection. A lower mean indicates that apples were harvested more rapidly (Pérolat et al., 2017). This meteoric can be considered as an indirect measure of the indirect mechanism in hypothesis 2, leading to decreased sustainability. For agent *i,* it is calculated as:

$$S = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} t^i \,|r_t^i > 0\right]$$

Additionally, we introduce a measure of total contribution to the public good for the cleanup game, denoted as P, which is defined as the total number of wastes cleaned by all agents:

$$P = \sum_i p_i$$

We use the waste cleaned (P) as a measure of the direct mechanism in hypothesis 1

Experiments

In both the Cleanup and Harvest environments, we train five independent agents without any parameter sharing between them, to better emulate the complexity of human cooperation. This setup forces each agent to develop its own strategies for cooperation or defection, which evolve over time through actions such as moving, consuming apples, fining others, and cleaning waste. Each agent also needs to learn how its actions are perceived by others, leading to cooperative or competitive dynamics.

We use five baseline PPO agents with no inequity aversion to establish a benchmark. We then introduce agents with either advantageous ($\alpha = 0.05, \beta = 5$) or disadvantageous ($\alpha = 5, \beta = 0.05$) inequity aversion. The experiments include various combinations: one PPO with four advantageous agents, two PPO with three advantageous agents, and all five agents being advantageous. Similarly, for disadvantageous inequity aversion, we consider one PPO with four disadvantageous agents, two PPO with three disadvantageous agents, and all five agents being disadvantageous. This variety in agent configurations allows us to observe how different levels and types of inequity aversion influence cooperative behaviour in these environments.

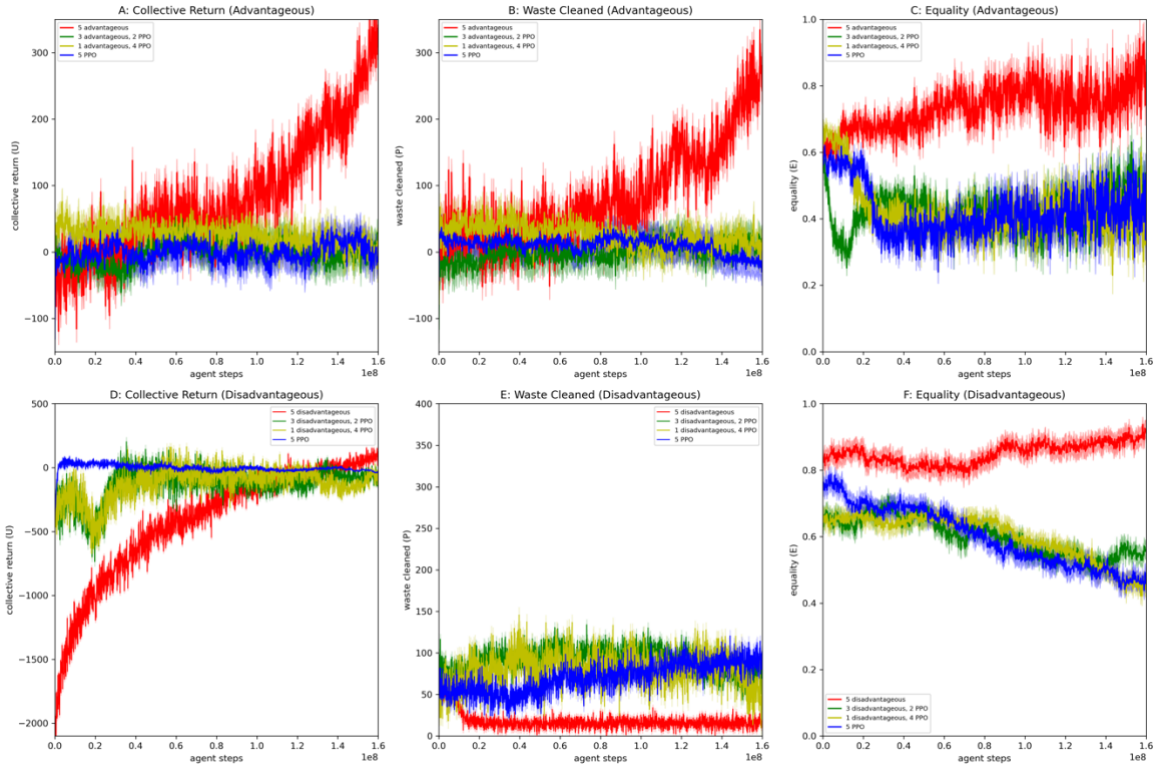
Ethics

The collected data is in adherence to the Research Ethics Policy of the London School of Economics.

Results

Our results demonstrate that advantageous inequity aversion can effectively address certain intertemporal social dilemmas by offering a well-timed intrinsic reward, thereby eliminating the need for punitive measures. This mechanism's success is contingent upon a sufficient proportion of the population being advantageous-inequity-averse. In contrast, even a small number of disadvantageous inequity averse agents can foster mutual cooperation by imposing timely penalties on defectors. Notably, advantageous inequity aversion excels in resolving public goods dilemmas (cleanup), while disadvantageous inequity aversion proves more effective in tackling commons dilemmas (harvest). In both scenarios, the baseline PPO agents were unable to consistently achieve socially beneficial outcomes. These findings suggest that different forms of inequity aversion can be strategically leveraged depending on the specific nature of the social dilemma, highlighting the nuanced role that these mechanisms play in facilitating cooperation.

Figure 3.

