COMENIUS UNIVERSITY IN BRATISLAVA

FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

Multiagent model of active inference

Master Thesis

2025

Armin Nouri, BA

COMENIUS UNIVERSITY IN BRATISLAVA

FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

Multiagent model of active inference

Master Thesis

Study Programme: Cognitive Science

Field of Study: Computer Science

Department: Department of Applied Informatics

Supervisor: doc. RNDr. Martin Takáč, PhD

Bratislava 2025

Armin Nouri, BA

Comenius University Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Armin Nouri
Study programme:	Cognitive Science (Single degree study, master II. deg., full time form)
Field of Study:	Computer Science
Type of Thesis:	Diploma Thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Multiagent model of active inference

Annotation: Active Inference is a process where an agent actively samples its environment to reduce uncertainty about its state and thus minimize its Free Energy (Friston, 2010). The agent does this by updating its internal model to better predict sensory inputs and by acting in ways that bring its sensory inputs in line with its predictions. In a collective environment, the agent has to take into account actions of other agents. Two-agent models exist for various tasks and games, such as iterative prisoner's dilemma (Demekas et al., 2023). The goal of this thesis is to extend the framework to cover more than two agents.

Aim:
1. Gain a theoretical insight into free energy principle and active inference framework.
2. Reimplement a two-agent model of active inference, such as Demekas et al. (2023).
3. Extend the model to more than two (at least three) agents.

Literature:
Demekas, D., Heins, C., & Klein, B. (2023). An analytical model of active inference in the iterated prisoner's dilemma. In Communications in computer and information science (pp. 145–172).
Friston, K. The free-energy principle: a unified brain theory?. Nat Rev Neurosci 11, 127–138 (2010).
Parr, T. Pezzulo, G., Friston, K. (2022): Active Inference: The Free Energy Principle in Mind, Brain, and Behavior. The MIT Press.

Supervisor:	doc. RNDr. Martin Takáč, PhD.
Department:	FMFI.KAI - Department of Applied Informatics
Head of department:	doc. RNDr. Tatiana Jajcayová, PhD.
Assigned:	27.03.2025

Approved: 04.04.2025 prof. Ing. Igor Farkaš, Dr.
Guarantor of Study Programme

Student Supervisor

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Armin Nouri
Študijný program: kognitívna veda (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Multiagent model of active inference
Multiagentový model s aktívnou inferenciou

Anotácia: Aktívna inferencia je proces, pri ktorom agent aktívne vzorkuje svoje prostredie, aby znížil neistotu svojho stavu a minimalizoval tak svoju voľnú energiu (Friston, 2010). Agent to robí tak, že aktualizuje svoj vnútorný model, aby lepšie predpovedal zmyslové vstupy, a koná tak, aby jeho zmyslové vstupy boli v súlade s jeho predikciami. V kolektívnom prostredí musí agent brať do úvahy konanie iných agentov. Modely dvoch agentov existujú pre rôzne úlohy a hry, napríklad iteratívnu väzňovu dilemu (Demekas a kol., 2023). Cieľom tejto práce je rozšíriť rámec tak, aby pokrýval viac ako dvoch agentov.

Cieľ: 1. Získať teoretický prehľad princípu voľnej energie aktívnej inferencie.
2. Reimplementovať model aktívnej inferencie s dvoma agentmi, napr. Demekas a kol. (2023).
3. Rozšíriť model na viac ako dvoch (aspoň troch) agentov.

Literatúra: Demekas, D., Heins, C., & Klein, B. (2023). An analytical model of active inference in the iterated prisoner's dilemma. In Communications in computer and information science (pp. 145–172).
Friston, K. The free-energy principle: a unified brain theory?. Nat Rev Neurosci 11, 127–138 (2010).
Parr, T. Pezzulo, G., Friston, K. (2022): Active Inference: The Free Energy Principle in Mind, Brain, and Behavior. The MIT Press.

Vedúci: doc. RNDr. Martin Takáč, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: doc. RNDr. Tatiana Jajcayová, PhD.

Dátum zadania: 27.03.2025

Dátum schválenia: 04.04.2025

prof. Ing. Igor Farkaš, Dr.
garant študijného programu

.....
študent

.....
vedúci práce

Abstract

This study examines how agents operating under the Active Inference framework can autonomously develop cooperative behavior in the Iterated Prisoner's Dilemma through purely inferential processes, without recourse to pre-programmed strategies or externally imposed reward functions. Each agent maintains a probabilistic generative model of state transitions, refines its beliefs by accumulating Dirichlet pseudo-counts, and selects actions by minimizing Expected Free Energy, a principled trade-off between pursuing preferred outcomes and reducing uncertainty about the environment.

In two-agent simulations, belief updates drive a characteristic progression: early trials are marked by an irregular alternation of cooperation and defection, followed by a transient dominance of unilateral defection, culminating in a stable regime of mutual cooperation. Systematic parameter sweeps reveal that neither very slow nor overly rapid learning fosters robust coordination; instead, an intermediate learning rate combined with moderate decision noise and minimal emphasis on epistemic gain yields the highest cooperation rates. Under these balanced conditions, agents spontaneously exhibit Win–Stay tendencies as a natural outgrowth of their free-energy minimization, while Lose–Shift patterns only emerge under slower learning dynamics, demonstrating how familiar heuristic motifs can arise without explicit coding.

When extended to a three-agent environment, introducing an additional player increases strategic complexity and initially undermines cooperation. Nevertheless, by suitably adjusting learning rates, triads of Active Inference agents can regain high levels of joint cooperation, underscoring the framework's scalability and capacity to accommodate greater social uncertainty through adaptive belief updating.

Across both two- and three-agent scenarios, our results illustrate that sophisticated cooperative behaviors and elements of classical strategies emerge organically from the interaction of generative modeling, Bayesian belief updating, and free-energy-based action selection. These findings position Active Inference as a compelling alternative to traditional reinforcement learning and fixed-rule approaches for modeling adaptive, socially intelligent agents in dynamic and uncertain multi-agent environments.

Keywords: Active Inference; Iterated Prisoner's Dilemma; cooperation; Expected Free Energy; Dirichlet learning; multi-agent adaptation.

Abstrakt

Táto práca skúma, ako sa môže u agentov používajúcich aktívnu inferenciu samostatne vyvinúť kooperatívne správanie v Iterovanej väzeňskej dileme prostredníctvom čisto inferenčných procesov bez použitia vopred naprogramovaných stratégií alebo externe zavedených funkcií odmeňovania. Každý agent si udržiava pravdepodobnostný generatívny model prechodov medzi stavmi, spresňuje svoje presvedčenia akumuláciou Dirichletových pseudo-počtov a vyberá akcie minimalizáciou očakávanej voľnej energie, čo je kompromis medzi sledovaním preferovaných cieľov a znižovaním neistoty o prostredí.

V simuláciách s dvoma agentmi vidíme charakteristický priebeh aktualizácie presvedčení: prvé pokusy sa vyznačujú nepravidelným striedaním spolupráce a zrádzania, po ktorom nasleduje prechodná prevaha jednostranného zrádzania, ktorá vyvrcholí stabilným režimom vzájomnej spolupráce. Systematický výber parametrov odhalil, že ani veľmi pomalé, ani príliš rýchle učenie nepodporuje stabilnú koordináciu; stredná rýchlosť učenia v kombinácii s miernym rozhodovacím šumom a minimálnym dôrazom na epistemický zisk prináša najvyššiu mieru spolupráce. Za týchto vyvážených podmienok agenty spontánne prejavujú tendencie Win-Stay ako prirodzený výsledok minimalizácie voľnej energie, zatiaľ čo vzory Lose-Shift sa objavujú len pri pomalšej dynamike učenia. To ukazuje, ako môžu známe heuristické motívy vzniknúť bez explicitného kódovania.

Po rozšírení na prostredie troch agentov sa zavedením ďalšieho hráča zvyšuje strategická zložitosť a spočiatku sa oslabuje spolupráca. Vhodným nastavením miery učenia však môžu trojice agentov s aktívnou inferenciou opäť dosiahnuť vysokú úroveň vzájomnej spolupráce, čo zdôrazňuje škálovateľnosť aktívnej inferencie a schopnosť prispôsobiť sa väčšej sociálnej neistote prostredníctvom adaptívnej aktualizácie presvedčení.

Naše výsledky v scenároch s dvoma aj tromi agentmi ukazujú, že sofistikované kooperatívne správanie a prvky klasických stratégií vznikajú organicky z interakcie generatívneho modelovania, bayesovskej aktualizácie presvedčení a výberu akcií na základe voľnej energie. Tieto zistenia stavajú aktívnu inferenciu do pozície presvedčivej alternatívy k tradičnému učeniu posilňovaním a k prístupom s pevnými pravidlami na modelovanie adaptívnych, sociálne inteligentných agentov v dynamických a neistých multiagentových prostrediach.

Kľúčové slová: Aktívna inferencia; Iterovaná väzeňská dilema; spolupráca; Očakávaná voľná energia; Dirichletovo učenie; adaptácia viacerých agentov.

Contents

1. Introduction 1
    1.1 Motivation 1
    1.2 Limitations of Reinforcement Learning in IPD 2
    1.3 Ecological and Empirical Complexity 3
    1.4 Active Inference as an Alternative 3
    1.5 Research Questions 4
    1.6 Contributions 5
    1.7 Significance of the Study 6

2. Literature Review 7
    2.1 Active Inference Framework 7
    2.2 Reinforcement Learning and IPD 9
    2.3 Existing Applications 11

3. Theoretical Framework 12
    3.1 Generative Models 12
    3.2 Learning Mechanisms 13

4. Methodology 14
    4.1 Overview of Methodology 14
    4.2 Computational Modeling Approach 15
        4.2.1 Generative Model Components 15
    4.3 Agent Behavior and Learning Algorithm 20
        4.3.1 State Inference 20
        4.3.2 Action Selection via Expected Free Energy 20
        4.3.3 Transition Model Learning 21
    4.4 Experimental Design 22
        4.4.1 Parameters and Conditions 22
        4.4.2 Simulation Setup 23
    4.5 Evaluation Criteria and Reproducibility 23
        4.5.1 Evaluation Criteria 23

4.6 Software, Libraries, and Computational Setup 25

5. Results 25
5.1 Baseline Two-Agent Learning Trajectory 26
5.2 Mapping Two-Agent Cooperation Across Cognitive Parameters 33
5.2.1 Cooperation as a Function of Learning Rate 33
5.2.2 Cooperation as a Function of Policy Precision 35
5.2.3 Epistemic Weighting (γ) Sweep 36
5.3 Emergent Strategy Patterns 38
5.4 Three-Agent Simulation 40
5.4.1 Learning-rate Sweep in Three-Agent Simulation 42

6. Discussion 43
6.1 Revisiting Baseline Two-Agent Dynamics 43
6.2 The Role of Learning-Rate Symmetry 44
6.3 Action-Precision and the Exploration–Exploitation Trade-off 45
6.4 Epistemic Weighting: Curiosity Versus Stability 45
6.5 Emergence of Classical Heuristics 46
6.6 Scaling Up: Three-Agent Dynamics 46
6.7 Broader Implications, Limitations, and Future Directions 47

7. Conclusion 48

Appendix A: Code Snippets and Definitions 53
A.1 IPD Environment Definition 53
A.2 Active Inference Agent Class 54
A.3 Simulation Helper Functions 56

Appendix B: Generative Model and Inference Equations 57
B.1 Generative Model 58
B.2 Bayesian State Inference 59
B.3 Learning via Dirichlet Updates with Decay 59
B.4 Expected Free Energy and Action Selection 60

Appendix C: Use of Artificial Intelligence Tools 61

List of Figures

Figure 5.1 Cumulative fractions of CC, CD, DC, DD over trials ……. 27
Figure 5.2 Expected Free Energy trajectories for both agents ……..… 29
Figure 5.3 Cumulative rewards over time ……..…… 30
Figure 5.4 Total cumulative rewards per agent …….. 30
Figure 5.5 Cumulative rewards over time (rolling average) ……..… 31
Figure 5.6 Total cumulative rewards per agent (bar chart) ……..… 31
Figure 5.7 Learning progress (KL divergence from uniform) ……..… 32
Figure 5.8 Cooperation rate vs. learning rates (η_1, η_2) ……..… 34
Figure 5.9 Cooperation rate vs. policy precision (α_1, α_2) ……..… 36
Figure 5.10 Cooperation rate vs. epistemic weighting (γ_1, γ_2) ……..… 37
Figure 5.11 Cumulative and rolling triple-cooperation dynamics (3-agent IPD) ……..… 40
Figure 5.12 Cumulative and rolling triple-cooperation dynamics (3-agent IPD 2) ……...41
Figure 5.13 Three-agent cooperation rate across learning-rate anchors ……...…… 43

List of Tables

Table 5.1 Mean strategy statistics …….. 39

1. Introduction

1.1 Motivation

Cooperation plays a crucial role in the organization of both human and animal societies. However, explaining why individuals choose to cooperate, especially in situations where self-interested behavior would lead to better immediate rewards, remains a central challenge in behavioral science. The Iterated Prisoner's Dilemma (IPD) offers a well-known framework for exploring this problem. While classical approaches based on game theory and reinforcement learning (e.g., Axelrod, 1984; Sandholm & Crites, 1996) have provided useful insights, they often rely on unrealistic assumptions such as perfect rationality, complete information, or fixed strategy sets. In reality, cooperative behavior emerges under uncertainty, through limited information, and in dynamic environments. As Raihani and Bshary (2011) noted, these real-world conditions require models that better account for bounded rationality, context sensitivity, and learning over time.

Active Inference Framework (AIF) offers a compelling alternative grounded in theoretical neuroscience and statistical physics. It provides a unifying framework in which agents act to minimize variational free energy, thereby aligning their beliefs, actions, and observations over time (Friston et al., 2016; Parr et al., 2022). Rather than relying on external reward signals or rigid heuristics, AIF enables adaptive behavior through continual Bayesian inference under a generative model. This perspective allows for goal-directed exploration (Parr et al., 2022), Theory of Mind modeling (Kaufmann et al., 2021), belief-driven learning (Demekas et al., 2023), and emergent coordination, all of which are critical in dynamic multi-agent environments like the IPD. Notably, AIF agents can encode preferences, uncertainty, and social constraints in their beliefs, making the framework especially suited for modeling strategic social interaction (Demekas et al., 2023).

Recent experimental and theoretical work has further shown that long-term cooperative behavior in the IPD is best understood not through fixed strategies, but through evolving probabilistic tendencies modulated by context and interaction structure (Montero-Porras et al., 2022; Martínez-Martínez & Normann, 2022). These insights position AIF as a

promising computational substrate for modeling cooperation that emerges organically from continual belief updates rather than prescriptive rules. Furthermore, AIF's roots in thermodynamics and variational inference provide a principled foundation for unifying perception, action, learning, and planning (Friston et al., 2006).

