Navigating the N-Person Prisoner's Dilemma: From the Tragedy Valley to the Reciprocity Hill with Adaptive Learning Agents

Chris Tcaci Chris Huyck[1]

[1] Middlesex University, London NW4 4BT UK
M00674787@mdx.ac.uk
[2] c.huyck@mdx.ac.uk
https://cwa.mdx.ac.uk/chris/chrisroot.html

Abstract. The N-Person Iterated Prisoner's Dilemma (N-IPD) poses a significant challenge to the emergence of cooperation due to diffused responsibility and obscured reciprocity. This paper investigates how agent-based learning models navigate this complex social dilemma. We demonstrate that simple reinforcement learning agents consistently fall into a "Tragedy Valley" of mutual defection in standard N-IPD neighbourhood interaction models. However, by enhancing agents with contextual awareness of their local environment and employing adaptive Multi-Agent Reinforcement Learning (MARL) algorithms like Hysteretic-Q and Wolf-PHC, high levels of sustained cooperation (over 85%) can be achieved. Furthermore, we explore the fundamental impact of interaction structure, contrasting the neighbourhood model with a pairwise interaction model where agents play repeated 2-player games. The pairwise model, by enabling direct reciprocity, facilitates a climb towards a "Reciprocity Hill," where cooperation is more readily established and maintained. Our findings highlight the critical roles of agent cognition, learning algorithms, and interaction structure in fostering cooperation in multi-agent systems.

Keywords: N-Person Prisoner's Dilemma · Agent-Based Modelling · Reinforcement Learning · Emergence of Cooperation · Tragedy Valley · Reciprocity Hill · Multi-Agent Systems.

1 Introduction

The Prisoner's Dilemma (PD) serves as a foundational paradigm in game theory, starkly illustrating the conflict between individual rational self-interest and mutually beneficial collective action [1]. In its simplest form, two individuals, unable to communicate, must independently choose whether to cooperate or defect. While mutual cooperation yields a good outcome for both, each player has an individual incentive to defect, leading to a suboptimal outcome if both choose to do so. The Iterated Prisoner's Dilemma (IPD), where the game is played repeatedly, opens the door for cooperation to emerge through strategies

based on reciprocity, as famously demonstrated by Axelrod's tournaments where Tit-for-Tat proved remarkably successful [1].

However, many real-world social and economic dilemmas—ranging from managing common-pool resources to international climate agreements and team collaborations—involve more than two interacting parties. The N-Person Iterated Prisoner's Dilemma (N-IPD) generalizes the IPD to scenarios with n participants [**?**,**?**]. This extension introduces significant complexities:

- **Diffused Responsibility and Payoffs:** The impact of a single agent's cooperative or defective action is spread across the group, diluting the direct consequences felt by any one individual.
- **Obscured Reciprocity:** It becomes harder to identify and respond to specific cooperators or defectors, making direct tit-for-tat like reciprocity challenging.
- **Increased Temptation to Free-Ride:** With many participants, an individual might be more tempted to defect, hoping to benefit from others' cooperation without contributing.

These complexities often lead rational, self-interested agents in N-IPD scenarios towards a "Tragedy Valley" of widespread defection, a concept echoing Hardin's "Tragedy of the Commons" [**?**]. Our computational explorations using agent-based models (ABMs) with standard reinforcement learning (RL) agents consistently confirm this pessimistic outcome in certain N-IPD structures. This paper investigates the cognitive and structural conditions that allow learning agents to escape this valley and, in more favorable settings, ascend a "Reciprocity Hill" where cooperation can flourish.

We present `npdl`, an agent-based simulation framework, to explore these dynamics. Our central argument is that the emergence of cooperation in the N-IPD is not solely dependent on sophisticated learning algorithms but is critically shaped by (a) the agents' ability to perceive **context** from their social environment, (b) the inherent **adaptability** of their learning mechanisms, and (c) the fundamental **interaction structure** of the dilemma itself.

The key takeaways from our investigation are:

1. **The "Tragedy Valley" vs. "Reciprocity Hill" (Interaction Structure - T1):** The structure of agent interactions is paramount.
   - In N-IPD *neighbourhood models*, where an agent's single choice affects a diffuse group payoff, most learning algorithms (including standard RL and simpler reactive strategies like Tit-for-Tat) tend to descend into the "Tragedy Valley" of defection.
   - In contrast, N-IPD *pairwise models*, where each agent effectively makes $N-1$ choices by engaging in distinct 2-player games with all others, direct reciprocity is clear. This structure facilitates climbing a "Reciprocity Hill" where cooperation is more readily established and maintained.
2. **Context is Crucial for Escaping the Valley (Cognitive Prerequisite - T2):** For agents operating in the challenging neighbourhood model, perceiving local social context (e.g., the proportion of cooperating neighbours) is a vital first step to avoid immediate and total defection.