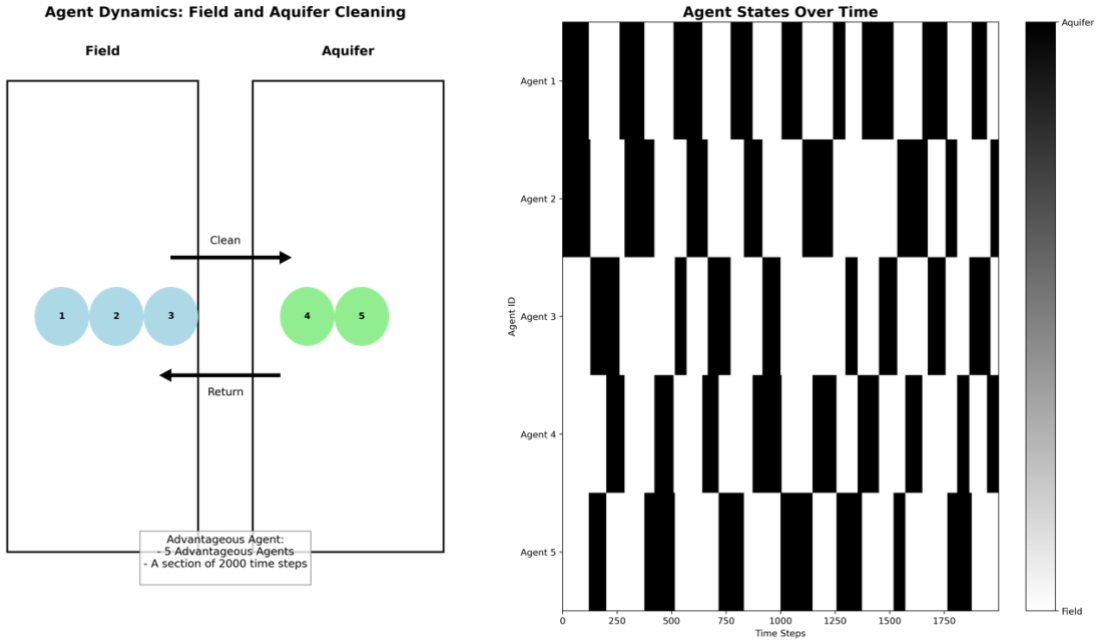


Note. In the Cleanup game, Advantageous inequity aversion promotes mutual cooperation. In (A), we compare the collective returns (metric U) of PPO agents with those of agents exhibiting advantageous inequity aversion, (B) illustrates their contributions to the public good (metric P), and (C) tracks the level of equality throughout the training process (metric E). In contrast, panels (D, E, and F) show that disadvantageous inequity aversion does not lead to enhanced cooperation between agents in the cleanup game.

Agents integrated advantageous inequity aversion outperform PPO in maintaining cooperative behaviour across both public goods (cleanup) and commons games (harvest), with the effect being especially notable in the cleanup game, exhibit higher collective returns (U), contribute more in waste cleaning(P), and achieve a greater level of equality (E) among agents. These findings support the direct pathway outlined in Hypothesis 1, which suggests that advantageous inequity aversion directly encourages cooperative behaviour by modifying the perceived payoffs, thus reducing the incentives for defection and promoting actions that

benefit the group as a whole (Figure 3). In this scenario, groups of five advantageous-inequity-averse agents are able to consistently cooperate and coordinate without explicit communication, with at least two agents consistently cleaning large amounts of waste, which leads to a significantly higher collective return (Figure 4, a video of this dynamic is available at the project's repo). This behaviour closely mirrors human cooperation, where individuals often take turns contributing to collective tasks, ensuring that the group benefits while no single individual bears the entire burden, which is a manifestation of guilt (advantageous inequity aversion). This turn-taking dynamic is a well-documented strategy in human social interactions, reflecting an evolved mechanism for sustaining cooperation and maximizing collective outcomes over time (Hamlin et al., 1995).

Figure 4.

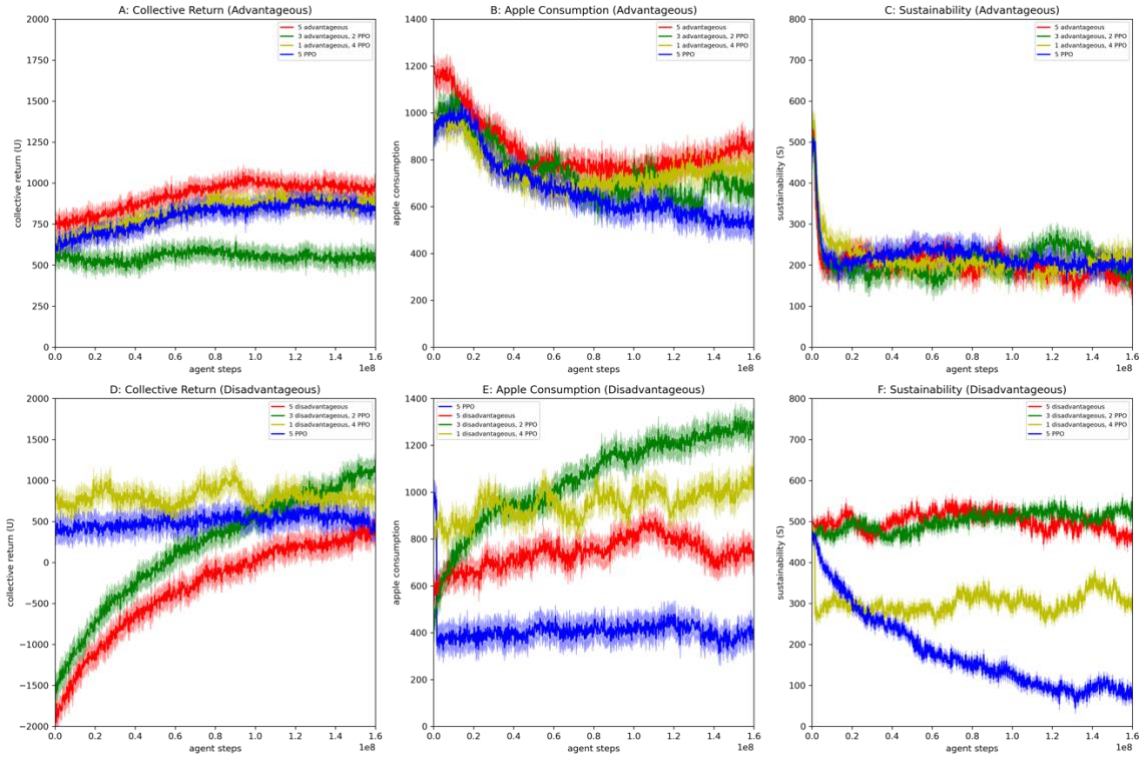


Note. The illustration on the right visually represents the turn-taking behaviour observed among advantageous agents in the cleanup game. For ease of understanding, we consider an agent to be engaged in cleaning waste when it moves to the aquifer and in collecting apples when it is in the field. On the left side, we have a 2000-step segment of the cleanup game, where each agent's position is plotted. Black represents the agent being at the aquifer (engaged in cleaning, shown as black), while white indicates that the agent is in the field (collecting apples, shown as white). This segment shows how the agents alternately take turns, ensuring that the aquifer is cleaned, and apples are collected.

This suggests that advantageous inequity aversion facilitates better temporal credit assignment: Turn-taking behaviour required agents to sacrifice short term return through collecting apples, which suggest agents could be rewarded in a way that encourages long-term cooperative behaviour. To further investigate this empirically, we introduced a delay in the intrinsic reward signal to verify the temporal credit assignment hypothesis (hypothesis 3) through the direct mechanism of inequity aversion. As shown in Figure 6, this delay

diminishes the effectiveness of advantageous inequity aversion, providing empirical evidence that the stable cooperation observed among agents with advantageous inequity aversion is indeed achieved through improved temporal credit assignment. This finding implies that when the reward signal is delayed, advantageous agents do not experience the immediate negative feedback associated with defection, thereby failing to link short-term defection with long-term negative consequences. This observation supports the hypothesis that inequity aversion facilitates more accurate temporal credit assignment through direct pathways, hence, reinforcing the connection between immediate actions and their delayed impacts on group outcomes.

Figure 5.