Cooperation is fundamental to emerging complex social, economic, and biological behavior. However, explaining how cooperative behavior arises and stabilizes remains a central challenge across cognitive science, artificial intelligence, and evolutionary biology (Raihani & Bshary, 2011; Grujić et al., 2012; Galesic et al., 2023). This puzzle is exemplified by the Iterated Prisoner's Dilemma (IPD), a strategic game in which agents repeatedly decide whether to cooperate or defect, with outcomes shaped by the joint action of all players.

IPD reveals a fundamental tension, although mutual cooperation can maximize group outcomes, short-term incentives often favor unilateral defection. Classical strategies such as Tit-for-Tat (Axelrod, 1984), Grim Trigger (Nowak & Sigmund, 1990), and Win-Stay, Lose-Shift (Nowak & Sigmund, 1993) have been proposed to resolve this dilemma by promoting reciprocity and contingent behavior. These strategies offer strong performance under idealized conditions but assume perfect rationality, fixed strategy rules, and full observability of other agents' behavior.

As Akin (2015) emphasizes, many classical strategies depend on unrealistic assumptions such as the agent's ability to identify game structure and apply backward reasoning perfectly. These assumptions limit their applicability to real-world social scenarios, where noise, ambiguity, and bounded rationality are the norm. Baek and Kim (2008) further show that traditional strategies can be highly fragile under even minor disturbances, failing to sustain cooperation when agents are uncertain or probabilistic in their responses.

1.2 Limitations of Reinforcement Learning in IPD

Reinforcement learning (RL) offers a data-driven alternative but faces challenges in dynamic, non-stationary multi-agent settings. For example, Sandholm and Crites (1996) found that Q-learning agents frequently fail to reach cooperative equilibria, especially when paired with other learners. Without centralized coordination or aligned priors, RL agents may oscillate between suboptimal outcomes or converge to mutual defection.

Moreover, reinforcement learning (RL) often lacks principled mechanisms for uncertainty-aware exploration, relying instead on heuristic approaches such as ε-greedy or softmax action selection. This limitation has been noted in multi-agent IPD contexts, where heuristic-driven exploration can fail to capture the uncertainty inherent in strategic adaptation (Sandholm & Crites, 1996; Gergely, 2022).

1.3 Ecological and Empirical Complexity

Natural cooperation among animals, humans, or artificial agents often unfolds under partial observability, incomplete information, and diverse social norms. Raihani and Bshary (2011) argue that ecological realism is rarely captured by classical IPD models, which assume strict reciprocity and infinite memory. In contrast, real-world agents must infer intentions, estimate trustworthiness, and adapt to changing social landscapes.

Empirical studies reinforce this complexity. For example, Grujić et al. (2012) show that human participants in multiplayer IPD settings exhibit heterogeneous, context-sensitive strategies. Their behavior is not governed by strict rules but is modulated by prior outcomes, emotional states, and social framing. Fogel (1993) further emphasizes that evolutionary dynamics and adaptive exploration, not hard-coded strategies, better explain the emergence of stable cooperation in nature.

1.4 Active Inference as an Alternative

The Active Inference provides a promising framework for modeling adaptive, belief-driven cooperation. Rooted in the Free Energy Principle (Friston, 2010), AIF frames perception, action, and learning as processes of Bayesian inference under a generative model. Rather than maximizing extrinsic rewards, AIF agents minimize variational free energy, aligning internal beliefs with observations and preferences.

Expected Free Energy (EFE) plays a central role in Active Inference: it combines risk minimization, favoring outcomes aligned with prior preferences, and epistemic value, seeking actions that reduce uncertainty about the environment or other agents. This dual structure enables agents to integrate exploration and exploitation in a principled manner, without the need for externally imposed heuristics (Parr et al., 2022, Ch. 2; Demekas et al., 2023).

1.5 Research Questions

This thesis investigates the potential of Active Inference as a robust, belief-based framework for modeling cooperation in repeated social dilemmas, with a primary focus on the Iterated Prisoner's Dilemma. At its core lies the question of whether agents that minimize variational free energy, rather than follow explicitly coded rules, can develop robust cooperative behavior akin to classical strategies such as Tit-for-Tat (Axelrod & Hamilton, 1981) or Pavlov/Win-Stay-Lose-Shift (Nowak & Sigmund, 1993). By replacing hard-wired heuristics with continual belief updating and policy evaluation, we seek to determine if Active Inference agents can adapt flexibly to changing circumstances and generate cooperation through purely inferential means.

A second dimension of this work examines how cooperation emerges and endures within groups of interacting agents. Drawing on empirical and theoretical insights suggesting that factors such as group size, the structure of interactions critically shape cooperative dynamics (Grujić et al., 2012; Martinez-Martinez & Normann, 2022), we ask: under what combinations of internal cognitive parameters (epistemic‑weight parameter, inverse precision, learning rate), Active Inference agents converge on stable cooperative equilibria? To address this, we systematically vary parameters governing belief updating, such as asymmetric learning rates that capture differences in how agents weigh new versus prior information. Through these multi-agent simulations, we aim to map the regions of parameter space where cooperation flourishes and identify thresholds beyond which defection becomes the dominant strategy.

Another specific cognitive parameter, policy precision (α), modulates the stochasticity of action selection, and epistemic weighting (γ), which governs the balance between goal-directed planning and information-seeking exploration. Building on the computational framework established by Demekas et al. (2023), we explore how variations in these parameters influence both the speed and robustness of cooperative convergence. Recent experimental work by Galesic et al. (2023) highlights that individuals' choices in social dilemmas are driven as much by their social expectations and personal norms as by material payoffs, pointing to the importance of modeling cognitive dissonance and projection processes. By embedding these insights into our simulations, we investigate

how agents' prior beliefs about others and their willingness to revise those beliefs when faced with contradictory evidence shape the evolution of cooperation over repeated interactions.

In a complementary investigation, we probe whether the inferential machinery underpinning Active Inference can give rise to well-known cooperative heuristics without direct encoding. Inspired by earlier evolutionary studies (Baek & Kim, 2008; Fogel, 1993), we examine whether agents engaging in Dirichlet learning over policy priors can implicitly discover patterns analogous to Tit-for-Tat or Win-Stay-Lose-Shift. This approach treats classical strategies not as prescriptions but as emergent regularities of the underlying inference process: if cooperation consistently maximizes expected free energy under certain conditions, will the resulting policy posterior mirror these canonical strategies? Verifying this hypothesis would lend credence to Active Inference as a unifying theory capable of reproducing a broad repertoire of social behaviors.

Finally, we extend the standard two-player IPD to scenarios involving three interacting agents to bridge our modeling efforts with real-world complexity. Building on the empirical finding that cooperation rates decline as the number of participants increases is attributed to heightened strategic uncertainty and coordination challenges (Martinez-Martinez & Normann, 2022). We simulate incremental expansions from dyadic to triadic settings. Our goal is to determine whether adding a third player introduces sufficient inferential ambiguity to destabilize cooperative regimes, and if so, to characterize the mechanisms by which belief updating and policy selection falter under multi-party uncertainty. In doing so, this thesis provides both a rigorous computational account of cooperative dynamics and novel predictions about the cognitive and environmental prerequisites for sustained cooperation in complex social systems.

1.6 Contributions

This thesis makes several key contributions to the intersection of probabilistic inference, social decision-making, and strategic interaction. First, it introduces a novel multi-agent Active Inference framework for simulating the Iterated Prisoner's Dilemma. This framework is grounded in full Expected Free Energy-based action selection and incorporates Dirichlet learning over transition dynamics, extending the analytical foundation laid out by Demekas et al. (2023). Second, through a series of simulation

studies, it demonstrates that robust cooperation can arise purely from probabilistic belief updating, without any reliance on hard-coded strategies, scalar reward signals, or explicit equilibrium-seeking routines, a finding consistent with the patterns of adaptive social behavior reported by Montero-Porras et al. (2022) and Grujić et al. (2012). Finally, by showing that Active Inference can successfully model cooperation under conditions of uncertainty and partner heterogeneity, this work positions Active Inference as a compelling alternative to traditional reinforcement learning approaches in contexts where ecological validity and dynamic partner modeling are paramount (Sandholm & Crites, 1996; Gergely, 2022).

1.7 Significance of the Study

This study makes a substantive contribution to the evolving dialogue at the crossroads of cognitive modeling, social decision-making, and computational neuroscience by bringing the Active Inference Framework to bear on the Iterated Prisoner's Dilemma, a prototypical model of cooperation and conflict. Rather than relying on static heuristics or externally imposed value functions, Active Inference offers a fully generative, belief-driven account of decision-making under uncertainty. By constructing a multi-agent simulation, cooperation must arise through agents' inferential processes rather than through pre-programmed strategies. This work deepens our understanding of how adaptive behavior can emerge organically in dynamic, partially observable environments.

One of the primary theoretical advances of this thesis is its extension of Demekas et al.'s (2023) analytical model into a scalable simulation platform capable of handling multiple interacting agents. This innovation tests the limits of active inference in increasingly complex settings and enriches the theoretical toolkit available for studying the emergence and stability of cooperation through updating probabilistic beliefs. In bridging disciplines, the thesis weaves together social science, game theory, and artificial intelligence perspectives, demonstrating that Active Inference can serve as a unified lens for examining strategic behavior, belief revision, and social learning across these fields.

Methodologically, the study offers a flexible simulation framework that makes it possible to probe the impact of key cognitive parameters such as learning rates, Policy precision, and the weight placed on epistemic exploration on collective outcomes. By systematically varying these parameters and observing the resulting shifts in cooperation and defection

patterns, the research provides new tools for probing adaptation and coordination in complex social systems.

Ultimately, the findings establish Active Inference not just as a conceptually elegant theory, but as a practically powerful approach for modeling how agents form and adjust beliefs, select strategies, and sustain cooperative relationships. This work lays the groundwork for future applications in artificial multi-agent systems and the study of human-centered social interactions.

2. Literature Review

2.1 Active Inference Framework

AIF is a theoretical framework based on the Free Energy Principle (FEP), which says that all living or self-organizing systems try to avoid being surprised by what they sense in the world (Friston et al., 2006; Parr et al., 2022). Surprise means getting sensory input that doesn't match the system's expectations. For example, if an agent expects to feel warm but suddenly feels cold, that's surprising. To avoid this, the agent either updates its beliefs to better match reality or takes action to make the world more like what it expected, like putting on a jacket. This constant adjustment helps the agent stay in a safe and stable state. Since computing genuine surprise is generally intractable, agents minimize a tractable upper bound; variational free energy, which integrates both model accuracy and complexity in a principled way (Parr et al., 2022, Ch. 2).

From a computational standpoint, Active Inference agents infer hidden states of the environment and act to realize preferred observations by performing approximate Bayesian inference (K. Friston et al., 2015). Perception, action, and learning are unified to minimize variational free energy. Perception updates beliefs to explain incoming sensory data better, while action changes the external world to bring about expected (i.e., less surprising) sensory outcomes. Learning entails updating the agent's generative model itself to improve future inference and control. These processes collectively yield a closed action–perception–learning loop grounded in Bayesian and information-theoretic principles (Parr et al., 2022).

Crucially, AIF introduces EFE as a forward-looking objective function that agents use to evaluate policies. EFE decomposes into epistemic value, which motivates actions that resolve uncertainty, and pragmatic value, which favors outcomes consistent with prior preferences (Parr et al., 2022, Ch. 2). This formulation allows AIF to reconcile exploration and exploitation naturally. This integration is critical in the context of strategic interactions like the IPD. Traditional models often rely on fixed strategies such as Tit-for-Tat (Axelrod & Hamilton, 1981) or Pavlov (Nowak & Sigmund, 1993), which respond deterministically to an opponent's past actions. However, real agents frequently adjust their behavior based on patterns in ongoing interaction (Wedekind & Milinski, 1996). Stewart and Plotkin (2012) showed that the strategy space in IPD is more complex and responsive than such fixed rules imply. Although the agents in this thesis do not explicitly model others' intentions, they do adapt their behavior through belief updates about state transitions, allowing them to implicitly track and respond to regularities in their partner's behavior. This belief-based adaptation supports a more flexible and probabilistic form of strategy selection, which aligns with the richer behavioral patterns observed in empirical studies (Montero-Porras et al., 2022; Grujić et al., 2012).

Active Inference models allow agents to learn how to cooperate over time without being given fixed rules. This idea fits well with Akin's (2015) argument that good strategies in the IPD come from adapting and learning, not from following hardcoded instructions. There, Akin emphasizes that cooperation in IPD emerges when players know transition dynamics and strategy spaces, not from rigid rules. AIF agents implement this insight directly, as their generative models enable them to learn and internalize transition probabilities without predefined scripts.

Active Inference can be seen as an advanced version of reinforcement learning, but it doesn't rely on value functions to work**.** In reinforcement learning, value functions estimate how much reward an agent can expect in the future if it takes a specific action in a particular state. These functions guide the agent to choose actions that maximize long-term rewards. However, in Active Inference, agents don't need to calculate future rewards this way. Instead, they make decisions by minimizing expected free energy, which naturally combines their preferences and the need to reduce uncertainty. This allows AIF agents to act adaptively without explicitly defining or learning value functions (Friston et al., 2020). Baek & Kim (2008) criticize classical RL strategies in IPD as overly rigid and suggest the necessity of flexible, inference-based approaches to cooperation. They argue that

intelligent strategies must adapt to changing interaction histories, a property that AIF provides inherently through continuous belief updating.

Demekas et al. (2023) implement this framework in the context of the IPD, showing how agents initialized with symmetrical generative models transition from defection to cooperation purely through inference-based learning of transition structures. Their agents achieve coordination through belief convergence, not payoff maximization, reinforcing the utility of Active Inference for modeling social behavior under uncertainty.

Taken together, AIF offers a comprehensive, unified account of generative, recursive, and adaptive behavior. AIF is uniquely suited for modeling cooperation in uncertain and dynamic multi-agent contexts like the Iterated Prisoner's Dilemma by grounding action selection in variational inference and enabling belief-based reasoning over hidden social dynamics.

2.2 Reinforcement Learning and IPD

Reinforcement learning offers a robust, data-driven approach to strategy discovery in the Iterated Prisoner's Dilemma (IPD). However, when agents learn independently in this multi-agent setting, they routinely encounter deep-seated obstacles to cooperation. One of the most fundamental is environmental non-stationarity: as each agent updates its policy, the effective "game" faced by its counterpart shifts, violating the stationary Markov assumptions that underlie classical RL convergence proofs. Early work by Tan (1993) demonstrated that treating co-learners as part of a static environment "generally encounters convergence issues," a point echoed by Sandholm and Crites (1996), who observed that while a Q-learning agent can learn to exploit a fixed Tit-for-Tat opponent, two Q-learners playing each other almost immediately collapse to mutual defection rather than discovering a cooperative equilibrium (Sandholm & Crites, 1996; Tan, 1993). Foerster et al. (2017) later showed that deep Q-learning with naive experience replay exacerbates this moving-target problem, requiring elaborate stabilization techniques merely to achieve convergence in simple multi-agent benchmarks (Foerster et al., 2017).