3. **Adaptive MARL Can Navigate the Valley (Learning Mechanism - T3):** Even within the difficult neighbourhood model, advanced Multi-Agent Reinforcement Learning (MARL) algorithms—particularly those incorporating optimism (like Hysteretic-Q) or adaptive learning rates (like Wolf-PHC)—can enable agents to learn resilient cooperative strategies and achieve high, sustained cooperation. Standard RL often fails where these succeed.

This paper will first provide a brief background on the N-IPD and relevant learning approaches (Section 2). We then describe the `npdl` simulation framework and its distinct interaction models (Section 3), followed by our experimental methodology (Section 4). Results supporting our key takeaways are presented in Section 5. Finally, we discuss the broader implications of these findings for understanding and fostering cooperation in multi-agent systems (Section 6) and conclude with future research directions (Section 7).

2 Background and Related Work

This section briefly reviews key concepts from game theory, focusing on the N-Person Prisoner's Dilemma (N-IPD), and introduces the Agent-Based Modelling (ABM) and Multi-Agent Reinforcement Learning (MARL) approaches employed in our study. While the broader field of learning and adaptation informs this work, our focus on abstract agent learning in game-theoretic N-IPD scenarios distinguishes it from prior research centered on specific neural architectures or spiking neuron models (e.g., [2,**?**,**?**,**?**]).

2.1 The N-Person Prisoner's Dilemma (N-IPD)

The N-IPD extends the classic two-person dilemma to $N \geq 2$ players, presenting a more complex challenge for cooperation [**?**,**?**]. Let n_c be the number of players in a group who choose to cooperate (C), and $N - n_c$ be the number who choose to defect (D). The payoff to an individual cooperator is $P_C(n_c)$ and to an individual defector is $P_D(n_c - 1)$ (when considering a defector, they are not part of the n_c cooperators in their own payoff calculation from that group). The dilemma is typically characterized by two conditions:

- **Dominance of Defection:** For any individual, defecting yields a higher personal payoff than cooperating, regardless of how many others cooperate. Formally, if an agent considers switching from cooperate to defect, its payoff increases: $P_D(n_c) > P_C(n_c + 1)$ where n_c is the number of *other* cooperators (if the agent defects) versus $n_c + 1$ (if the agent cooperates).
- **Deficient Equilibrium:** Universal cooperation yields a better payoff for every individual than universal defection: $P_C(N) > P_D(0)$.

This inherent conflict between individual rationality (defect) and collective benefit (cooperate) often leads to the "Tragedy of the Commons" [**?**], where shared resources are depleted. Our concept of the "Tragedy Valley" directly reflects this gravitational pull towards mutual defection observed in N-IPD simulations.

2.2 Agent-Based Modelling (ABM) for Social Dilemmas

Agent-Based Modelling (ABM) offers a powerful computational methodology for studying complex social systems from the bottom up [**?**,**?**]. By simulating the actions and interactions of autonomous, heterogeneous agents according to predefined rules within a specified environment, ABM allows researchers to observe emergent macroscopic phenomena, such as the rise or fall of cooperation. It is particularly well-suited for exploring the N-IPD due to its ability to model local interactions, diverse agent strategies (including learning), and the non-linear dynamics that often characterize social dilemmas. Axelrod's pioneering tournaments using ABM for the 2-player IPD provided early insights into the conditions favoring cooperative strategies [1].

2.3 Reinforcement Learning in Multi-Agent Systems (MARL)

Reinforcement Learning (RL) is a class of machine learning where agents learn to make optimal sequences of decisions by interacting with an environment and receiving feedback in the form of rewards or punishments [**?**]. Standard Q-learning is a foundational RL algorithm that learns the value of taking a particular action in a given state. However, when applied to multi-agent systems (MARL), where multiple agents are learning simultaneously, standard RL algorithms face significant challenges, primarily due to the non-stationarity of the environment: each agent's policy changes as it learns, thereby changing the environment from the perspective of other agents [**?**]. This can destabilize learning and prevent convergence to cooperative equilibria. To address these issues within the N-IPD context, our work explores more advanced MARL techniques:

- **Hysteretic Q-learning:** This algorithm employs asymmetric learning rates, specifically using a higher learning rate for positive updates (when an action leads to a better-than-expected outcome) and a lower one for negative updates. This "optimism" can help sustain cooperation by making agents less reactive to occasional defections and quicker to reinforce mutually beneficial actions [**?**].
- **Win-or-Learn-Fast Policy Hill-Climbing (WoLF-PHC):** WoLF-PHC dynamically adjusts an agent's learning rate based on its performance relative to an average policy. If an agent is "winning" (performing better than average), it learns more cautiously (lower learning rate); if "losing," it learns more rapidly (higher learning rate). This adaptability helps agents converge in non-stationary settings [**?**].