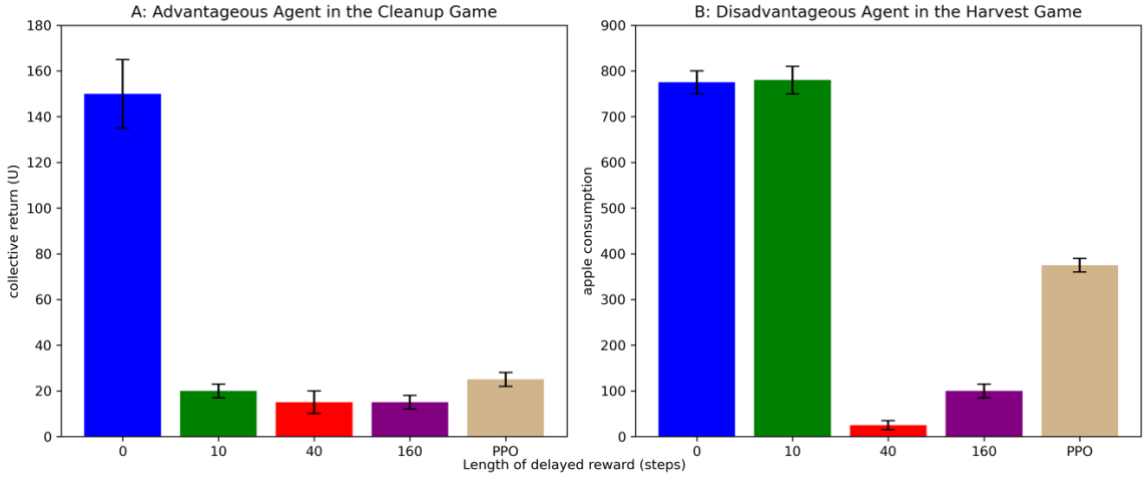


Note. In the harvest game, both advantageous and disadvantageous inequity aversion contributes to enhanced cooperation. When all five agents exhibit advantageous inequity aversion, we observed a slight improvement over baseline (PPO) across key social outcome metrics: collective return (U, shown in Figure 5A), apple consumption (Figure 5B), and sustainability (S, shown in Figure 5C). However, the impact of disadvantageous inequity aversion is more substantial, offering a significant improvement over baseline, even when only one out of five agents exhibit this trait. The results are reflected in collective return (U, Figure 5D), apple consumption (Figure 5E), and sustainability (S, Figure 5F).

Disadvantageous inequity averse agents exhibit superior performance compared to PPO baseline in maintaining cooperation through punitive actions in commons games (Figure 5). The results displayed indicate that agents exhibiting disadvantageous inequity aversion

achieve higher collective returns (U), better management of common-pool resources (as evidenced by measured apple consumption), and more sustainable strategies (S). These agents effectively employ punishment mechanisms to deter defection, validating the indirect pathway proposed in Hypothesis 2. Specifically, even a single disadvantageous inequity averse agent can effectively fine defectors, thereby promoting a cooperative and sustainable outcome. This supports the idea that disadvantageous inequity aversion can indirectly foster cooperation by imposing penalties on defectors, which aligns individual actions with the group's long-term interests. (Figure 5A). This behaviour closely mirrors human strategies in common dilemmas, as extensive research has shown that humans are often motivated to punish defectors (Fehr & Gächter, 2000; Yamagishi, 1986). We examine whether disadvantageous inequity aversion improve temporal credit assignment by adding gradual delays in reward signal for disadvantageous inequity averse agents in the harvest game, Figure 6 demonstrates that the punitive mechanism's effectiveness relies on the disadvantageous inequity aversion signal being closely timed with consumption of apples, enabling timely and effective policing as collective return diminishes with increasing delay in the reward signal. This indicates that inequity aversion can help align short-term incentives with the long-term impacts of over-consumption, thereby promoting cooperative behaviour. This supports the idea that indirect mechanisms can facilitate robust temporal credit assignment, as proposed in hypothesis 3.

Figure 6.



Note. Inequity aversion enhances cooperative behaviour by refining temporal credit assignment. (A) shows the collective return when there is a delay in implementing advantageous inequity aversion in the Cleanup game, while (B) displays the effects of delayed disadvantageous inequity aversion on apple consumption in the Harvest game.

However, the same advantage is not observed in the cleanup game, where disadvantageous inequity aversion does not yield the same positive impact on cooperation. The underlying reason for this divergence indicates that the effectiveness of inequity aversion may be context-dependent, particularly influenced by the specific dynamics of the social dilemma at hand. This context sensitivity suggests certain limitation of the design of our inequity aversion mechanisms: while they can be effective tools for modelling cooperation, their implementation must be carefully tailored to the characteristics of the environment in which they are applied.

Discussion

Our study explores how integrating inequity aversion into MARL models can replicate cooperation of humans in complex social dilemmas. By creating more realistic, dynamic environments such as the Cleanup and Harvest games, and modelling agents with either advantageous or disadvantageous inequity aversion, the research investigates the theorical mechanisms through which cooperative behaviour in humans can emerge and persist. The findings indicate that advantageous inequity aversion facilitates cooperation by directly modifying perceived payoffs and promoting collective action, while disadvantageous inequity aversion works indirectly by imposing penalties on defectors, thus aligning short-term actions with long-term group interests. We demonstrated that stable cooperation under inequity aversion is achieved through enhanced temporal credit assignment. By gradually introducing delays in the reward signal, we observed that the collective return (U), our metric for assessing cooperative performance, decreases as the delay increases.

The Cleanup game illustrates that advantageous inequity aversion provides an unambiguous and effective feedback signal for agents, encouraging them to contribute to the public good by cleaning up waste. This mechanism works directly by creating a straightforward link between cooperative behaviour and rewards: when agents clean the waste, they diminish the negative rewards associated with advantageous inequity aversion, as this action enables more apples for everyone, including themselves (Foerster et al., 2018). On the other hand, disadvantageous inequity aversion and punishment operate through an indirect mechanism that lacks this clear feedback property: While punishment can help agents explore new strategies by penalizing non-cooperative behaviour, it does not inherently make waste cleaning a more attractive option. As a result, the direct benefits of contributing to the public good are less apparent when relying

solely on disadvantageous inequity aversion and punishment. Our agents' behaviour closely mirrors human cooperation observed in real-life scenarios (Henrich et al., 2005). We demonstrate this by comparing our model's behaviour with human experimental results from previous studies where cooperation was tested in the cleanup game (McKee et al., 2021). In these human experiments, participants were able to cooperate effectively when other players were identifiable, and their contributions could be tracked. This reflects a similar dynamic in our model, where advantageous inequity aversion enables agents to make decisions that benefit the collective good, with the knowledge that their actions are observed and valued by others through advantageous utility. However, under conditions of anonymity, cooperation significantly declined in the human experiments. This situation is analogous to the baseline PPO scenario in our study. In both cases, the lack of visibility and traceability of actions disrupts the correct assignment of credit, leading to a breakdown in cooperative behaviour. Without the ability to accurately assign credit or blame, agents (or participants) struggle to maintain consistent cooperation, as there is no clear reinforcement of positive social behaviours (Hamlin et al., 1995). This comparison highlights how transparency and accountability are crucial for sustaining cooperation in both human and artificial agents.