Even when independent RL converges, it typically settles on the risk-dominant Nash equilibrium of mutual defection rather than the cooperative outcome. This coordination failure arises because defecting is a safe best response to defection, whereas cooperation exposes the agent to potentially irreversible "sucker" penalties if the partner deviates.

Claus and Boutilier (1998) characterized the resulting equilibrium-selection challenge, noting that standard RL carries no intrinsic bias toward the cooperative basin of attraction. They showed that whether Q-learning agents converge to cooperation or defection hinges sensitively on hyperparameters like memory depth, discount factor, and exploration schedule—parameters that must be painstakingly tuned to promote reciprocity (Claus & Boutilier, 1998). Without centralized critics, aligned priors, or specialized opponent models, independent learners gravitate toward the safer but socially suboptimal defect-defect outcome.

Compounding these difficulties is the reliance of RL on undirected, heuristic exploration schemes—ε-greedy randomness or softmax action selection—that treat exploration as mere noise. In the IPD, uncoordinated exploratory cooperations often backfire: one agent's chance cooperation may be met with defection by an exploiting partner, yielding a low payoff that reinforces further defection. Sandholm and Crites (1996) observed that unless ε remains high for many thousands of episodes, Q-learners "converge prematurely to mutual defection," effectively locking in the very outcome that exploration was meant to overcome (Sandholm & Crites, 1996).

Finally, traditional RL's reactive, model-free policies lack foresight and theory of mind. Without an explicit generative model of how one's current actions will influence the other agent's future behavior, RL agents cannot plan through the opponent's adaptation or deliberately signal cooperative intent. They simply reinforce actions that yielded high immediate rewards, blind to longer-term strategic gains that would follow from initial sacrifices. Advanced extensions such as Learning with Opponent Learning Awareness (LOLA) partly address this by anticipating the opponent's learning updates. Still, they represent significant algorithmic add-ons beyond the vanilla independent RL paradigm (Jaques et al., 2019). The need for such extensions underscores the inadequacy of fundamental RL when agents must shape each other's behavior to foster cooperation.

These intertwined limitations, environmental non-stationarity, suboptimal equilibrium selection, heuristic exploration, and lack of opponent modeling, explain why independent RL agents in IPD often oscillate between transient strategies or converge to persistent defection unless augmented with centralized coordination, aligned priors, or bespoke exploration heuristics. Understanding and overcoming these barriers remains a central challenge in multi-agent learning and motivates the exploration of alternative frameworks,

such as active inference, which embeds uncertainty modeling, epistemic exploration, and planning as inference directly into the decision-making process.

2.3 Existing Applications

Applications of Active Inference to multi-agent and game-theoretic contexts have expanded in recent years, particularly in modeling social dilemmas like the Iterated Prisoner’s Dilemma. One of the most analytically rigorous contributions is from Demekas et al. (2023) in their paper "An Analytical Model of Active Inference in the Iterated Prisoner’s Dilemma". Their model demonstrates how two AIF agents configured with symmetric generative models and learning rates can transition from initial defection to mutual cooperation through iterative free energy minimization. Crucially, cooperation arises not from explicitly encoded strategies but through belief updating over transition dynamics. This progression is mathematically tied to the agents' learning rate (η), which controls the speed of Dirichlet learning over the agent's beliefs about how hidden states transition over time. The simulations presented by Demekas et al. (2023) demonstrate that when two Active Inference agents are configured with symmetric preferences and deterministic learning (i.e., no action stochasticity and identical learning rates), their behavior naturally converges to a pattern that is behaviorally equivalent to the classical Win-Stay, Lose-Shift (WSLS) strategy. This convergence is not due to any hardcoded heuristic or rule-following mechanism. Still, it arises organically from the agents’ continual minimization of expected free energy and belief updating via Dirichlet learning. In this setup, agents are more likely to repeat actions that previously led to preferred outcomes (i.e., "win"), and to switch actions when outcomes violate their expectations (i.e., "lose"). Importantly, this alignment with WSLS arises purely from the generative model's structure and the learning's statistical dynamics. The analytical clarity of this model, along with its ability to reproduce well-known cooperative strategies, makes it a foundational reference for the present thesis. It is a critical benchmark for exploring how Active Inference can be extended to more complex, multi-agent scenarios.

Together, these studies reveal a converging trajectory: whether grounded in Active Inference or not, successful models of long-term cooperation increasingly depend on probabilistic reasoning, dynamic structure, and adaptive response to uncertainty. The

current thesis builds on ideas of Demekas et al. (2023) by extending their AIF model to multi-agent simulations, additionally by adding several controllable parameters, investigating how cooperation emerges, stabilizes, or collapses when belief-driven agents interact in structured or stochastic environments. This expands the scope of Active Inference in social modeling, highlighting its capacity to unify learning and perception.

3. Theoretical Framework

3.1 Generative Models

In Active Inference, generative models are internal probabilistic structures that an agent uses to predict its environment, anticipate outcomes of actions, and update beliefs. These models specify the joint probability distribution over hidden states, observations, and actions. By encoding prior beliefs, transition dynamics, and expected observations, generative models allow agents to infer hidden causes of sensory inputs and select actions expected to fulfill their prior preferences (Parr et al., 2022, Ch. 2 & 4).

In the Iterated Prisoner's Dilemma, generative models become essential for capturing the latent dynamics of social interaction. Rather than relying on fixed strategy tables like Tit-for-Tat or WSLS, an Active Inference agent uses its generative model to infer whether its opponent will likely cooperate probabilistically and selects actions accordingly by minimizing expected free energy. This entails balancing epistemic value (resolving uncertainty) and pragmatic value (realizing preferred outcomes).

The analytical model introduced by Demekas and colleagues casts each participant in the two-agent Iterated Prisoner's Dilemma as an Active Inference–driven decision-maker employing a discrete-time Partially Observable Markov Decision Process (POMDP). In this setup, every agent maintains a generative model of its environment composed of several key components. First, an A-matrix captures how likely each possible sensory observation is, given the true hidden state of the world. Agents then rely on a B-matrix to represent the probabilities of transitioning from one hidden state to another; these transition beliefs are not fixed but are continually refined via Dirichlet learning as agents accumulate experience over successive rounds. Prior preferences over outcomes that the agent hopes or fears to observe are encoded in a C-vector, while the D-vector encodes the agent's beliefs about the hidden state at the very start of each trial. Initially, agents have uninformative,

uniformly biased priors across all hidden states and observations. As play unfolds, the learning rate parameter η governs how quickly and decisively those priors are adjusted. It determines the timing and the granularity with which agents update their transition beliefs and shape their emerging cooperative or competitive strategies.

Notably, the model uses a "memory-one" setup, meaning agents base their decisions mostly on the last round. This is similar to what Baek & Kim (2008) found; many successful strategies in the Prisoner's Dilemma rely on short-term memory. However, active inference agents don't stick to one rule, unlike strategies that always follow the same rule (like 'if they cooperate, then I cooperate'). Instead, they learn by estimating how likely different outcomes are based on their past experiences. This means they can make more flexible decisions instead of always doing the same thing. This helps them switch between cooperating and defecting more smoothly, depending on what they believe is best.

Akin (2015) emphasizes that strategic success in IPD requires more than mimicking surface behavior; it demands internal consistency and understanding of reciprocal structure. Generative models in AIF provide such structure, embedding expectations about reciprocity, transitions, and uncertainty into a coherent inferential system.

3.2 Learning Mechanisms

In Active Inference (AIF), learning is defined as the process of updating the parameters of the generative model such as those governing transition and observation probabilities, to reduce free energy across time (Parr et al., 2022, Ch. 6). More specifically, learning involves changing the agent's beliefs about how things change over time (the B matrix) and how observations are linked to hidden causes (the A matrix), based on new experiences. This process helps the agent reduce free energy over time, making its predictions more accurate and less surprising. As the agent gathers more evidence, its model improves, allowing it to adapt to new situations, plan better, and act more effectively in uncertain environments.

A key feature of AIF learning is its integration with perception and action, all following the same basic idea: reducing surprise. This means that when an agent updates what it believes about the world (perception), chooses what to do next (action), or changes how its model works (learning), it's all part of the same process. The agent adjusts its internal model

based on what it sees to stay better prepared for the future. Instead of using fixed rules or strategies, it keeps learning and adapting from experience (Parr et al., 2022).

This approach diverges sharply from reinforcement learning (RL), which typically relies on scalar reward signals and Q-value updates, without representing structural uncertainty. Sandholm and Crites (1996) and Gergely (2022) show that reinforcement learning (RL) agents often struggle in multi-agent settings like the IPD because other agents are also learning and changing, which makes the environment unstable and hard to predict. In contrast, Active Inference handles this challenge by allowing agents to represent and update their uncertainty about how the world works and which actions are best. This makes their behavior more stable, adaptive, and easier to understand over time.

In the analytical AIF model of the IPD by Demekas et al. (2023), learning is implemented via Dirichlet updates to the B matrix. Each agent maintains a probabilistic belief over state transitions conditioned on its own and its partner's previous actions. As the game progresses, the agent refines its model of the opponent's behavior not via trial-and-error optimization, but by accumulating evidence and updating beliefs about transition contingencies.

4. Methodology

4.1 Overview of Methodology

This chapter outlines the computational methodology used to simulate and analyze emergent cooperation in the Iterated Prisoner's Dilemma using the Active Inference framework. The simulations investigate how agents, equipped with generative models and driven by EFE minimization, can learn to cooperate over time without relying on fixed strategies or externally imposed reinforcement schedules.

Two simulation environments are implemented: two agents (2A-IPD) and three agents (3A-IPD). In both settings, agents repeatedly interact, observe outcomes, update their internal beliefs about hidden states, and select actions that minimize expected free energy. Each agent maintains a belief over hidden states, predicts the consequences of actions using a transition model, and encodes preferences over outcomes as softmax-normalized prior distributions. Agents select actions by applying a softmax over the negative expected

free energy (EFE), which combines a pragmatic risk term reflecting preference alignment and an epistemic term that promotes uncertainty reduction. This formulation follows the general Active Inference framework and its formal development in Parr et al. (2022).
Learning occurs through Dirichlet-based updates to the agent's transition model, allowing each agent to adapt its expectations over time. By systematically varying internal parameters such as learning rate, policy precision, and epistemic weighting, the simulations explore how different cognitive profiles and group sizes influence the emergence and stability of cooperation. Insights from prior work shape this chapter, Demekas et al. (2023), which explored Active Inference in the context of the Iterated Prisoner's Dilemma. The modeling approach also reflects core principles of the Active Inference framework (Friston et al., 2010; Parr et al., 2022). These works provided valuable context and inspiration.

4.2 Computational Modeling Approach

The computational approach employed in this study is based on modeling agents as active inference systems embedded within a repeated social dilemma framework, here the IPD. Two separate simulation environments were implemented: one with two interacting agents (2A-IPD) and one with three interacting agents (3A-IPD). In both cases, agents interact synchronously across discrete trials, selecting actions to minimize expected free energy (EFE) and updating their internal generative models based on experience.

4.2.1 Generative Model Components

Each agent in the simulation maintains an internal generative model that it uses to predict environmental dynamics, evaluate the consequences of actions, and guide behavior. This generative model is structured into four core components: the observation model (A), the transition model (B), the preference model (C), and the initial state prior (D). This section describes each component in detail. more details along with the matrix dimensions etc. can be found in Appendix B.

Observation Model (A): The observation model defines the likelihood of observing a particular sensory outcome given the hidden state. In these simulations, observations are assumed to be perfect and noiseless, meaning the observed outcome corresponds precisely

to the actual environmental state. Therefore, the A-matrix is implemented as an identity matrix. This setup ensures that the agent's posterior belief over hidden states (qs) collapses to a one-hot distribution after each observation, simplifying the inference process.

Transition model (B): The transition model in an Active Inference agent is captured by the B-matrix, which formalizes the agent's probabilistic understanding of how the hidden state of the environment evolves as a consequence of its own choices. In practical terms, for each combination of a current state (s) and an action (a) that the agent might take, the B-matrix provides a categorical distribution over the possible following states s′, written $B(s' \mid s, a)$. This compact representation is the basis of the agent's capacity to predict the environment's future course and evaluate the expected consequences of different policies.

The B-matrix is initialized with uniform probabilities across all admissible transitions to avoid imparting any unwarranted bias at the outset. In other words, before any experience is acquired, the agent treats each potential next state as equally likely for every state–action pair, reflecting maximal epistemic uncertainty. As the agent engages in repeated interactions, selecting actions according to its current policy, observing the resulting state transitions, and then looping back to update its beliefs, data gradually erodes this initial ignorance. The mechanism for belief revision is Dirichlet learning: each time the agent observes a transition from state s to state s′ under action a, it increments an associated pseudo-count for that particular transition. After adjusting these counts, it renormalizes them to yield a new categorical distribution. Thanks to the conjugacy between Dirichlet priors and the categorical likelihood, the posterior distribution over transition probabilities after each update remains Dirichlet, ensuring mathematical tractability and interpretability (Parr et al., 2022).

Throughout many rounds of interaction, this process sculpts the B-matrix into a structured map that accurately reflects the actual statistics of the environment's dynamics. Early on, when pseudo-counts are low, the agent's model remains diffuse, favoring exploration and epistemic actions. However, as more data accumulate, particularly if transitions follow consistent patterns, the pseudo-counts for frequently observed transitions grow, sharpening the agent's predictions and making its policy more exploitative. In effect, Dirichlet learning endows the Active Inference agent with a self-calibrating model of the world: one that begins in a state of deliberate ignorance, and then, through continual Bayesian updating,

transforms into an increasingly precise guide for both action planning and cooperative coordination.

Preference Model (C): The preference model, represented by the C-vector, encodes the agent's subjective evaluation of different hidden states regarding desirability or utility. Each entry in the C-vector reflects a raw scalar reward associated with a particular state, capturing the agent's internal valuation of possible outcomes (e.g., mutual cooperation versus unilateral defection). When designing payoff vectors for the Prisoner's Dilemma, one must respect the canonical ordering of rewards: temptation (T) > reward (R) > punishment (P) > sucker's payoff (S), and the additional "mutual‑cooperation bonus" condition, $2R > T + S$, which ensures that two cooperators fare better overall than an alternating sequence of exploitative moves (Lin et al., 2020). For both the two- and three-agent simulations, we fixed payoffs Reward Punishment so that they satisfy: $T > R > P > S$ and $2R > T + S$:

In our two-agent runs, we chose: R, S, T, P = 3.0, 0.5, 4.0, 1.0, so that defecting against a cooperator (T=4) beats cooperation (R=3), which in turn beats mutual defection (P=1), which itself beats being exploited (S=0.5), and $2\times3.0=6.0$ is greater than T+S=4.5.