These algorithms represent attempts to endow agents with more sophisticated learning capabilities to navigate the complexities of multi-agent interactions.

3 The `npdl` Simulation Framework and Interaction Models

We developed `npdl`, a Python-based ABM platform. Key components include agent architecture and distinct interaction models.

3.1 Agent Architecture

Agents use learning strategies. Standard Q-learning agents perceive states based on their local neighbourhood. The proportion$_d iscretizedstaterepresentation, quantifyingneighbourcooperat$
$QorWolf - PHC$.

3.2 Interaction Models: Neighbourhood vs. Pairwise

npdl simulates two N-IPD interaction structures:

1. Neighbourhood Model: Agents interact with local network neighbours. Payoffs are from N-player functions based on neighbourhood cooperation. This represents diffuse public good scenarios and often leads to the "Tragedy Valley."
2. Pairwise Model: Each agent plays a 2-player IPD against every other agent. Total payoff sums these dyadic interactions. This emphasizes direct reciprocity, allowing strategies like Tit-for-Tat (TFT) to function effectively. This structure facilitates climbing the "Reciprocity Hill."

The pairwise model required careful agent memory handling for reactive strategies (per-opponent history) and RL agents (aggregate signals).

4 Methodology and Experiments

Simulations typically involved $N = 30$ agents, 500 rounds, Small-World networks, and standard PD payoffs ($R = 3, S = 0, T = 5, P = 1$). We evaluated:
Baseline Q-learning agents with minimal (basic) and contextual (proportion$_d iscretized)staterep$
$QandWolf-PHC(oftenagainstTFTagents), globalcooperationbonuses, andbothNeighbourhoodandPairwis$

5 Results

This section presents key experimental results.

5.1 The Tragedy Valley and the Importance of Context

Standard Q-learning agents with a ‘basic‘ state (no neighbour information) rapidly converged to near-zero cooperation (the "Tragedy Valley"). Providing proportion$_d iscretizedstate(fractionofcooperatingneighbours)improvedperformancetounstable\ 50\%cooper$

5.2 Adaptive MARL Achieves High Cooperation in Neighbourhood N-IPD

Optimized Hysteretic-Q and Wolf-PHC agents achieved high, sustained cooperation (over 85-90%) in the N-IPD neighbourhood model, even against TFT agents. Hysteretic-Q’s optimism and Wolf-PHC’s adaptive learning rates were effective.

5.3 Impact of Interaction Structure: Pairwise Model and the Reciprocity Hill

The pairwise interaction model, with explicit direct reciprocity, fundamentally alters the strategic landscape. Initial observations and theory suggest this structure makes the "Reciprocity Hill" more accessible, as feedback for cooperation/defection is immediate and unambiguous. RL agents benefit from clearer underlying reward signals.

6 Discussion

Our experimental results shed light on the critical factors influencing the emergence and stability of cooperation in the N-Person Iterated Prisoner's Dilemma, painting a narrative of challenges and pathways towards collective benefit. The concepts of the "Tragedy Valley" and the "Reciprocity Hill" serve as useful metaphors for the different dynamic landscapes agents encounter.

T1: Interaction Structure as the Primary Determinant { The Valley and The Hill. Perhaps our most fundamental insight is the profound impact of the interaction structure itself. The standard N-IPD *neighbourhood model*, where an agent makes a single choice and its payoff is determined by the collective actions within its local group, inherently presents a difficult path to cooperation. The benefits of an individual's cooperation are diffused, while the costs are borne individually. Direct, targeted reciprocity is obscured, making it hard for simple reciprocal strategies like Tit-for-Tat to gain a foothold or for learners to accurately assign credit for good outcomes. This environment readily leads agents into the "Tragedy Valley" of mutual defection. Even if agents possess some learning capabilities, the path out of this valley is steep and fraught with the risk of being exploited.