In the Harvest game, agents must refrain from immediate consumption rather than actively contribute to a public good, which creates distinct challenges for encouraging cooperative behaviour. Advantageous inequity aversion in this context provides an inconsistent signal for promoting sustainability. This inconsistency arises because the agents' responses are highly sensitive to the specific and rapidly changing distribution of apples, making the feedback less reliable for guiding the exploration of effective policies (Dietz et al., 2003). Punishment, by contrast, serves as a more precise shaping mechanism in the Harvest game. It operates by discouraging overconsumption at critical times and locations, offering immediate and clear

feedback that aids agents in learning sustainable behaviours (Oliver, 1980). However, the cooperation facilitated by disadvantageous inequity aversion in the harvest game is inefficient, as it results in substantial resource losses due to the imposition of fines (Henrich et al., 2010). This inefficiency is evident when comparing the results across different agent configurations. Specifically, the setup with all five agents exhibiting disadvantageous inequity aversion results in notably lower apple consumption compared to configurations with a mix of disadvantageous agents and PPO agents, such as the 1 disadvantageous and 4 PPO combination or the 2 disadvantageous and 3 PPO combination. This suggests that while disadvantageous inequity aversion can promote cooperation, it does so at the cost of overall resource efficiency. This inefficiency parallels observations in human behaviour within laboratory matrix games of common pools resource management, where punitive measures, while promoting cooperation, lead to considerable resource wastage (Fehr & Gächter, 2000; Yamagishi, 1986). By contrast, in the cleanup game, agents with advantageous inequity aversion can effectively resolve the public good social dilemma without incurring the significant resource losses associated with punitive strategies. However, the success of this approach requires that a substantial portion of the population consists of fairness-oriented, advantageous inequity averse agents. This dependency mirrors the role of cultural norms in modulating fairness within human societies, where a critical threshold of fairness-minded individuals is necessary to sustain cooperation (Henrich et al., 2005). The hypothesis that fairness norms evolved as a mechanism to support continued human cooperation is consistent with these findings.

Interestingly, we observed turn-taking behaviour in waste cleanup when all five agents are set to exhibited advantageous inequity aversion a phenomenon that mirrors patterns seen in human experiments across various public goods games, including cleanup scenarios (Schmidt & Richardson, 2008). This finding is particularly notable because such behaviour had not been

documented in previous studies involving modelling using artificial agents without explicit training for turn taking. This turn-taking behaviour among agents can be attributed to the nature of advantageous inequity aversion (Blake et al., 2015). Agents motivated by this mechanism seek to reduce the negative utility associated with outperforming others. Consequently, agents take turns cleaning the waste to ensure that everyone benefits from the resulting increased apple production, thereby balancing their rewards and minimizing inequity, which resembles coordination without explicit communication between agents (Henrich et al., 2010). In this dynamic, an agent perceives cleaning as a means to boost the collective reward and, indirectly, its own reward by maintaining a favourable comparison with others. When one agent engages in cleanup, others are free to harvest apples, and the roles then switch, ensuring that no single agent continuously bears the cost of cleanup. This coordination strategy emerges as a stable solution to the public goods problem, reflecting similar behaviours observed in human groups where individuals alternate in performing essential tasks to sustain collective benefits (Herrmann et al., 2007). This suggests that advantageous inequity aversion not only promotes cooperation but also facilitates the emergence of sophisticated, human-like strategies such as turn-taking in complex social dilemmas.

Taking a broader perspective on our findings, we observe a distinct divergence in the effectiveness of advantageous and disadvantageous inequity aversion across different types of social dilemmas, a phenomenon that is also widely documented behavioural studies. In the Cleanup game (a public goods dilemma), advantageous inequity aversion emerged as the superior mechanism for fostering cooperation. This suggests that in scenarios where active contribution to a shared resource is required, guilt or discomfort from outperforming others serves as a more powerful motivator than the fear of being left behind. Conversely, in the Harvest game (a commons dilemma), disadvantageous inequity aversion proved more effective.

This indicates that when the challenge is to restrain consumption rather than actively contribute, the fear of others gaining an unfair advantage is a stronger driver of cooperative behaviour (Brosnan & De Waal, 2014). This asymmetry in the effectiveness of different forms of inequity aversion reveals a nuanced relationship between the nature of the social dilemma and the psychological mechanisms that best promote cooperation. It suggests that the optimal approach to fostering cooperation may depend critically on whether the dilemma requires positive action (contribution) or restraint (conservation). This finding challenges the notion of a one-size-fits-all approach to promoting cooperation and implies that successful interventions in real-world social dilemmas may need to be tailored to the specific type of collective action required (Herrmann et al., 2007). Furthermore, this asymmetry provides new insights into the potential evolutionary origins of these different forms of inequity aversion. It suggests that advantageous and disadvantageous inequity aversion may have evolved to solve different types of cooperation problems faced by our ancestors, rather than being general-purpose mechanisms for promoting fairness (Janssen & Rollins, 2012).

Building upon our findings, we can discern fundamental aspects of cooperation through our approach of integrating inequity aversion into multi-agent reinforcement learning. By modelling agents with advantageous and disadvantageous inequity aversion, we've demonstrated how these basic social preferences can engender complex cooperative behaviours, including turn-taking and sustainable resource management via punishment, without explicit training (Agapiou et al., 2023). This emergent cooperation closely mirrors patterns observed in human societies, suggesting that inequity aversion may be a core mechanism driving cooperative behaviour (Gächter & Herrmann, 2008, 2009). The success of these models in fostering cooperation across different types of social dilemmas - public goods and commons problems - indicates that sensitivity to fairness and equity could be a universal factor in

promoting collective action. However, the asymmetry in effectiveness between advantageous and disadvantageous inequity aversion in different contexts reveals a more nuanced picture. This suggests that these mechanisms may have evolved to address specific types of social challenges rather than serving as all-purpose cooperation facilitators (Williams & Moore, 2016). From an evolutionary perspective, these results support the hypothesis that inequity aversion may have evolved as a cognitive adaptation to facilitate cooperation in early human groups, with different forms of inequity aversion potentially evolving to solve distinct cooperation problems (Duéñez-Guzmán et al., 2023). This evolutionary foundation may explain why fairness considerations remain so deeply ingrained in human psychology and continue to play a crucial role in shaping cooperative behaviours in modern societies (Jaderberg et al., 2019).

Our findings offer insights for social science research on human cooperation by providing a computational and theorical framework that bridges evolutionary theory, behavioural economics, and social psychology. By demonstrating how simple fairness preferences can lead to sophisticated cooperative strategies, our work contributes to a more nuanced understanding of the theorical mechanisms underlying human social behaviour and collective action in complex, real-world scenarios. Moreover, it highlights the importance of considering the specific nature of social dilemmas when designing interventions or policies aimed at promoting cooperation, as different contexts may require different approaches to effectively leverage our innate sense of fairness and equity.