In our three-agent setup, every possible combination of cooperations and defections is scored so that the classic Prisoner's Dilemma ordering ($T > R > P > S$) and the mutual‑cooperation bonus ($2R > T + S$) still hold. Cooperation by all three players yields a reward of R = 3, while mutual defection gives P = 1. Whenever exactly one player defects against two cooperators, that lone defector earns the temptation payoff T = 4, and the solitary cooperator in the opposite scenario receives the sucker's payoff S = 0.5. Every other mixed outcome—where two players' choices differ but there isn't a single clear exploiter or sole "sucker" is assigned an intermediate value of 2.5. This interpolation ensures that each agent's payoff vector across the eight joint‑action profiles faithfully generalizes the two-player PD structure to three participants.

For the two-agent simulations, we collapsed the classic Prisoner's Dilemma payoffs into a simple four-entry vector $[R,S,T,P] = [3.0, 0.5, 4.0, 1.0]$, where each entry corresponds respectively to the joint outcomes CC, CD, DC, and DD. In this scheme, defecting against a cooperator ($T = 4.0$) strictly dominates mutual cooperation ($R = 3.0$), which in turn beats

mutual defection (P = 1.0), which itself beats being exploited (S = 0.5). Moreover, the requirement 2R>T+S2R > T+S2R>T+S (i.e. 6.0 > 4.5) guarantees that sustained cooperation yields higher cumulative payoffs than an alternating exploiter–sucker sequence. By encoding preferences in this way, each agent's generative model naturally drives it toward cooperative choices once its beliefs over transitions become sufficiently confident.

Extending to three players introduces eight joint outcomes, but we preserve the same PD logic by assigning each agent a payoff of 3.0 when all three cooperate and 1.0 when all three defect. A lone defector facing two cooperators earns the temptation payoff of 4.0, while a single cooperator among defectors suffers the sucker's payoff of 0.5. All remaining mixed profiles, those without a unique exploiter or sole sucker, are given an intermediate reward of 2.5. These assignments yield each agent's C-vector across the eight states (CCC, CCD, CDC, CDD, DCC, DCD, DDC, DDD), for example Agent 1's preferences are: [3.0, 0.5, 0.5, 0.5, 4.0, 2.5, 2.5, 1.0], and analogous permutations for Agents 2 and 3. This construction ensures the canonical ordering T> R> P> S and the mutual-cooperation bonus 2R> T+S hold in every player's payoff structure, seamlessly generalizing the two-player PD to our three-agent setting.

To integrate these preferences into the probabilistic machinery of Active Inference, particularly in the computation of expected free energy, these raw values are transformed into a normalized probability distribution using a softmax function:

$$\mathrm{Pref}(s) = \frac{e^{C(s)}}{\sum_{s'} e^{C(s')}} \qquad (4.1)$$

This transformation yields a differentiable and normalized prior distribution over states, called Pref(s), which serves as the reference distribution in the Kullback–Leibler (KL) divergence term of expected free energy. The KL divergence measures how much the predicted future state distribution p(s'|a) diverges from the agent's preferred distribution Pref(s):

$$Risk(a) = D_{\mathrm{KL}}\left[p(s' \mid a) \| P_{\mathrm{ref}}(s)\right] \qquad (4.2)$$

This term penalizes actions expected to lead to outcomes that differ from the agent's internal preferences. When p(s'|a) closely matches Pref(s), the KL divergence is low; when

they differ significantly, the divergence increases. This way, the agent is incentivized to choose actions that steer expected outcomes toward subjectively valued states.

In this way, the C-vector shapes behavior by encoding what the agent expects or wants to experience, influencing action selection through anticipated outcomes rather than direct reward feedback. Different agents are assigned distinct C-vectors to reflect asymmetries in preference, which is significant in multi-agent settings like the Iterated Prisoner's Dilemma. For example, the states "cooperate-defect" (CD) and "defect-cooperate" (DC) may have opposite subjective values for each player, depending on their perspective in the interaction. These preference asymmetries are crucial for modeling role-specific strategic behavior and social dynamics.

Prior Belief (D): The prior belief over initial states is represented by the D-vector, which specifies the agent's assumptions about the probability of being in each possible hidden state before any observations are made. This prior is initialized as a uniform distribution across all states in the present simulations.

This setting reflects a state of maximal uncertainty, assuming all initial states are equally likely without prior evidence. The uniform prior ensures that no bias is introduced at the outset of inference, and that observed outcomes from the first trial entirely shape the agent's beliefs.

Although the D-vector is only used during the initial timestep before any observations, it plays a key role in the agent's Bayesian inference by combining with the observation model to yield posterior state beliefs:

$$q_s(s) \propto A(o_0, s) \cdot D(s) \quad (4.3)$$

In this formulation, the D-vector acts as the prior in Bayes' rule, while the observation model A is the likelihood. In future work or simulations with noisy or ambiguous observations (i.e., non-identity A-matrices), the D-vector could be adapted to encode informative priors or historical expectations about the environment's initial configuration.

Agents update their beliefs about the current state based on the most recent observation and their D-vector. Given the assumption of perfect observations (identity A-matrix), the

posterior belief collapses into a one-hot distribution centered on the observed state at each timestep.

4.3 Agent Behavior and Learning Algorithm

Each agent in the simulation operates as a belief-driven decision-making system. It selects actions by minimizing expected free energy (EFE) and updates its internal generative model based on observed outcomes. This section details the agent's inference, action selection, and learning processes..

4.3.1 State Inference

At each timestep, the environment observes the current game state (e.g., CC, CD, DC, or DD). The environment initializes the game on the first trial by randomly selecting a joint outcome from all possible states. This reflects the agent's initial uncertainty and corresponds to a uniform prior over states.

Upon receiving the observation, each agent updates its belief over hidden states using its prior (D-vector) and the observation model (A-matrix). Because we assume perfect perception, every observation tells the agent which state occurred. When the agent applies Bayes' rule using its uniform prior and an identity‑matrix likelihood, all the probability mass collapses onto the observed state. In other words, the agent's belief vector becomes "one-hot": it assigns probability one to the observed state and zero to all others. This ensures that, before planning or learning, there is no ambiguity about which joint action just took place. From the second time step onward, the environment's state is entirely determined by the actions taken by both agents (or all three in the 3A-IPD). The belief update remains deterministic: each observation directly specifies the new state, and the agent's belief is updated using Bayes' rule.

4.3.2 Action Selection via Expected Free Energy

The agent evaluates each possible option by computing its expected free energy (EFE) to choose an action. The EFE associated with an action a comprises two terms: a risk term and an epistemic value term.

The risk term is the KL divergence between the predicted future state distribution p(s′|a) and the agent's preferred distribution Pref(s). This term penalizes actions that are expected to lead to undesirable outcomes.

The epistemic value is the entropy of p(s′|a), which reflects how much uncertainty the action is expected to reduce.

The expected free energy is calculated as:

$$G(a) = D_{\mathrm{KL}}\left[p(s' \mid a) \| P_{\mathrm{ref}}(s)\right] - \gamma \cdot H\left[p(s' \mid a)\right] \quad (4.4)$$

Here, γ (gamma) is a positive scalar that controls the influence of the epistemic term.

The predicted distribution over the following states, p(s′|a), is computed by marginalizing over the agent's current belief $q_s(s)$ and its transition model B:

$$p(s' \mid a) = \textstyle\sum_s q_s(s) \cdot B(s' \mid s, a) \quad (4.5)$$

In the two-agent setting, $q_s(s)$ is a one-hot vector due to perfect observation (since A=I), so this reduces to a single column of the B-matrix. The same logic applies in the three-agent version, although the state's dimensionality and transition space increases.

Action selection is implemented by applying a softmax function over the negative expected free energy values:

$$P(a) \propto \exp(-\alpha \cdot G(a)) \quad (4.6)$$

where α (alpha) is the policy precision. A higher value of α results in more deterministic behavior, while a lower value increases stochasticity, allowing for more exploratory actions.

4.3.3 Transition Model Learning

After taking an action and observing the resulting new state, each agent updates its internal transition model using a Dirichlet learning rule. The model maintains a pseudo-count tensor denoted as B(s', s, a), which stores the agent's accumulated evidence about transitions from state s to state s', conditioned on action a.

At each timestep, the agent increments the pseudo-count corresponding to the observed transition by a small amount equal to the learning rate η (eta). Each time an agent sees a transition, it scales down all its previous transition counts by a factor of $(1 - \eta)$, adds η to the count for the newly observed transition, and finally renormalizes so everything sums to one. This makes recent transitions count more while gradually "forgetting" older ones, like an exponential moving average over past experiences.

After updating the pseudo-counts, the agent re-normalizes them to obtain a valid probability distribution over future states. This normalized B-matrix represents the agent's learned model of how the environment evolves in response to its actions.

Using fractional increments instead of full counts allows smoother and more gradual adaptation. The learning rate effectively controls how strongly new experiences influence the model: lower values lead to slower, more stable learning, while higher values allow faster adaptation but may increase overfitting.

This learning mechanism enables agents to build increasingly accurate transition models over time and supports flexible adaptation in uncertain or dynamic environments.

4.4 Experimental Design

The experimental design aims to systematically explore how different cognitive and behavioral parameters affect cooperation's emergence and stability in two-agent (2A-IPD) and three-agent (3A-IPD) Active Inference simulations. This section details the parameters varied, the structure of the simulation runs, and how cooperation and agent dynamics are measured.

4.4.1 Parameters and Conditions

Three core parameters control each agent's behavior:

- **Learning rate (η)**: Governs the speed of belief updating in the transition model via Dirichlet learning. Higher η results in faster adaptation but may overfit to recent

transitions.

- **Policy precision (α)**: Controls the determinism of the softmax action policy. Higher values make agents more deterministic in selecting actions that minimize EFE.

- **Epistemic weighting (γ)**: Controls the relative influence of entropy (uncertainty resolution) in the EFE computation. Higher γ promotes exploratory behavior.

Simulation runs systematically to vary these parameters and investigate how cooperation emerges across different cognitive profiles.

4.4.2 Simulation Setup

Each simulation consists of repeated trials in which agents observe the current state, infer beliefs, select actions, receive the resulting state, and update their models.

In the two-agent simulation, each agent chooses between “Cooperate” and “Defect.” The outcome state depends on the pair of actions. In the three-agent simulation, each agent still chooses between the same two actions, but the joint state space grows to 8 possible combinations (e.g., CCC, CCD, CDC, etc.). Transition learning and EFE computation are extended accordingly.

4.5 Evaluation Criteria and Reproducibility

This section outlines how the simulation outcomes are evaluated and how reproducibility is maintained across experimental runs.

4.5.1 Evaluation Criteria

The primary objective of the simulations is to investigate the emergence and stability of cooperation among agents under the Active Inference Framework. Quantitative and qualitative metrics evaluate agent behavior, internal learning dynamics, and strategic tendencies.

We assess our simulations through a combination of outcome‑level, behavioral, and internal‑state metrics illuminating how cooperation emerges, fluctuates, and stabilizes under Active Inference. At the most macroscopic level, we compute the final cooperation fraction, the percentage of rounds in which all agents simultaneously choose to cooperate, which gives a simple summary of collective success or failure. To unpack individual decision tendencies, we examine action‑selection frequencies, plotting how often each agent cooperates versus defects; this reveals whether any agent develops a strong bias toward one choice over the other.

Beyond these overt behaviors, we track the evolution of each agent's expected free energy (EFE) for cooperation and defection. By charting EFE as a function of trial number, we gain insight into the shifting balance between risk (pragmatic value) and uncertainty resolution (epistemic value) that underlies each decision. In parallel, we quantify learning progress by measuring the KL divergence between the agent's learned transition model and its initial uniform prior. The rising divergence suggests that an agent's Dirichlet counts converge towards specific state-action contingencies, establishing robust, experience-driven expectations.

Although Active Inference does not directly maximize scalar rewards, we also record the cumulative sum of preference values encountered during play to gauge how well agents' experiences align with their encoded utilities. To understand how performance depends on our key control parameters, we generate parameter‑sweep heatmaps that show mean cooperation rates across grids of learning rates, policy precision, or epistemic weighting; these visual summaries reveal the regions of parameter space most conducive to sustained cooperation. Finally, to connect our belief‑based framework with classical IPD heuristics, we mine agents' action histories for Win-Stay, Lose-Shift, and Tit-for-Tat–style motifs. Identifying these emergent patterns demonstrates whether simple inferential rules can reproduce the strategic signatures found in decades of IPD research. These complementary measures provide a rich, multi‑level portrait of how belief updating, risk assessment, and exploration–exploitation trade-offs conspire to produce cooperative dynamics under Active Inference.

4.6 Software, Libraries, and Computational Setup

The entire simulation framework was implemented in Python 3.10 using a modular architecture to ensure consistency and reproducibility. Core components such as the environment, agent logic, inference, learning, and visualization were organized into clearly separated modules.

All simulations were run on a 2020 Apple MacBook Air with an Apple M1 chip (8-core CPU) and 8 GB of unified memory. Due to the efficiency of the model architecture and the modest computational demands of the parameter sweeps, all experiments were executed using CPU-based computation without the need for high-performance hardware.

Random number generation was controlled using fixed seeds to ensure deterministic and replicable results across repeated runs. All key simulation parameters, including learning rates, action precision, epistemic weighting, and episode length, were explicitly defined, logged, and varied systematically during experimentation.

The project used the following open-source libraries:

- **NumPy 1.24** for numerical computation and matrix operations,
- **Matplotlib 3.7** for plotting all visual outputs (time series, histograms, heatmaps, etc.),
- **SciPy 1.10** for computing Kullback–Leibler divergence via scipy.special.rel_entr.

This setup enabled transparent, reproducible exploration of agent behavior under varied internal parameter configurations and interaction structures.

5. Results

This chapter presents the empirical findings from our Active Inference simulations of the Iterated Prisoner's Dilemma, beginning with the simplest two‑agent scenario. We first

demonstrate how two identical agents, each endowed only with generic priors and a shared learning rate, naturally progress through an “oscillate-defect-cooperate” sequence as they update their beliefs and refine their transition models over repeated play. This baseline establishes the characteristic learning trajectory that underlies all subsequent analyses.