In stark contrast, the *pairwise model* of N-IPD fundamentally alters this landscape. Here, each agent effectively engages in $N-1$ distinct 2-player IPD games with every other participant. This structure brings clarity and directness to reciprocity. A defection from agent B towards agent A directly impacts A's payoff from that specific interaction, and A can retaliate or forgive B in their subsequent dyadic game without "punishing" other innocent bystanders. This clear cause-and-effect makes it much easier for cooperative norms, supported by reciprocal strategies, to emerge and stabilize. The "Reciprocity Hill" becomes a more accessible and sustainable state because individual incentives are better aligned with mutual cooperation through direct, accountable interactions. Our framework's ability to model both structures allowed us to highlight this critical difference.

T2: Contextual Awareness { A Perceptual Foothold Against the Valley's Slope. For agents operating within the challenging neighbourhood model,

the ability to perceive their immediate social context is a crucial first defense against an immediate slide into the Tragedy Valley. Our results (Section 5A) clearly showed that Q-learning agents with no information about their neighbours' actions ('basic' state) invariably defected. However, simply providing them with a discretized proportion of cooperating neighbours ('proportion$_{d}$iscretized'state)allowedforasignificantlyhigher, albeitunstable, levelofcooperat

T3: Adaptive MARL { The Cognitive Tools to Climb in Difficult Terrain. While context is necessary, more sophisticated cognitive tools in the form of adaptive learning are often required to navigate the N-IPD neighbourhood model successfully and sustain high levels of cooperation. Standard reinforcement learning, represented by basic Q-learning, struggles with the non-stationarity of the multi-agent environment. In contrast, adaptive MARL algorithms like Hysteretic-Q and Wolf-PHC demonstrated the capacity to achieve and maintain robust cooperation (Section 5B). Hysteretic-Q, with its optimistic approach of learning more readily from positive experiences (mutual cooperation) than from negative ones (being defected upon), fosters resilience. It allows agents to "forgive" occasional defections and maintain cooperative overtures, preventing spirals of retaliation that can trap agents in the Tragedy Valley. Wolf-PHC, by dynamically adjusting its learning rates based on whether it is "winning" or "losing," shows adaptability to the changing strategies of other agents. This careful modulation of learning helps it find and stabilize cooperative equilibria that elude simpler learners. These algorithms, therefore, represent mechanisms by which agents can, through more nuanced learning, not only avoid the worst of the Tragedy Valley but actively "climb" towards more cooperative states even in the challenging diffuse-payoff structure of the neighbourhood N-IPD. Their success underscores that the type of learning matters significantly; it's not just about learning, but *how* agents learn in a social context.

The consistent failure of standard exploration strategies like UCB1 (detailed in the original report but not explicitly shown with a figure here) further emphasizes the difficulties posed by non-stationarity in MARL within the N-IPD. Additionally, while external factors like global incentives can dramatically shift the landscape towards cooperation (as also shown in our broader study), the focus of these core takeaways is on the inherent learnability and structural properties of the dilemma itself and the cognitive capabilities of the agents. Our findings collectively suggest that understanding and promoting cooperation in complex multi-agent systems requires attention not only to individual learning capacities but also, critically, to the very structure of their interactions. Limitations of this study include the abstraction of agent cognition and the focus on specific network structures and payoff parameters. Future work will expand the range of scenarios and delve deeper into the comparative dynamics of the neighbourhood versus pairwise models, further exploring the pathways from the Tragedy Valley to the Reciprocity Hill.

7 Conclusion

This paper demonstrated that while N-IPD in neighbourhood models leads to a "Tragedy Valley" for simple learners, cooperation can emerge with **contextual awareness** and **adaptive MARL algorithms** (Hysteretic-Q, Wolf-PHC). The **interaction structure** is critical: pairwise models, facilitating a "Reciprocity Hill," make cooperation more accessible. The npdl framework enables these explorations. Future work will further investigate these dynamics to understand and promote cooperation in complex multi-agent systems.

References

1. R. Axelrod and W. Hamilton, ‘‘The evolution of cooperation,’’ *Science*, vol. 211(4489), pp. 1390--1396, 1981.
2. C. Huyck, ‘‘Learning categories with spiking nets and spike timing dependent plasticity,’’ in *SGAI 2020*, pp. 139--144, 2020.
3. C. Huyck and C. Samey, ‘‘Extended category learning with spiking nets and spike timing dependent plasticity,’’ in *SGAI 2021*, pp. 33--43, 2021.
4. C. Huyck and O. Erekpaine, ‘‘Competitive learning with spiking nets and spike timing dependent plasticity,’’ in *SGAI 2022*, pp. 153--166, Springer, 2022.
5. P. Diehl and M. Cook, ‘‘Unsupervised learning of digit recognition using spike-timing-dependent plasticity,’’ *Frontiers in computational neuroscience*, vol. 9, p. 99, 2015.