Conclusion

Our work contributes to the theoretical understanding of human cooperation by offering a simplistic model that achieved stable human-like cooperation in intertemporal social dilemmas and do so in a way that does not require explicitly training separate cooperating and defecting agents or modelling their specific behaviours. This makes our mechanism more scalable to complex environments and larger populations of agents. Additionally, our research provides valuable insights into how simple traits such as inequity aversion can allow stable cooperation in temporally extended social dilemmas to emerge and persist. Notably, we demonstrate that complex coordination, such as turn-taking behaviour, can naturally arise from inequity aversion, highlighting the potential for sophisticated social strategies to develop without explicit training. Through our experiments, we show that both advantageous and disadvantageous inequity aversion can facilitate cooperation by enabling accurate temporal credit assignment, thereby promoting behaviours that benefit the group over time.

Our research has wide applications in social science studies, particularly in enhancing models of human behaviour in economic and social contexts. By demonstrating that inequity aversion can naturally lead to cooperation and coordination among agents, our findings can be used to develop more accurate models for understanding collective bargaining, public goods provision, and resource allocation within communities (Zheng et al., 2022). These insights can inform policies aimed at promoting sustainable resource management, where collective action is crucial for managing common-pool resources like fisheries, forests, and water systems (Ndousse et al., 2021). Furthermore, we introduce a framework for building socially informed agents, where agents operate under a dual paradigm of environmental loss and social loss. This framework goes beyond inequity aversion and social dilemmas, allowing the integration of

other social functions such as reputation management, social pressure, and imitation (Kruppa et al., 2019). This flexibility enables the modelling of a wide range of social science problems, providing a more comprehensive approach to understanding and simulating complex social dynamics.

In addition to social science applications, our approach also has important implications for the development of cooperative AI and multi-agent systems. By integrating inequity aversion into the reward structures of agents, we can create systems that foster cooperation without the need for explicit communication or centralized control, which is particularly useful in robotics and distributed AI environments where autonomous agents must work together effectively (Dafoe et al., 2021). Moreover, our research contributes to human-computer interaction (HCI) by offering a framework for designing interfaces and systems that enhance cooperation among users or between users and machines (Ashktorab et al., 2020). By incorporating fairness and equity concerns, these systems can improve user engagement and foster collaborative behaviour, making them more effective in both individual and group settings (Schelble et al., 2021). Lastly, the principles of inequity aversion and its impact on cooperation can be applied in educational and training programs to teach cooperation and coordination skills, helping learners understand the importance of fairness in group dynamics and fostering leadership skills in team-based environments (Haiguang et al., 2020; McKee et al., 2023). Overall, our research offers a versatile foundation for advancing studies and applications across social sciences, AI, HCI, and education.

Our method does come with certain limitations. One key issue is that our agents rely on outcomes rather than predictions to guide their decision-making. While this approach works

within the confines of our model, it is unrealistic for accurately capturing human behaviour, especially in environments with high uncertainty or stochasticity, for example: farming and fishery. Humans often use predictions, heuristics, and anticipatory strategies to navigate complex, unpredictable situations, something our model does not account for (Charpentier et al., 2020; Rilling & Sanfey, 2011). This reliance on immediate outcomes means our agents may miss out on replicating more nuanced human behaviours, such as planning for future contingencies, adapting to changing environments, or learning from the probabilistic nature of real-world interactions (Sanfey, 2007). Furthermore, by not incorporating elements like social influence, peer pressure, or the ability to shift between advantageous and disadvantageous inequity aversion depending on context, our model may overlook critical factors that drive human cooperation and decision-making (Cialdini & Goldstein, 2004). These gaps highlight directions of more sophisticated models in the future that better reflect the complexities of human thought processes, social dynamics, and the spatial contexts within which these interactions occur.

Future research could build on our model by applying the socially informed agent framework beyond inequity aversion to other social dynamics, such as reputation management, social norms enforcement, and peer influence. For example, incorporating reputation systems could allow agents to make decisions based on their perceived standing within a group, encouraging long-term cooperative behaviour (Niu et al., 2021). Additionally, researchers could replace one or multiple agents with human players to see if the agent can learn stable cooperative strategies from human interactions, providing insights into how well these agents can integrate into human-centric environments (McKee et al., 2021).

Another avenue for future exploration involves developing agents that make decisions based on predictions of future behaviour, such as anticipating whether others will cooperate or defect, rather than relying solely on past rewards. This approach could better reflect human decision-making processes and align with research in behavioural economics, cognitive science, and neuroscience (Sajid et al., 2021). Incorporating spatial relationships into these models could also provide a more comprehensive understanding of how physical proximity and resource distribution influence cooperation (Koleff et al., 2003; Littman, 1994). There is also a growing trend in using large language models (LLMs) to replicate social science research, as these models often exhibit biases similar to those observed in humans (Bakhtin et al., 2022). The expanding community of researchers using LLMs as agents in multi-agent settings offers a unique opportunity to study cooperation among LLM agents, between LLMs and human agents, and to determine whether humans can distinguish between interactions with LLMs and other humans (Yocum et al., 2023).

In terms of analysis, our research provided a suite of social outcome metrics, such as collective return, sustainability, and equality, focusing primarily on the temporal aspects rather than spatial dynamics (Ceriani & Verme, 2012; Pérolat et al., 2017; Peysakhovich et al., 2014). future research could develop metrics specifically for spatial analysis. These metrics could include measures of how agents' spatial positioning affects their access to resources, their interactions with other agents, and their overall contributions to collective tasks. For example, spatial metrics could assess the density of agents in key resource areas, the distance between cooperative clusters, or the impact of spatial arrangement on the effectiveness of coordinated actions (Koleff et al., 2003). With these spatial metrics future studies could delve deeper into turn-taking coordination problems by examining how teams of "cleaners" and "collectors" are formed (Jaques et al., 2019). Researchers could also explore building explicit communication

channels for agents, akin to a chat system in a video game, and test the interactions of new agents, LLM agents, and human players within this environment. Overall, these directions offer possibilities for advancing our understanding of cooperation, coordination, and decision-making in both artificial and human systems.

Data and Code Availability:

The simulation data, figures and code to reproduce the experiment are available at: https://github.com/edluyuan/LSE_capstone_project

References

Acheson, J. M. (1981). Anthropology of fishing. *Annual Review of Anthropology. Volume 10*. https://doi.org/10.1146/annurev.an.10.100181.001423

Agapiou, J. P., Sasha Vezhnevets, A., Duéñez-Guzmán, E. A., Matyas, J., Mao, Y., Sunehag, P., Köster, R., Madhushani, U., Kopparapu, K., Comanescu, R., Strouse, D., Johanson, M. B., Singh, S., Haas, J., Mordatch, I., Mobbs, D., Leibo, J. Z., & contributions, E. (2023). Melting Pot 2.0. *Arxiv.Org*. https://arxiv.org/abs/2211.13746

Ashktorab, Z., Liao, Q. V., Dugan, C., Johnson, J., Pan, Q., Zhang, W., Kumaravel, S., & Campbell, M. (2020). Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2). https://doi.org/10.1145/3415167

Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2003). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, *32*(1). https://doi.org/10.1137/S0097539701398375

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review*, *80*(4). https://doi.org/10.1017/S0003055400185016

Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., Roller, S., … Zijlstra, M. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, *378*(6624). https://doi.org/10.1126/science.ade9097

Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., Dunning, I., Mourad, S., Larochelle, H.,

Bellemare, M. G., & Bowling, M. (2020). The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, *280*. https://doi.org/10.1016/j.artint.2019.103216

Besley, T., & Coate, S. (2003). Centralized versus decentralized provision of local public goods: A political economy approach. *Journal of Public Economics*, *87*(12). https://doi.org/10.1016/S0047-2727(02)00141-X

Blake, P. R., McAuliffe, K., Corbit, J., Callaghan, T. C., Barry, O., Bowie, A., Kleutsch, L., Kramer, K. L., Ross, E., Vongsachang, H., Wrangham, R., & Warneken, F. (2015). The ontogeny of fairness in seven societies. *Nature*, *528*(7581). https://doi.org/10.1038/nature15703

Bloembergen, D., Tuyls, K., Hennes, D., & Kaisers, M. (2015). Evolutionary dynamics of multi-agent learning: A survey. In *Journal of Artificial Intelligence Research* (Vol. 53). https://doi.org/10.1613/jair.4818

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep Reinforcement Learning and Its Neuroscientific Implications. In *Neuron* (Vol. 107, Issue 4). https://doi.org/10.1016/j.neuron.2020.06.014

Brosnan, S. F., & De Waal, F. B. M. (2014). Evolution of responses to (un)fairness. In *Science* (Vol. 346, Issue 6207). https://doi.org/10.1126/science.1251776

Camerer, C. F. (2003). Behavioral game theory: Experiments in strategic interaction. In *Behavioral Game Theory: Experiments in Strategic Interaction*. https://doi.org/10.1016/j.socec.2003.10.009

Ceriani, L., & Verme, P. (2012). The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality*, *10*(3). https://doi.org/10.1007/s10888-011-9188-x

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, *117*(3). https://doi.org/10.1162/003355302760193904

Charpentier, C. J., Iigaya, K., & O'Doherty, J. P. (2020). A Neuro-computational Account of Arbitration between Choice Imitation and Goal Emulation during Human Observational Learning. *Neuron*, *106*(4), 687-699.e7. https://doi.org/10.1016/J.NEURON.2020.02.028

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*. https://doi.org/10.1146/annurev.psych.55.090902.142015

Claus, C., & Boutilier, C. (1998). Dynamics of reinforcement learning in cooperative multiagent systems. *Proceedings of the National Conference on Artificial Intelligence*.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021). Cooperative AI: machines must learn to find common ground. In *Nature* (Vol. 593, Issue 7857). https://doi.org/10.1038/d41586-021-01170-0

De Cote, E. M., Lazaric, A., Restelli, M., & Bonarini, A. (2006). Learning to cooperate in multi-agent social dilemmas. *Proceedings of the International Conference on Autonomous Agents*, *2006*. https://doi.org/10.1145/1160633.1160770

Dietz, T., Ostrom, E., & Stern, P. C. (2003). The Struggle to Govern the Commons. In *Science* (Vol. 302, Issue 5652). https://doi.org/10.1126/science.1091015

Dörner, D. (1990). The logic of failure. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *327*(1241). https://doi.org/10.1098/rstb.1990.0089

Duéñez-Guzmán, E. A., Sadedin, S., Wang, J. X., McKee, K. R., & Leibo, J. Z. (2023). A social path to human-like artificial intelligence. *Nature Machine Intelligence*, *5*(11).

https://doi.org/10.1038/s42256-023-00754-x

Eckel, C., & Gintis, H. (2010). Blaming the messenger: Notes on the current state of experimental economics. *Journal of Economic Behavior and Organization*, *73*(1). https://doi.org/10.1016/j.jebo.2009.03.026

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, *54*(2). https://doi.org/10.1016/j.geb.2005.03.001

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4). https://doi.org/10.1257/aer.90.4.980

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation on JSTOR. *The Quarterly Journal of Economics, 14*(3), 817–868. https://www.jstor.org/stable/2586885

Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., & Mordatch, I. (2018). Learning with opponent-learning awareness. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *1*.

Foerster, J., Nardell, N., Farquhar, G., Afouras, T., Torr, P. H. S., Kohli, P., & Whiteson, S. (2017). Stabilising experience replay for deep multi-agent reinforcement learning. *34th International Conference on Machine Learning, ICML 2017*, *3*.

Frey, B. S., & Bohnet, I. (1995). Institutions affect fairness: Experimental investigations. *Journal of Institutional and Theoretical Economics*, *151*(2).

Gächter, S., & Herrmann, B. (2008). Reciprocity, Culture and Human Cooperation: Previous Insights and a New Cross-Cultural Experiment about the Centre or contact. *Experimental Economics*, *364*(1518).

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: Previous

insights and a new cross-cultural experiment. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 364, Issue 1518). https://doi.org/10.1098/rstb.2008.0275

Greif, A. (2010). A Primer in Game Theory. In *Institutions and the Path to the Modern Economy*. https://doi.org/10.1017/cbo9780511791307.017

Haiguang, F., Shichong, W., Shushu, X., & Xianli, W. (2020). *Research on Human-Computer Cooperative Teaching Supported by Artificial Intelligence Robot Assistant*. https://doi.org/10.1007/978-3-030-41099-5_3

Hamlin, A., Ostrom, E., Gardner, R., & Walter, J. (1995). Rules, Games, and Common-Pool Resources. *The Economic Journal*, *105*(431). https://doi.org/10.2307/2235179

Hart, H. L. A. (1955). Are There Any Natural Rights? *The Philosophical Review*, *64*(2). https://doi.org/10.2307/2182586

Henrich, J., Boyd, R., Bowles, S., Camerer, C. F., Fehr, E., & Gintis, H. (2005). Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies. In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. https://doi.org/10.1093/0199262055.001.0001

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, *327*(5972). https://doi.org/10.1126/science.1182238

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence

hypothesis. *Science*, *317*(5843). https://doi.org/10.1126/science.1146282

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., Castañeda, A. G., Dunning, I., Zhu, T., McKee, K., Koster, R., Roff, H., & Graepel, T. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in Neural Information Processing Systems*, *2018-December*.

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castañeda, A. G., Beattie, C., Rabinowitz, N. C., Morcos, A. S., Ruderman, A., Sonnerat, N., Green, T., Deason, L., Leibo, J. Z., Silver, D., Hassabis, D., Kavukcuoglu, K., & Graepel, T. (2019). Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, *364*(6443). https://doi.org/10.1126/science.aau6249

Janssen, M. A. (2010). Introducing ecological dynamics into common-pool resource experiments. *Ecology and Society*, *15*(2). https://doi.org/10.5751/ES-03296-150207

Janssen, M. A. (2013). The role of information in governing the commons: Experimental results. *Ecology and Society*, *18*(4). https://doi.org/10.5751/ES-05664-180404

Janssen, M. A., Holahan, R., Lee, A., & Ostrom, E. (2010). Lab experiments for the study of social-ecological systems. *Science*, *328*(5978). https://doi.org/10.1126/science.1183532

Janssen, M. A., & Rollins, N. D. (2012). Evolution of cooperation in asymmetric commons dilemmas. *Journal of Economic Behavior and Organization*, *81*(1). https://doi.org/10.1016/j.jebo.2011.10.010

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P. A., Strouse, D. J., Leibo, J. Z., & de Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. *36th International Conference on Machine Learning, ICML 2019*, *2019-June*.

