

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

Chris Tcaci · Chris Huyck¹

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 `npd1`, 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, the N-IPD, agent-based modelling, and reinforcement learning relevant to our study. The work reported in this paper builds upon a broader understanding of learning and adaptation, though it diverges from prior work focused on spiking neural models such as those by [2,?,?] and mechanisms derived from Diehl and Cook [5], by focusing on abstract agent learning in game-theoretic scenarios.

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

The N-IPD extends the two-person dilemma to N players. Let n_c be the number of players who choose to cooperate (C). The payoff to a cooperator is $P_C(n_c)$ and to a defector (D) is $P_D(n_c)$. The dilemma is characterized by:

- **Dominance of Defection:** $P_D(n_c) > P_C(n_c + 1)$ for all $0 \leq n_c < N$. An individual always gains more by defecting.
- **Deficient Equilibrium:** $P_C(N) > P_D(0)$. Mutual cooperation is better for all than mutual defection.

This structure often leads to the "Tragedy of the Commons."

2.2 Agent-Based Modelling (ABM) for Social Dilemmas

Agent-Based Modelling (ABM) provides a bottom-up approach to studying complex systems by simulating autonomous agents. It is well-suited for exploring the N-IPD, allowing for heterogeneous strategies, local interactions, and emergent global patterns. Axelrod’s tournaments for the 2-player IPD set a precedent [1].

2.3 Reinforcement Learning in Multi-Agent Systems (MARL)

Reinforcement Learning (RL) enables agents to learn optimal actions through trial-and-error. Standard Q-learning faces challenges in multi-agent settings (MARL) due to non-stationarity. To address these, we implemented:

- **Hysteretic Q-learning:** Employs asymmetric, optimistic learning rates.
- **Win-or-Learn-Fast Policy Hill-Climbing (WoLF-PHC):** Adjusts learning rates based on performance.

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 `proportiondiscretizedstaterepresentation, quantifyingneighbourcooperation` or `Wolf - 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 (`proportiondiscretizedstaterepresentation` and `Wolf - PHC`) (often against TFT agents), global cooperation bonuses, and both Neighbourhood and Pairwise

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_{discretizedstate}(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 results highlight several key points. The **"Tragedy Valley"** is a common outcome in neighbourhood N-IPD for simple learners due to diffused incentives. **Contextual awareness** is crucial; agents need to perceive local social cues. **Adaptive MARL algorithms** like Hysteretic-Q and Wolf-PHC can overcome these challenges in neighbourhood N-IPD through sophisticated learning. Most fundamentally, **interaction structure** is a powerful determinant: the pairwise model’s direct reciprocity creates a "Reciprocity Hill," making cooperation inherently easier to achieve and sustain than in the diffuse neighbourhood model. The failure of standard exploration strategies like UCB1 (not shown here but detailed in original report) highlights non-stationarity challenges. Global incentives can also significantly boost cooperation. Limitations include abstracted cognition and specific parameter choices.

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.