Building on the canonical “oscillate-defect-cooperate” trajectory, we next investigate how each of our three cognitive parameters, learning rate (η), policy precision (α), and epistemic weighting (γ), shapes cooperation. To do so, we vary one parameter at a time while holding the other two at reference settings that we determined from pilot sweeps. In particular, we fixed $\eta = 0.6$ because it produces neither trivial collapse into universal defection nor runaway cooperation, but rather a balanced mix of outcomes that makes subtler effects visible. With η locked at 0.6, we then explore a range of α values, from low to high precision, to see how the noisiness of action selection boosts or suppresses mutual cooperation, settling on a default of $\alpha = 6$ for all subsequent experiments, since it lies in the mid‑range where cooperation is neither entirely random nor fully deterministic—finally, keeping both $\eta = 0.6$ and $\alpha = 6$. For all our η and α‑sweeps, we hold $\gamma = 0$ (i.e., no epistemic drive). Then, when it’s time to explore γ itself, we start at $\gamma = 0$, corresponding to zero explicit curiosity, and increase it stepwise, so we can directly observe how an agent’s intrinsic drive to resolve uncertainty promotes cooperation. At each stage, we display heatmaps of the mean cooperation rate, overlay one‑sided t-test markers where cooperation significantly exceeds our 0.60 threshold, and discuss the resulting patterns in depth.

Finally, we extend this parameter‑sweep protocol to the three‑agent environment, focusing exclusively on variations in learning rate. Holding action precision (α) and epistemic weighting (γ) at their pilot-justified reference values, we sweep η across the same grid and produce a corresponding series of heatmaps. Comparing these three-agent results to their two-agent counterparts allows us to isolate how increasing group size interacts with learning speed to promote or inhibit cooperation under Active Inference.

5.1 Baseline Two-Agent Learning Trajectory

Before delving into how particular parameters modulate cooperation, it is crucial to establish the canonical learning dynamics that emerge when two otherwise identical Active Inference agents play the IPD with no extra randomness or curiosity bonus (here, $\eta = 0.6$, $\alpha = 6$, $\gamma = 0$). Figure 5.1 presents the cumulative fraction of each joint action state over trials.

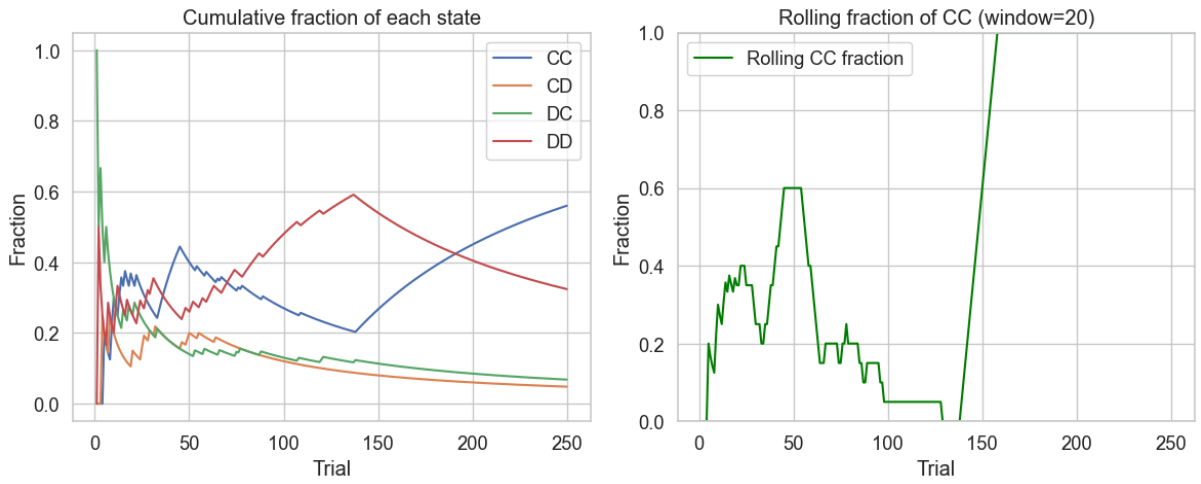

Figure 5.1 Cumulative fractions of CC, CD, DC, DD over trials

When we plot the cumulative state fractions for $\eta = 0.6$, $\alpha = 6$, and $\gamma = 0$, we see that after the first random move, the agents settle into an alternating pattern of cooperation and defection ("oscillation") for roughly the first fifty rounds. Following that, defection outcomes, especially mutual defection (DD), become more frequent for a time, reflecting how both agents' still-incomplete transition models briefly over-weight the safety of defection. Only after about 100 trials does the learning signal finally tip the scales: the agents' beliefs converge around the value of mutual cooperation, and CC quickly becomes the dominant outcome, climbing steadily to occupy over 50 percent of all plays by the final rounds. In other words, instead of a straightforward climb to cooperation, we observe an extended oscillatory phase, a secondary defection-dominant phase, and a final transition into sustained cooperation driven by each agent's ongoing belief updates and free-energy minimization.

It is instructive to examine the agents' expected free energy (EFE) over time for each policy to peek inside their reasoning as they learn to cooperate. Figure 5.2 plots, for a representative run (η=0.6, α=6, γ=0), the EFE that Agent 1 and Agent 2 assign to cooperation (blue) versus defection (red) at each trial.

What immediately stands out is the high degree of volatility in both curves during the first 50–60 trials. Neither agent has much evidence about how their opponent will respond at this early stage, so their Dirichlet counts over the transition model remain nearly uniform. In this “unknowing” state, defection habitually carries the lower expected free energy, reflecting that defecting is the safer, easier way to secure immediate reward with uniform priors and no epistemic bonus.

As play continues, however, each agent’s posterior over transition probabilities sharpens. We observe a series of brief inflection points in the blue curve, where the cooperation EFE momentarily dips below defection. These transient crossovers correspond exactly to the intermittent cooperative moves that give rise to the “oscillate” phase in their behavior. Each time an agent tests cooperation and sees it reciprocated, its model gains confidence that cooperation is not only possible but preferable, pulling the cooperative EFE downward.

Eventually, around trial 130 in this example, a clear and sustained separation emerges: the EFE for cooperation settles firmly below the EFE for defection, and both agents thereafter persist in cooperating. From this point onward, the two curves flatten out in parallel, illustrating that both agents have reached a stable consensus in their beliefs and, consequently, their policy evaluations.

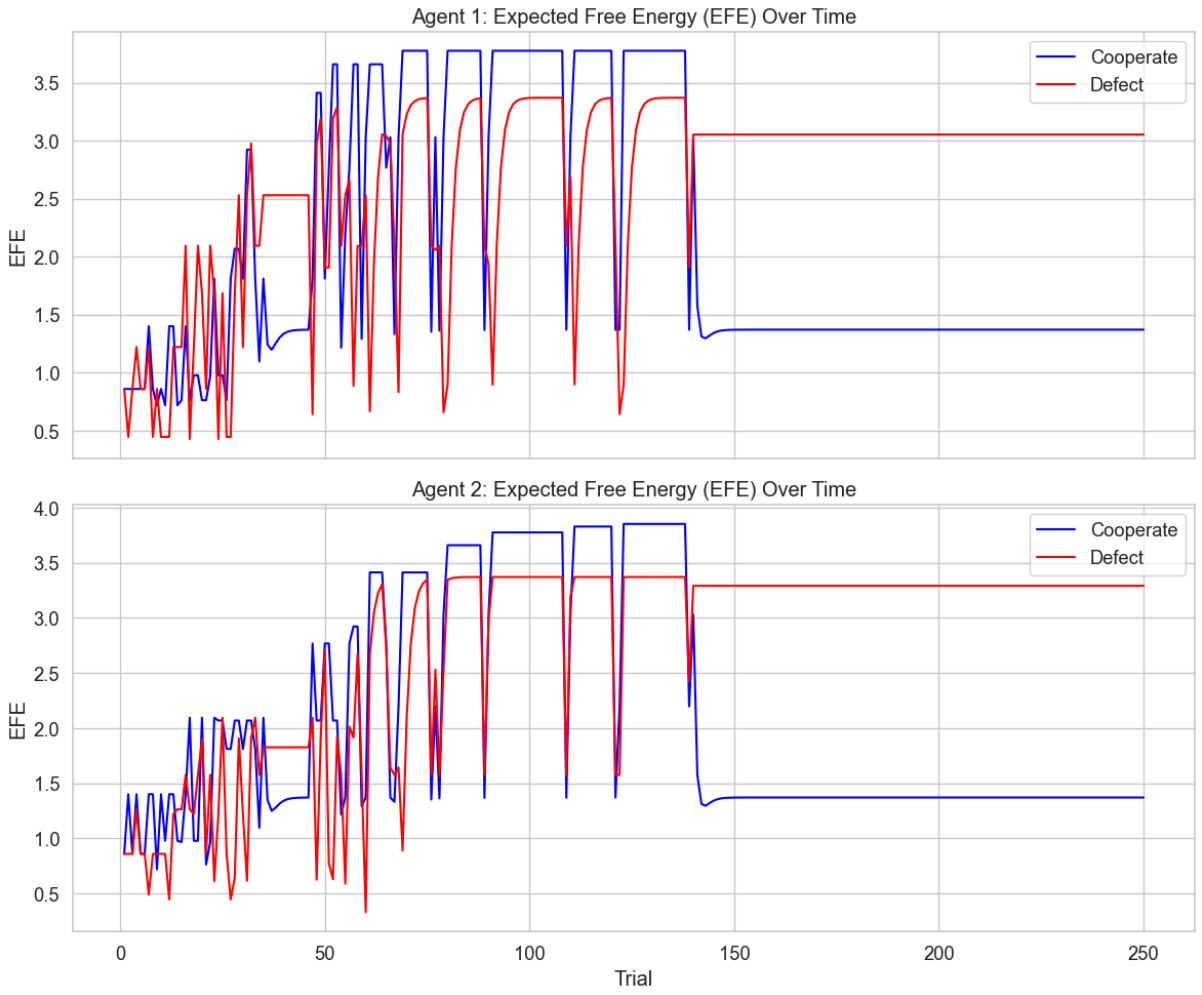

Figure 5.2 Expected Free Energy Trajectories for Both Agents

Having seen how belief updating transforms cooperation from a high-risk gamble into the lowest-cost policy, these EFE trajectories reveal the precise inferential mechanics at work: first a period of oscillatory "testing," then a phase dominated by defection as the model refines its priors, and finally a sustained turn toward cooperation once the cooperative EFE permanently undercuts defection.

Because we just saw how expected free energy (EFE) trajectories predict when each agent settles on cooperation, it's worth checking that those internal dynamics translate into different payoffs. In the first pair of panels, both agents share $\eta = 0.6$, $\alpha = 6$, and $\gamma = 0$. Their EFE curves swung into alignment simultaneously, and, not surprisingly, their cumulative rewards rose in lockstep once CC became dominant. By trial 140, the line graphs have straightened into a steady cooperative slope, and the final bar chart shows

nearly identical totals (within a few points) for Agent 1 and Agent 2. (Figure 5.3 and Figure 5.4)

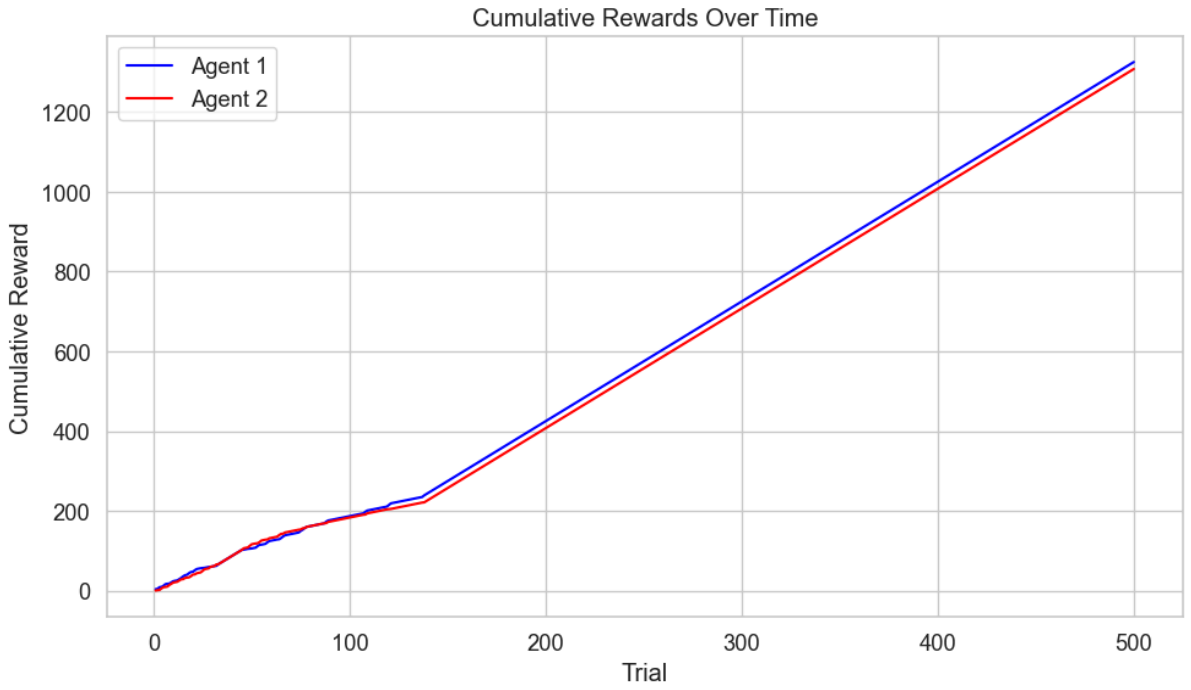

Figure 5.3 Cumulative Rewards Over Time

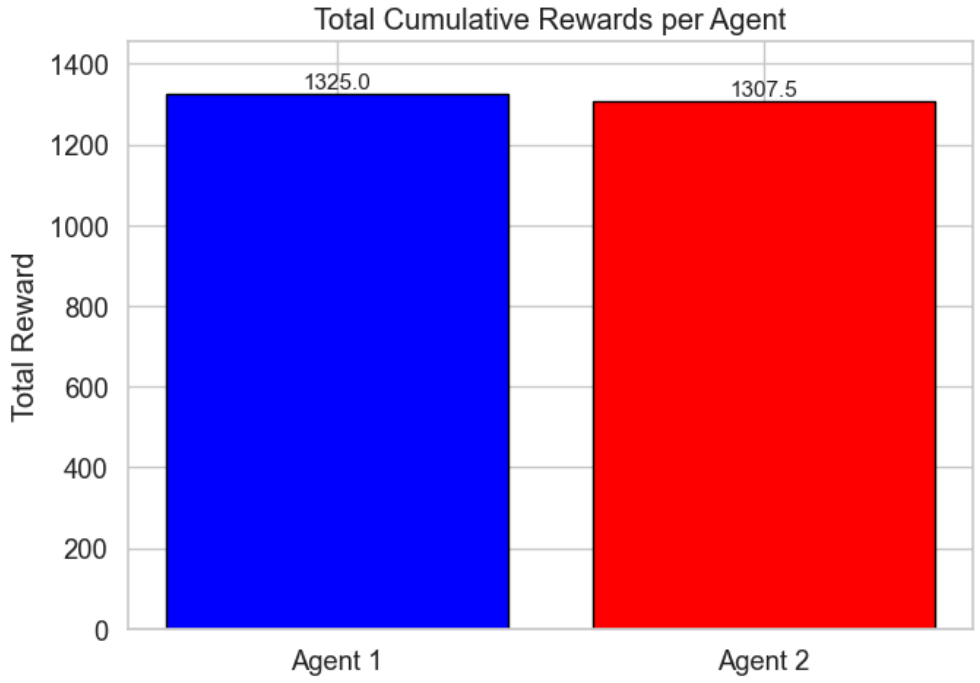

Figure 5.4 Total Cumulative Rewards per Agent

In the second experiment, everything stays the same except that Agent 2 learns more slowly ($\eta = 0.2$). This advantage for Agent 1 shows up directly in the reward curves: Agent 1's total climbs earlier and consistently outpaces Agent 2. The bar chart makes the gap plain: Agent 1 ends the run with a noticeably larger haul. (Figure 5.5 and Figure 5.6)