Johanson, M., Hughes, E., … F. T. preprint arXiv, & 2022, U. (2022). Emergent bartering behaviour in multi-agent reinforcement learning. *Arxiv.Org*, 2022–2027. https://arxiv.org/abs/2205.06760

Kearns, M. J., & Singh, S. P. (2000). Bias-Variance Error Bounds for Temporal Difference Updates. *Proceedings of the 13th Annual Conference on Computational Learning Theory*.

Kelly, C. (2019). Discourse on the Origin of Inequality. In *The Rousseauian Mind*. https://doi.org/10.4324/9780429020773-16

Klosko, G. (1987). The Principle of Fairness and Political Obligation. *Ethics*, *97*(2). https://doi.org/10.1086/292843

Koleff, P., Gaston, K. J., & Lennon, J. J. (2003). Measuring beta diversity for presence–absence data. *Journal of Animal Ecology*, *72*(3), 367–382. https://doi.org/10.1046/J.1365-2656.2003.00710.X

Kollock, P. (1998). Social dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology*, *24*. https://doi.org/10.1146/annurev.soc.24.1.183

Köster, R., Hadfield-Menell, D., Hadfield, G. K., & Leibo, J. Z. (2020). Silly rules improve the capacity of agents to learn stable enforcement and compliance behaviors. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *2020-May*.

Kruppa, J. A., Gossen, A., Oberwelland Weiß, E., Kohls, G., Großheinrich, N., Cholemkery, H., Freitag, C. M., Karges, W., Wölfle, E., Sinzig, J., Fink, G. R., Herpertz-Dahlmann, B., Konrad, K., & Schulte-Rüther, M. (2019). Neural modulation of social reinforcement learning by intranasal oxytocin in male adults with high-functioning autism spectrum

disorder: a randomized trial. *Neuropsychopharmacology*, *44*(4). https://doi.org/10.1038/s41386-018-0258-7

Leibo, J. Z., Hughes, E., Lanctot, M., & Graepel, T. (2019). *Autocurricula and the Emergence of Innovation from Social Interaction: A Manifesto for Multi-Agent Intelligence Research*. http://arxiv.org/abs/1903.00742

Leibo, J. Z., Perolat, J., Hughes, E., Wheelwright, S., Marblestone, A. H., Duéñez-Guzman, E., Sunehag, P., Dunning, I., & Graepel, T. (2019). Malthusian reinforcement learning. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *2*.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *1*.

Lerer, A., & Peysakhovich, A. (2017). *Maintaining cooperation in complex social dilemmas using deep reinforcement learning*. http://arxiv.org/abs/1707.01068

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. *Proceedings of the 11th International Conference on Machine Learning, ICML 1994*. https://doi.org/10.1016/B978-1-55860-335-6.50027-1

Macy, M. W., & Flache, A. (2002). Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(SUPPL. 3). https://doi.org/10.1073/pnas.092080099

McKee, K. R., Hughes, E., Zhu, T. O., Chadwick, M. J., Koster, R., Castaneda, A. G., Beattie, C., Graepel, T., Botvinick, M., & Leibo, J. Z. (2021). *A multi-agent reinforcement learning model of reputation and cooperation in human groups*.

http://arxiv.org/abs/2103.04982

McKee, K. R., Tacchetti, A., Bakker, M. A., Balaguer, J., Campbell-Gillingham, L., Everett, R., & Botvinick, M. (2023). Scaffolding cooperation in human groups with deep reinforcement learning. *Nature Human Behaviour*, *7*(10). https://doi.org/10.1038/s41562-023-01686-7

Ndousse, K., Eck, D., Levine, S., & Jaques, N. (2021). Emergent Social Learning via Multi-agent Reinforcement Learning. *Proceedings of Machine Learning Research*, *139*.

Niu, Y., Paleja, R., & Gombolay, M. (2021). Multi-agent graph-attention communication and teaming. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *2*.

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*(6685). https://doi.org/10.1038/31225

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, *364*(6432). https://doi.org/10.1038/364056a0

Nowak, Martin A., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, *359*(6398). https://doi.org/10.1038/359826a0

Nowak, Martin A., & Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature*, *355*(6357). https://doi.org/10.1038/355250a0

Ohtsuki, H., Hauert, C., Lieberman, E., & Nowak, M. A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, *441*(7092). https://doi.org/10.1038/nature04605

Oliver, P. (1980). Rewards and Punishments as Selective Incentives for Collective Action:

Theoretical Investigations. *American Journal of Sociology*, *85*(6). https://doi.org/10.1086/227168

Olson, M. (1965). The logic of collective action. In *Public goods and the theory of groups*.

Ostrom, E. (2009). A general framework for analyzing sustainability of social-ecological systems. In *Science* (Vol. 325, Issue 5939). https://doi.org/10.1126/science.1172133

Pérolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., & Graepel, T. (2017). A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in Neural Information Processing Systems*, *30*. https://youtu.

Peysakhovich, A., & Lerer, A. (2018). Prosocial learning agents solve generalized stag hunts better than selfish ones. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *3*.

Peysakhovich, A., Nowak, M. A., & Rand, D. G. (2014). Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nature Communications 2014 5:1*, *5*(1), 1–8. https://doi.org/10.1038/ncomms5939

Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*, *62*. https://doi.org/10.1146/annurev.psych.121208.131647

Rushworth, M. F. S., & Walton, M. E. (2009). Neuroeconomics: Decision Making and the Brain. *Neuron*, *63*(2). https://doi.org/10.1016/j.neuron.2009.07.005

Sajid, N., Ball, P. J., Parr, T., & Friston, K. J. (2021). Active inference: demystified and compared. In *Neural Computation* (Vol. 33, Issue 3). https://doi.org/10.1162/neco_a_01357

Sandholm, T. W., & Crites, R. H. (1996). Multiagent reinforcement learning in the Iterated Prisoner's Dilemma. *BioSystems*, *37*(1–2). https://doi.org/10.1016/0303-2647(95)01551-

Sanfey, A. G. (2007). Social decision-making: Insights from game theory and neuroscience. In *Science* (Vol. 318, Issue 5850). https://doi.org/10.1126/science.1142996

Schelble, B. G., Flathmann, C., McNeese, N., & Canonico, L. B. (2021). Understanding human-AI cooperation through game-theory and reinforcement learning models. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2020-January*. https://doi.org/10.24251/hicss.2021.041

Schlager, E., Blomquist, W., & Shui Yan Tang. (1994). Mobile flows, storage, and self-organized institutions for governing common-pool resources. *Land Economics*, *70*(3). https://doi.org/10.2307/3146531

Schmidt, R. C., & Richardson, M. J. (2008). Dynamics of interpersonal coordination. *Understanding Complex Systems*, *2008*. https://doi.org/10.1007/978-3-540-74479-5_14

Sunehag, P., Czarnecki, W. M., Lanctot, M., Lever, G., Zambaldi, V., Sonnerat, N., Gruslys, A., Jaderberg, M., Leibo, J. Z., Tuyls, K., & Graepel, T. (2017). Value-Decomposition networks for cooperative multi-agent learning. In *arXiv*.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An Introduction. In *Learning* (Vol. 3, Issue 9).

Terry, J. K., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L., Perez, R., Horsch, C., Dieffendahl, C., Williams, N. L., Lokesh, Y., & Ravi, P. (2021). PettingZoo: A Standard API for Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems*, *18*.

Wedekind, C., & Milinski, M. (2000). Cooperation Through Image Scoring in Humans. *Science*, *288*(5467), 850–852. https://doi.org/10.1126/SCIENCE.288.5467.850

Williams, A., & Moore, C. (2016). A longitudinal exploration of advantageous and disadvantageous inequality aversion in children. *Journal of Experimental Child Psychology*, *152*. https://doi.org/10.1016/j.jecp.2016.07.006

Wilson, J. A., Acheson, J. M., Metcalfe, M., & Kleban, P. (1994). Chaos, complexity and community management of fisheries. *Marine Policy*, *18*(4). https://doi.org/10.1016/0308-597X(94)90044-2

Wu, J., Balliet, D., & Van Lange, P. A. M. (2016). Reputation, Gossip, and Human Cooperation. In *Social and Personality Psychology Compass* (Vol. 10, Issue 6). https://doi.org/10.1111/spc3.12255

Wunder, M., Littman, M., & Babes, M. (2010). Classes of multiagent Q-learning dynamics with ε-greedy exploration. *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*.

Yamagishi, T. (1986). The Provision of a Sanctioning System as a Public Good. *Journal of Personality and Social Psychology*, *51*(1). https://doi.org/10.1037/0022-3514.51.1.110

Yocum, J., Christoffersen, P., Damani, M., Svegliato, J., Hadfield-Menell, D., & Russell, S. (2023). Mitigating Generative Agent Social Dilemmas. *NeurIPS 2023 Foundation Models for Decision Making Workshop, 2023•openreview.Net*. https://openreview.net/forum?id=5TIdOk7XQ6

Yoelia, E., Hoffmanbc, M., Randcd, D. G., & Nowak, M. A. (2013). Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(SUPPL2). https://doi.org/10.1073/pnas.1301210110

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The

Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. *Advances in Neural Information Processing Systems*, *35*.

Yu, C., Zhang, M., Ren, F., & Tan, G. (2015). Emotional multiagent reinforcement learning in spatial social dilemmas. *IEEE Transactions on Neural Networks and Learning Systems*, *26*(12). https://doi.org/10.1109/TNNLS.2015.2403394

Zhang, K., Yang, Z., & Başar, T. (2021). Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. In *Studies in Systems, Decision and Control* (Vol. 325). https://doi.org/10.1007/978-3-030-60990-0_12

Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., & Socher, R. (2022). The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances*, *8*(18). https://doi.org/10.1126/sciadv.abk2607

Appendix

Formally definition of intertemporal social dilemma

An Intertemporal Social Dilemma is a type of sequential social dilemma where the optimal strategy for an individual player in the short term involves defection, but such behaviour is suboptimal in the long term when considering the collective outcome. We formally define intertemporal social dilemmas as follows:

Consider a Markov game *M* with *N* players, each player *i* has a policy π_i belonging to either the set of cooperating policies Π_c or defecting policies Π_d. The state space is denoted by *S* and the action space is denoted by *A.* A strategy profile $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$ *i*s a collection of policies for all players, where each $\pi_\iota \in \Pi_c \cup \Pi_d$. Let $R_c(l)$ and $R_d(l)$ denote the average payoff for cooperating and defective policies when there are *l* cooperating players\

Intertemporal Tension:

Let π_{ki} denote the policy that maximizes player *i*'s expected return over the next *k* steps, starting from the initial state $s_0 \in S$. Formally in a short time horizon *k,* the optimal policy π_{ki} maximizes:

$$\pi_{ki} = arg \max_{\pi_i \in \Pi_c \cup \Pi_d} \mathbb{E}\left[\sum_{t=0}^{k} r_i(s_t, \pi(s_t)) \mid s_0\right]$$

Where $r_i(s_t, \pi(s_t))$ is the reward for player *i* at time *t* under the policy profile π.

Markov game *M* with the policy sets Π_c and Π_d forms an Intertemporal Sequential Social Dilemma if the following conditions hold:

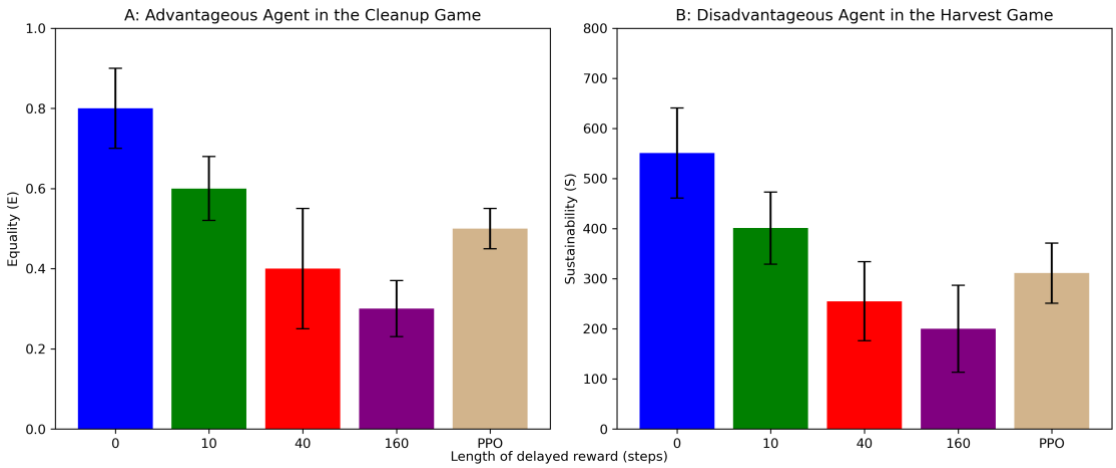
Short-Term Optimality of Defection: For sufficiently small k the policy π_{ki} that maximizes short-term return is a defecting policy, i.e. $\pi_{ki} \in \Pi_d$

Long-Term Suboptimality of Defection: If all players adopt defecting policies in the long term, the collective outcome is worse than if all had cooperated. Formally, for large k:

$$\mathbb{E}\left[\sum_{t=0}^{T} R_c(N)\right] > \mathbb{E}\left[\sum_{t=0}^{T} R_d(0)\right]$$

Where T represent a long-term horizon.

Supplementary Figure 1



Note. Illustration of how equality (A) and sustainability (B) metric diminishes when a delay in reward is induced in cleanup and harvest respectively