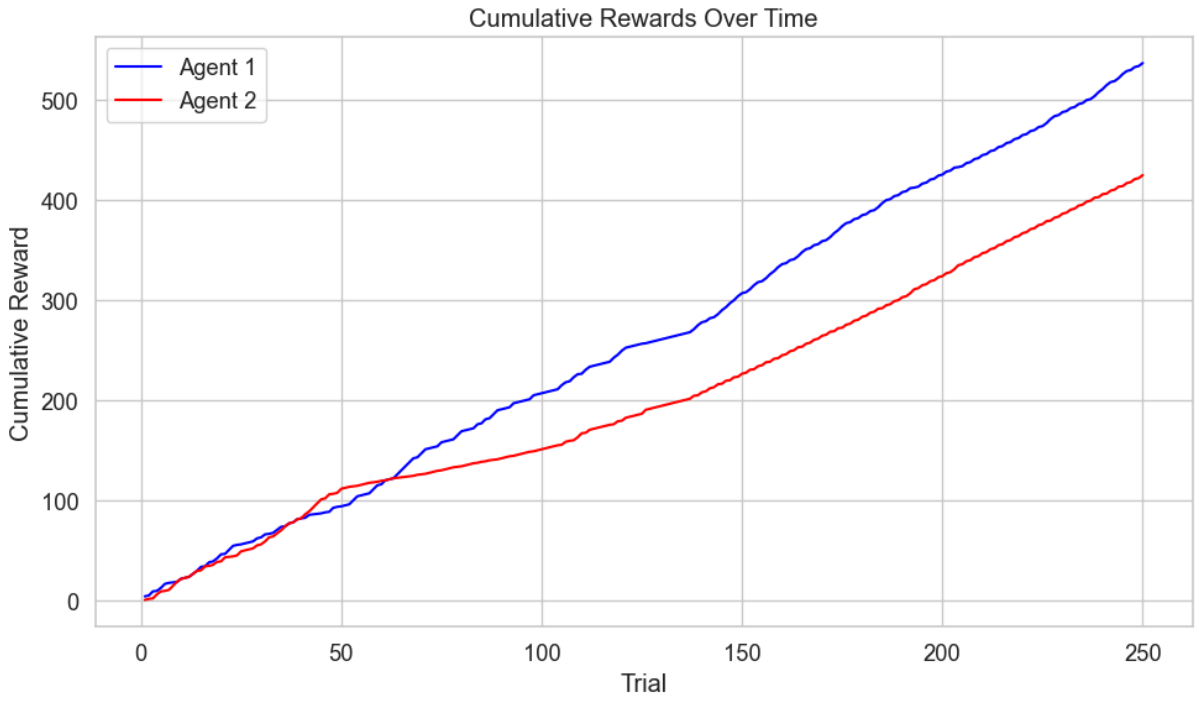

Figure 5.5 Cumulative Rewards Over Time

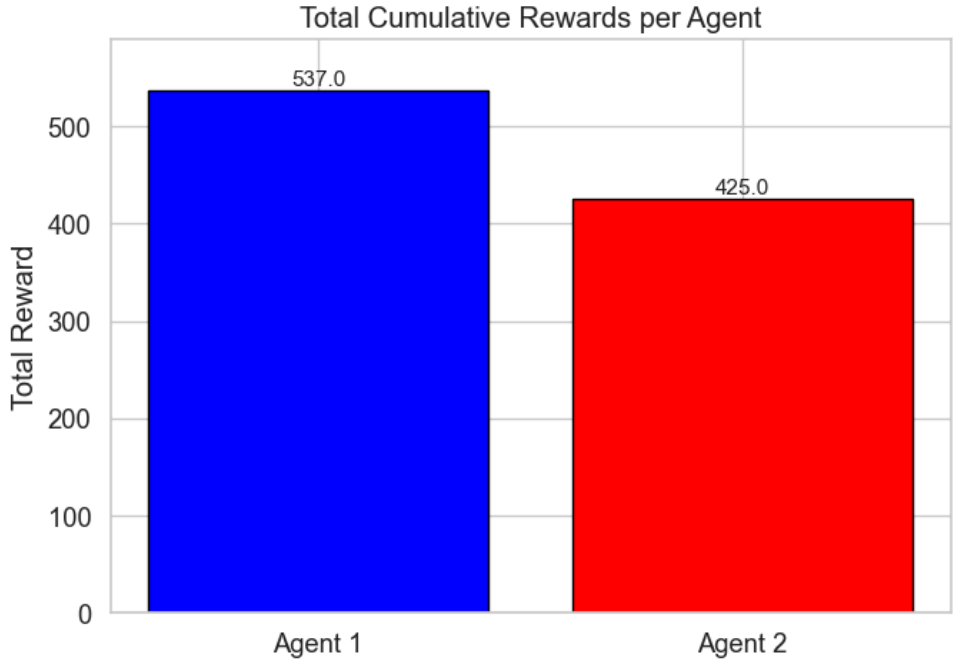

Figure 5.6 Total Cumulative Rewards per Agent

Putting these four plots side by side shows that when two agents infer and decide at the same speed, they almost earn the same rewards. A difference in learning rate alone, without touching any other decision parameters, shows how quickly an agent updates its beliefs under Active Inference, which can be just as consequential.

Both agents' learning trajectories, as measured by the KL divergence between their posterior Dirichlet parameters and the uniform prior, show the same basic pattern. Both divergences proliferate in the early rounds, indicating fast updating of expected state transitions, and then plateau once the agents have built accurate models of each other's behavior. The near-identical divergence profiles confirm that, under these homogeneous settings, both agents converge on the same internal model at approximately the same rate (Figure 5.7).

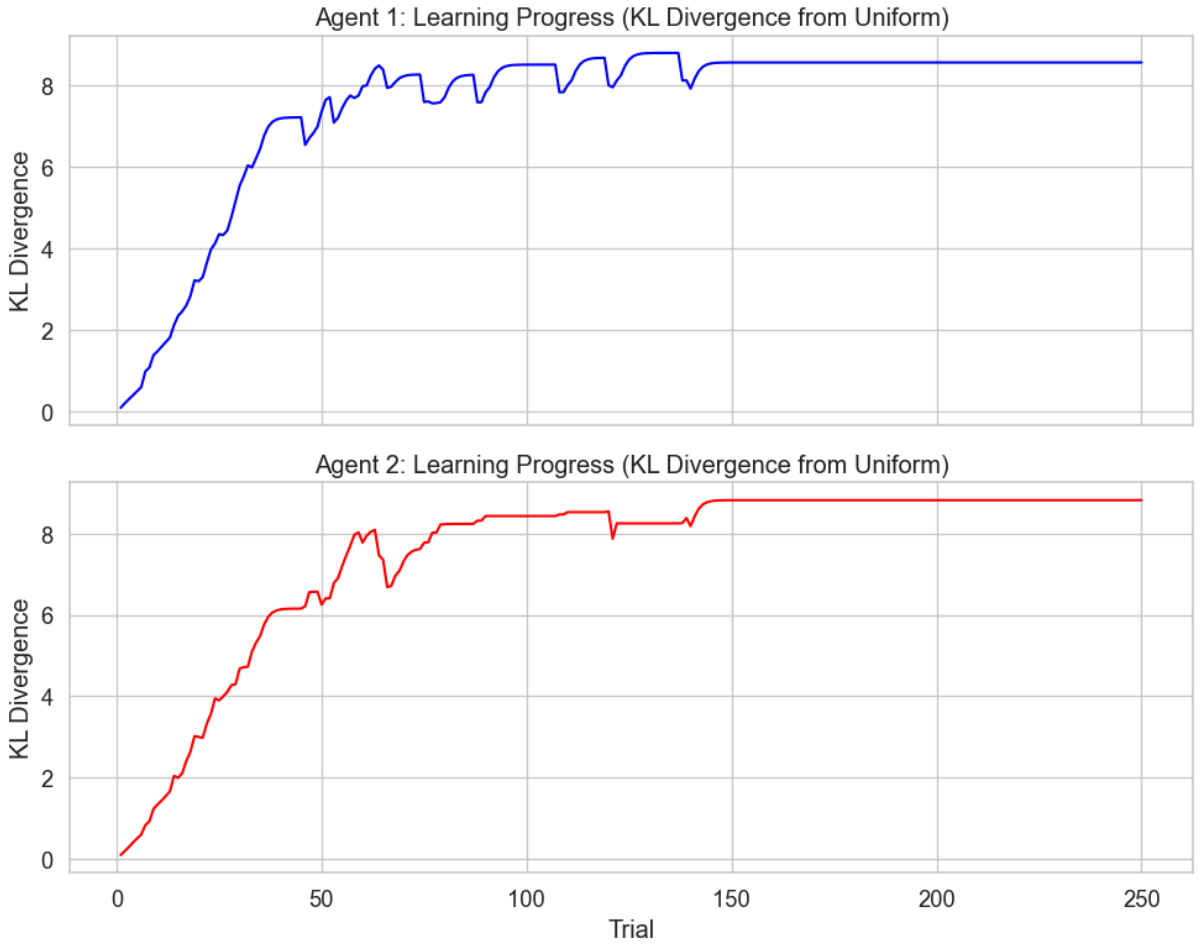

Figure 5.7 Learning Progress (KL Divergence from Uniform)

5.2 Mapping Two-Agent Cooperation Across Cognitive Parameters

In the next part of our study, we move beyond the single‑run dynamics and ask: How robust is cooperation when the agents' parameters are turned? In the two‑agent settings, we know that learning rate, policy precision, and epistemic drive each play a pivotal role, but how exactly do they shape the landscape of mutual cooperation? To answer this, we adopt a systematic sweep protocol. At each step, we hold two parameters at their pilot‑justified reference values (η=0.6, α=6, γ=0) and vary the remaining one across our predefined grid. This yields a family of heatmaps showing the mean joint‑cooperate rate over the η_1–η_2 plane (or α_1–α_2, γ_1–γ_2), with one‑sided t‑test markers highlighting regions where cooperation reliably exceeds our 0.60 threshold. We begin by charting the dependence on the learning rate itself, sweeping η_1 and η_2 before turning to policy precision and epistemic weighting. Together, these parameter sweeps reveal exactly where in parameter space Active Inference agents can sustain, amplify, or lose cooperative behavior in the two‑agent Iterated Prisoner's Dilemma.

5.2.1 Cooperation as a Function of Learning Rate

To see how robust the cooperation dynamic is when the two agents adapt at different speeds, we ran a grid sweep over their learning rates η_1 (along the x-axis) and η_2 (along the y-axis), holding both action precision (α=6) and epistemic weight (γ=0) constant, as demonstrated in the Figure 5.8. For each pair (η_1, η_2), we averaged the fraction of CC outcomes over 15 independent 500-trial runs. The resulting heatmap reveals a pronounced ridge of cooperation centered near the diagonal, indicating that agents who learn at similar rates are far more likely to settle into mutual cooperation than those whose learning speeds diverge.

Cooperation remains limited when either agent learns very slowly. In that regime, both agents' transition models linger close to their uninformative priors, making it impossible for them to form reliable expectations about one another's behavior; mutual defection thus predominates.

As both agents increase their learning rates into the moderate range, roughly between 0.4 and 0.7, their ability to align expectations improves dramatically. They accumulate enough experience to recognize and reinforce cooperative patterns, yet avoid overreacting to random fluctuations. At the sweet spot ($\eta_1 = \eta_2 \approx 0.6$), cooperation peaks, with agents sustaining mutual cooperation in nearly eighty percent of interactions. This plateau of high cooperation extends across a contiguous band of balanced learning rates, indicating that reciprocity and trust can emerge organically when both parties adapt at comparable speeds.

When learning rates become highly mismatched, the cooperative dynamic falters. A fast learner finds itself chasing the slow learner's outdated beliefs. In contrast, the slow learner remains stuck in stale assumptions and asymmetry that erode reciprocal expectations and drive cooperation rates back. Interestingly, cooperation partially rebounds when both agents adopt high learning rates (around 0.8 to 0.9).

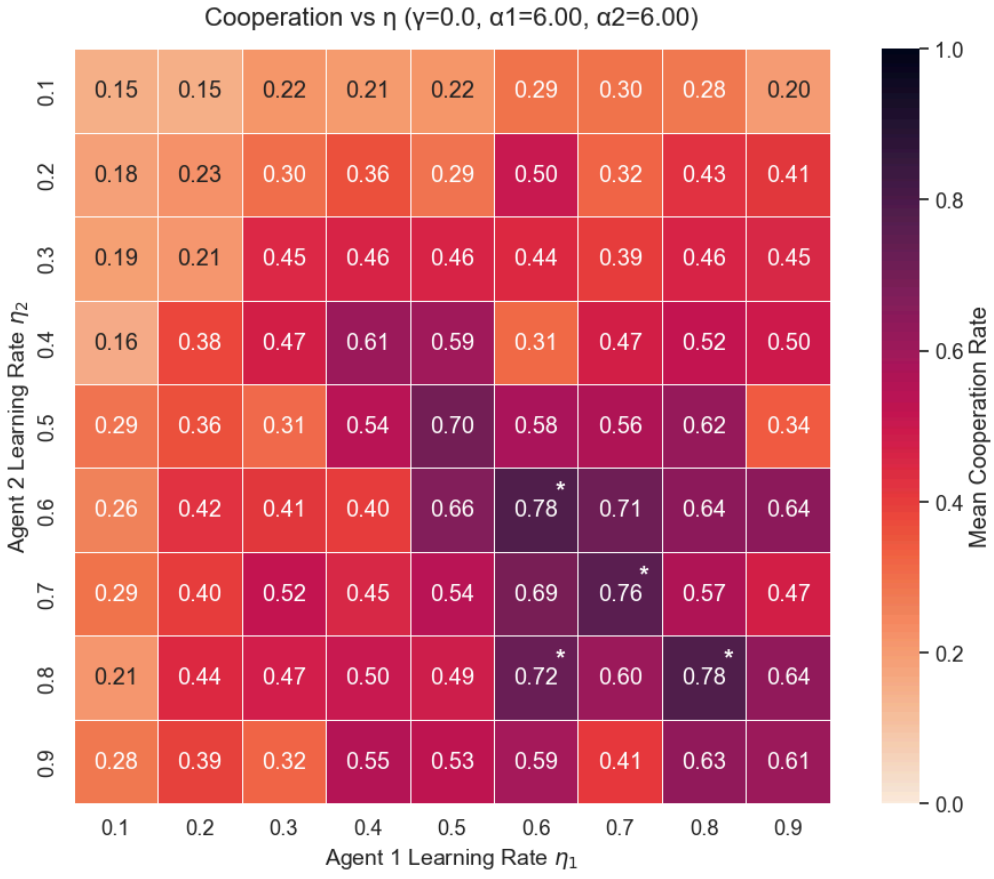

Figure 5.8. Cooperation Rate vs. Learning Rates (η_1, η_2). The heatmap illustrates how two Active Inference agents, each updating their transition beliefs via Dirichlet learning, negotiate

cooperation in a repeated Prisoner's Dilemma as a function of their learning rates, η_1 and η_2, under conditions of zero epistemic drive and high action-selection precision.

5.2.2 Cooperation as a Function of Policy Precision

Precision sweep (Figure 5.9) displays how cooperation rates vary as a function of agents' policy precision parameters (α_1 and α_2), with both agents sharing moderate learning rates ($\eta_1 = \eta_2 = 0.6$) and no explicit epistemic drive ($\gamma = 0$). Cooperation hovers around sixty to seventy percent along the lower edge of the precision spectrum when both α values lie around 5 or 6. In this regime, agents' choices retain a degree of stochasticity that allows them to align on cooperative moves occasionally, but the weak precision also leaves them vulnerable to occasional defection.

Cooperation peaks as both agents' precision parameters rise into the mid-range (approximately between 7 and 11). The single highest cooperation rate, around 0.79, occurs when $\alpha_1 \approx 7$ and $\alpha_2 \approx 8$. When the precision parameters become highly mismatched, one agent becomes very deterministic ($\alpha \approx 14$ or 15). At the same time, the other remains less precise; cooperation rates drop back into the forty-to-sixty percent range. The more deterministic agent locks in on its policy, leaving little room for responsive adaptation to a more exploratory partner. Interestingly, cooperation rebounds somewhat at the top end of the precision scale (both α around 14–15), climbing back toward the mid-sixty percent range. Both agents are nearly deterministic, so they tend to persist once a cooperative pattern is established. However, the lack of flexibility also leads to stubborn defection phases if expectations misalign.

Overall, these results underscore that Active Inference agents tuned to moderate-high and well-matched precision parameters can sustain robust cooperation without an explicit drive to explore—too little precision yields noisy interactions, while too much or mismatched precision risks brittle coordination. The sweet spot lies in the mid-range, where agents can reliably anticipate one another without becoming overly rigid.

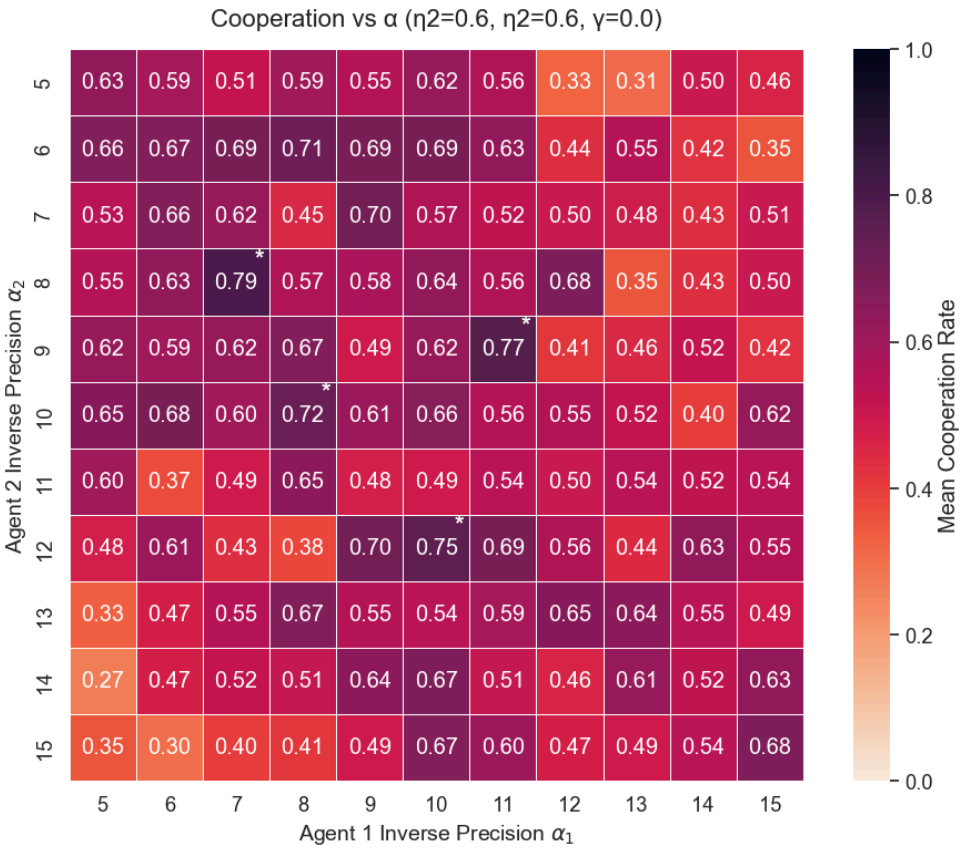

Figure 5.9 Cooperation Rate vs. Policy Precision (α_1, α_2)

5.2.3 Epistemic Weighting (γ) Sweep

The heatmap in Figure 5.10 reveals how agents' relative balance between goal‑directed exploitation and information‑seeking exploration, captured by their epistemic weightings γ, shapes cooperative dynamics when both learning rates (η=0.6) and decision precisions (α=6) are held constant. When neither agent places value on reducing uncertainty (γ≈0), cooperation is already quite strong, with mean rates around seventy percent, as they rely purely on updating transition beliefs through Dirichlet learning. Introducing a small dose of epistemic drive on one or both sides can further boost cooperation: the highest observed rate (~0.76) occurs when one agent's γ is very low (around 0.1). At the same time, the other is modest (around 0.4), and a similarly high plateau emerges for combinations in the region γ≈0.1–0.3. In this regime, limited curiosity helps agents refine their models just enough to reinforce mutual cooperation without destabilizing established expectations.

However, as both agents increase their epistemic weighting beyond this sweet spot, cooperation begins to wane. When γ climbs above roughly 0.5 for one or both agents,

mean cooperation rates drift toward the 0.50–0.60 range, indicating that excessive exploration can distract agents from consolidating cooperative norms. At the extreme, when both agents assign full weight to reducing uncertainty (γ=1), cooperation drops below fifty percent, reflecting a kind of over‑exploration in which agents continually probe rather than exploit stable reciprocal strategies. In addition, significant disparities in epistemic drive (for instance, an agent with γ≈0.9 paired with another at γ≈0.2) result in merely moderate cooperation, since the differing exploratory motivations diminish the common understanding expectations.

Overall, these results suggest that a balanced injection of epistemic value, enough to correct false beliefs but not so much as to chase novelty perpetually, can enhance cooperative behavior beyond what purely exploitative policies achieve. Too little uncertainty reduction prevents agents from discovering nuanced patterns in their partner's behavior, while too much leads to oscillatory dynamics that undermine the trust needed for sustained cooperation.

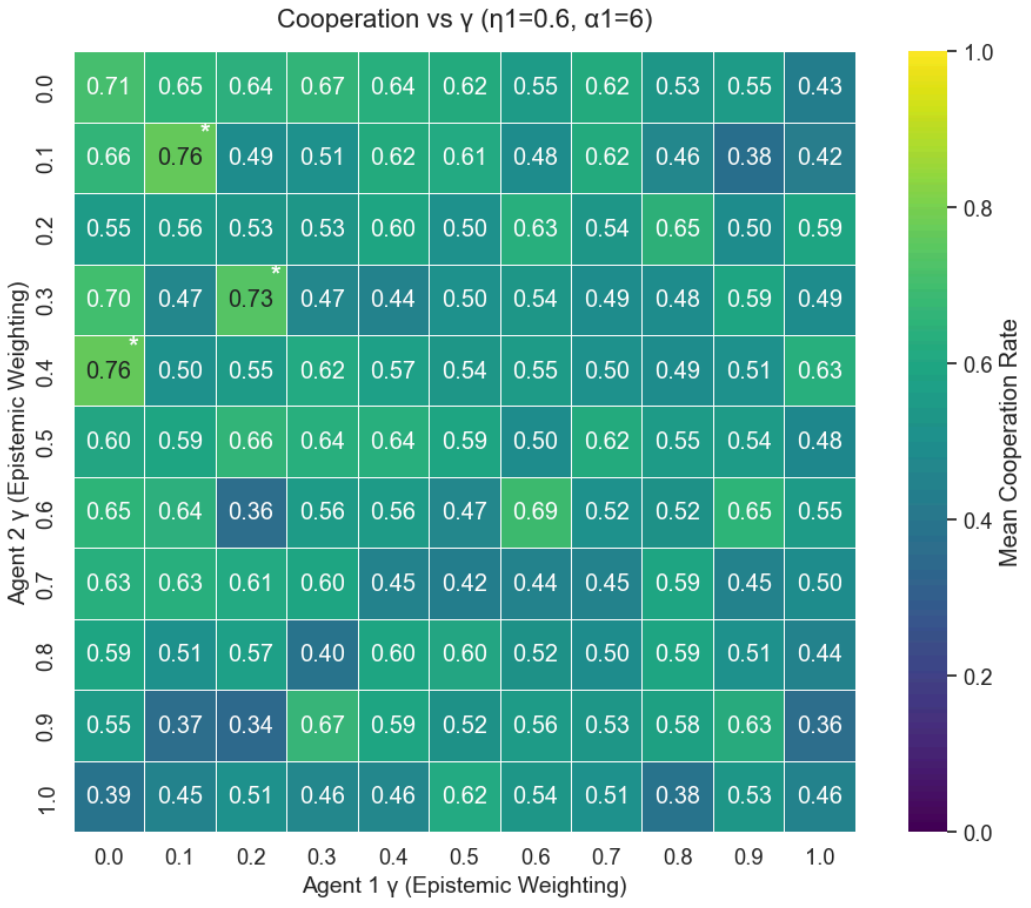

Figure 5.10 Cooperation Rate vs. Epistemic Weighting (γ_1, γ_2)

5.3 Emergent Strategy Patterns

To probe whether agents recover well-known IPD heuristics, such as Pavlov (Win–Stay, Lose–Shift) or reciprocal tit-for-tat tendencies, we computed the following conditional probabilities over trial sessions:

- **P(C | C)**: likelihood of cooperating given cooperation in the previous round
- **P(D | D)**: probability of defecting given defection in the previous round
- **Win–Stay**: probability of repeating the previous action when that action yielded a "win" (i.e., highest payoff)
- **Lose–Shift**: probability of switching actions after a "loss" (i.e., lowest payoff)

Across the four learning-rate conditions used in this experiment ($\eta = 0.1, 0.4, 0.6, 0.8$), we see a clear shift in how agents respond to both cooperative and defective outcomes and how reliably they adhere to "win-stay" and "lose-shift" heuristics relative to what would be expected by chance (0.5) (Table 5.1).

When learning is very slow ($\eta = 0.1$), agents rarely sustain cooperation after a cooperative move: the observed probability of cooperating given the partner cooperated sits at just 0.40, a value significantly below chance ($t = -9.36$, $p < 0.001$). By contrast, they are very consistent in reciprocating defection, with $P(D|D) \approx 0.65$ ($t = 10.03$, $p < 0.001$). Likewise, "win-stay" behavior is robust (mean = 0.61, $t = 6.48$, $p < 0.001$) and "lose-shift" remains above chance (mean = 0.57, $t = 4.26$, $p < 0.001$), suggesting that even sluggish learners still adhere to simple outcome-based rules, but struggle to build cooperative momentum.

With a moderate learning rate ($\eta = 0.4$), the probability of sustaining cooperation climbs to 0.59 (SEM = 0.047), though it narrowly misses statistical significance ($t = 1.90$, $p = 0.067$). Defection reciprocity falls slightly below the half-chance threshold (mean = 0.57, $p = 0.132$), indicating a breakdown of predictable punishment. At the same time, "win-stay" remains strong and statistically reliable (mean = 0.67, $t = 3.83$, $p = 0.001$), while "lose-shift" dips below chance (mean = 0.46, $p = 0.286$), suggesting that agents become somewhat less inclined to switch strategies following a loss.

When learning accelerates further (η = 0.6), agents begin to lock in on cooperative reciprocity once more: P(C|C) reaches 0.64 and just achieves significance (t = 2.05, p = 0.05), while P(D|D) falls markedly below chance (mean = 0.37, t = –2.68, p = 0.012), reflecting a surprising unwillingness to punish defection. Both "win‑stay" (mean = 0.67, t = 2.54, p = 0.017) and "lose‑shift" (mean = 0.37, t = –3.18, p = 0.004) are statistically significant, but now in opposite directions: agents tend to persist after wins yet are exceptionally slow to adapt after losses.

At the highest learning rate tested (η = 0.8), cooperative stickiness continues to strengthen—P(C|C) = 0.68, p = 0.013—while defection reciprocity remains low (P(D|D) = 0.34, p = 0.005). "Win‑stay" behavior is still above chance (mean = 0.66, p = 0.034), but "lose‑shift" shows no reliable deviation from 0.5 (mean = 0.44, p = 0.270).

Metric	η = 0.1	η = 0.4	η = 0.6	η = 0.8
P(C|C) mean	0.401	0.589	0.639	0.679
P(C|C) p-value	<0.001	0.067	0.050	0.013
P(D|D) mean	0.652	0.573	0.371	0.340
P(D|D) p-value	<0.001	0.132	0.012	0.005
Win–Stay mean	0.609	0.666	0.669	0.657
Win–Stay p-value	<0.001	0.001	0.017	0.034
Lose–Shift mean	0.573	0.463	0.368	0.437
Lose–Shift p-value	<0.001	0.286	0.004	0.270

Table 5.1 Mean strategy statistics and one‑sided p‑values for four behavioral metrics across learning rates

Taken together, these findings paint a nuanced picture: slow learners default to defection but still follow simple win‑stay/lose‑shift rules, moderate learners begin to cultivate cooperation albeit unevenly, and fast learners reinforce cooperation after mutual

cooperation yet become highly forgiving of defections persisting in "win-stay" while largely abandoning "lose-shift." This suggests that increasing the learning rate enhances cooperative stickiness at the expense of punitive consistency, and that the balance of these tendencies depends critically on how rapidly agents update their transition beliefs.

5.4 Three-Agent simulation

Here, we turn to our three-agent simulations and begin by examining how often the group settles into full cooperation over play. As in the two-player case (Sect. 5.1), we track the "cooperation fraction", the proportion of trials where all three agents choose cooperation across the run. This metric provides a straightforward readout of whether and when the network escapes the temptation to defect and converges on the socially optimal outcome. (Figure 5.11)

Because we've already seen (in Sect. 5.1) how homogeneous learning rates, high action-precision, and zero epistemic weighting lead two agents to oscillate-defect-cooperate, here we ask: under those same parameter settings ($\eta = 0.6$, $\alpha = 6$, $\gamma = 0$), does a trio of identical AIF agents manage to coordinate on CCC (all cooperate) state, or does the extra strategic uncertainty introduced by a third player inhibit that transition? The plot below shows the trial-by-trial cooperation fraction for our canonical three-agent configuration.

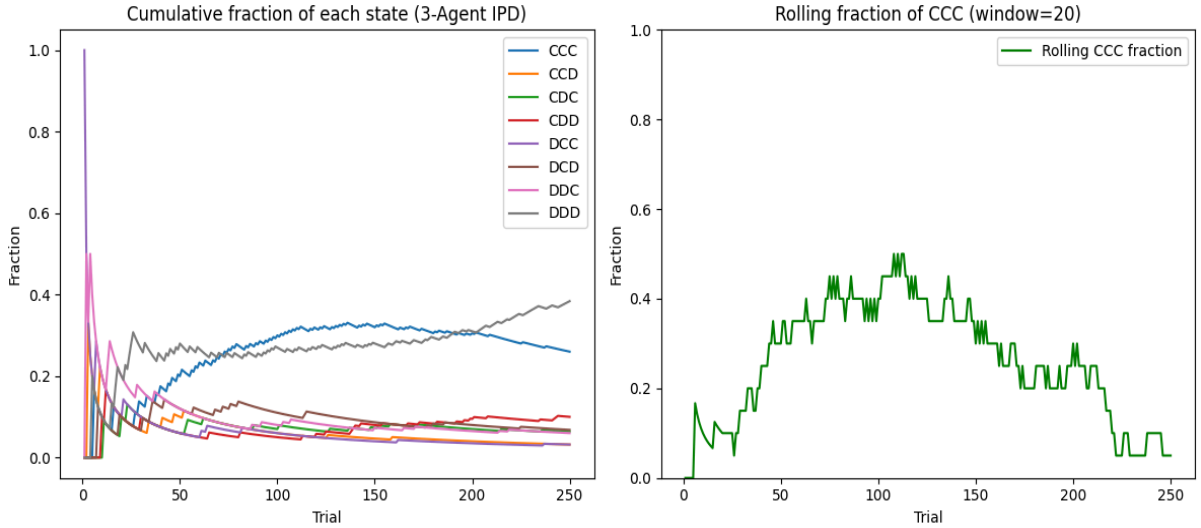

Figure 5.11 Cumulative and Rolling Triple Cooperation Dynamics (3-Agent IPD)

When three identical agents (η=0.6, α=6, γ=0) play the iterated dilemma, the early rounds are dominated by one-off mixed outcomes, particularly states where one agent defects against two cooperators, but these quickly give way to a slow build-up of full cooperation (CCC). CCC becomes one of the most frequent patterns by mid-experiment, though it never completely eclipses the other outcomes. In the latter stages, cooperation gradually loses ground as mass defection (DDD) climbs back to prominence.

CCC's rolling-window view shows this arc: cooperation starts rare, rises to a clear peak in the middle trials, and then declines toward the end. In sum, three-way coordination can emerge under these parameters, but it remains transient, with collective defection ultimately reclaiming dominance.

With only the inverse-precision doubled from 6 to 12, the three-agent system rapidly locks into cooperation (Figure 5.12). In the cumulative plot, all eight asymmetric and defection-heavy states that dominated the early trials are swept away by trial 50, and "CCC" rises to well over 80 % of all outcomes by mid-experiment. "DDD" collapses towards zero in perfect synchronicity.

The rolling-window view makes this even clearer: CCC climbs steeply from its first appearances to hit and sustain a perfect 1.0 fraction for long stretches between trials 60 and 180 (with only a brief dip back to ≈0.85). In short, increasing α made the agents more decisive in favor of cooperation, producing a stable, virtually unshakable cooperative equilibrium.

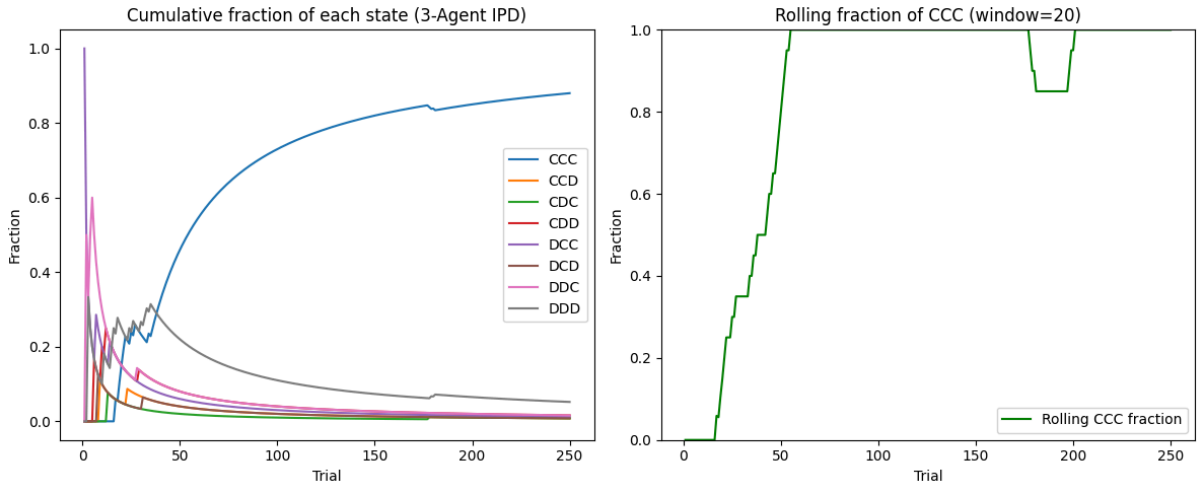

Figure 5.12 Cumulative and Rolling Triple Cooperation Dynamics (3-Agent IPD)

5.4.1 Learning rate sweep in three-agent simulation

In the three‑agent extension of our iterated Prisoner's Dilemma, each player's ability to learn from past outcomes can dramatically reshape the group's cooperative dynamics. In the two‑agent case, cooperation often emerges smoothly, but adding a third learner introduces new tensions: one player's learning speed can either reinforce or undermine the others' inclinations to cooperate. In this section, we explore how varying the learning rate η_3 of the third agent modulates the overall cooperation landscape as a function of the first two agents' learning rates (η_1 and η_2). The resulting contour plots provide a clear visual map of these spots of cooperation, highlighting how moderate pacing in one learner can upset otherwise stable cooperative outcomes.

As demonstrated in Figure 5.13, you notice that at η_3 low (0.10) the entire plane is pale yellow, indicating that when the third agent learns very slowly, triadic cooperation rarely takes hold regardless of the other two rates. η_3 modest (0.30): A small patch of greenish-blue emerges near the mid-range of η_1 and η_2, showing that moderate learning by the third agent lets all three reach cooperation but only in a narrow band of learning-rate combinations. η_3 intermediate (0.50): That cooperative patch widens and deepens; now a broader swath of $\eta_1 \approx 0.4–0.6$ and $\eta_2 \approx 0.3–0.7$ sustains higher CCC fractions, while edges remain yellow. η_3 higher (0.70): Cooperation peaks most sharply: a pronounced dark-blue appears at mid-range η_1/η_2, revealing that faster but not too fast learning by the third agent optimally aligns all three. η_3 very high (0.90): The blue region disappears once more toward just a few areas, indicating that when the third agent learns almost immediately, it again disrupts coordination except in limited parameter combinations.

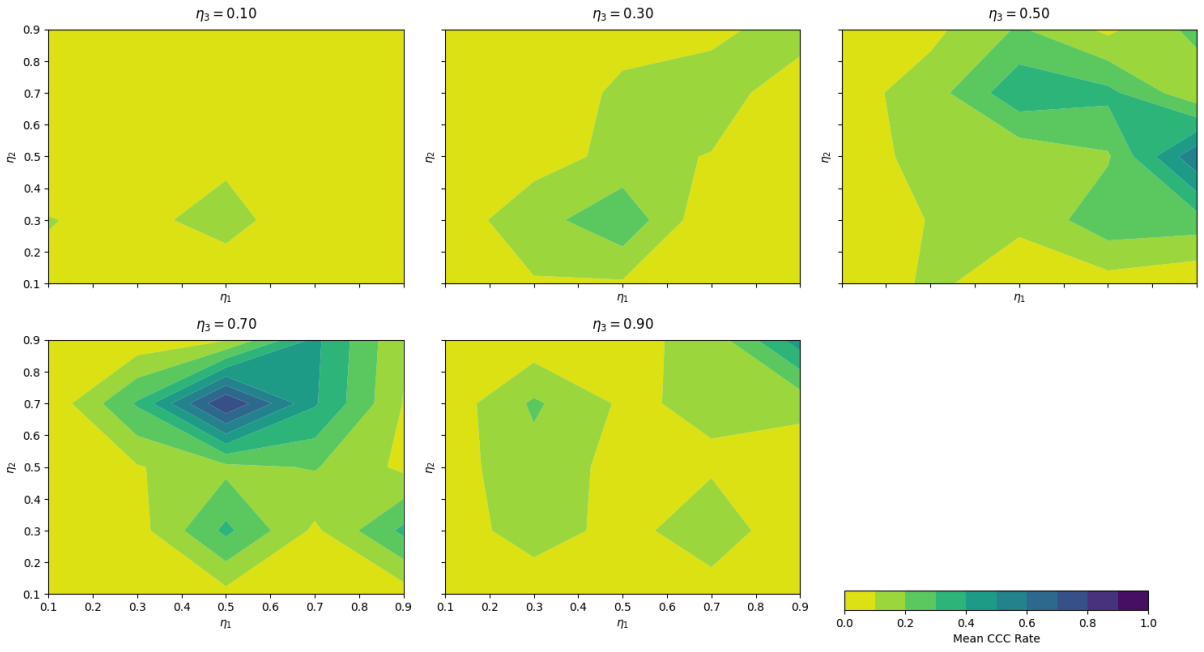

Figure 5.13 Three-Agent Cooperation Rate Across Learning‑Rate Anchors

6. Discussion

In this chapter, we interpret our simulation results in the context of Active Inference theory and prior work on cooperation in repeated social dilemmas. We first revisit the baseline two‑agent dynamics, then unpack how each cognitive parameter, learning rate (η), policy precision (α), and epistemic weighting (γ), shapes cooperative outcomes. We examine the effects of scaling from two to three agents, and finally reflect on broader implications, limitations, and avenues for future research.

6.1 Revisiting Baseline Two-Agent Dynamics

Our canonical two-agent simulation under η = 0.6, α = 6, and γ = 0 produced a characteristic three-phase progression: an initial oscillatory interplay of cooperation and defection, a transient surge in unilateral defection, and finally sustained mutual cooperation. In the early trials, agents' beliefs about the transition model (B-matrix) remain near their uniform priors, rendering both "cooperate" and "defect" similarly plausible. As a result, CD and DC outcomes alternate under maximally uncertain EFE calculations. As the Dirichlet learning continues, agents develop sharper expectations. As the trials progress, the risk term of EFE begins to favor cooperation modestly, triggering an increase in CC events. Consequently, CC frequency ramps up rapidly. Agents effectively discover that

cooperation best satisfies their C-vector preferences while minimizing free energy. The cooperation rate fluctuates around a stable level.

As agents cooperate more often, their total reward (payoffs) rise faster than during the earlier, more exploratory phase. At first, mutual cooperation (CC) is rare, so testing defection still looks "safe" in the EFE calculation. But once enough CC trials have occurred, the extra reward from each new cooperative turn outweighs the occasional benefit of exploring defection. From that point onward, the agents' free-energy minimization consistently favors cooperation, and we see their cumulative reward curves steepen exactly when CC frequency jumps (Figure 5.3). This positive feedback locks in a high‑cooperation regime for the rest of the simulation.

6.2 The Role of Learning-Rate Symmetry

When we swept η_1 and η_2 jointly, a marked prominence of intensive coordination emerged along the diagonal where $\eta_1 \approx \eta_2$. Outside this band, cooperation collapsed into persistent defection, prolonged oscillation. This demonstrates that symmetry in the pace of belief updating is critical: if one agent learns much faster, it "gets ahead" of its partner, over‑committing to a transition model that the slower learner cannot yet track which breaks the conditions for the typical oscillatory period that leads to cooperation (CC), which occurs when agents are cooperate simultaneously (Demekas et al., 2023). The result is a mismatch in predicted payoffs and an inability to lock into reciprocal CC.

When one agent learns much faster than its partner, its transition model "outpaces" the slower learner's beliefs, so the two never build a shared expectation of what comes next, and cooperation breaks down. In contrast, when both agents update at similar rates, their B‑matrices drift together, allowing mutual predictions to align and cooperation to take hold. In other words, it isn't how quickly agents learn in isolation, but how closely their learning speeds match that determines whether they can lock into a cooperative equilibrium.

6.3 Action-Precision and the Exploration–Exploitation Trade-off

The action‑precision sweep revealed an inverted‑U pattern in mean cooperation rates: when α is very low (around 5–6), agents remain overly stochastic, so exploratory behavior dominates and there is insufficient exploitation of nascent cooperative cues, keeping CC rates near 0.60. As α increases into the mid‑range (7–12), cooperation peaks, mean CC rises into the 0.70 – 0.83 band because the softmax becomes sharp enough to capitalize on emerging free‑energy gradients while allowing occasional exploration to avoid premature lock‑in. However, pushing α beyond about 12 makes action selection overly deterministic: agents simply repeat whichever action enjoyed a slight early advantage, losing the flexibility needed to adapt to beneficial cooperation norms or to break defection stalemates. From an Active Inference standpoint, α scales the influence of negative expected free energy on choice probabilities, so that small values flatten the distribution (excessive exploration) and large values collapse it (excessive exploitation). Our findings align with Parr et al. (2022), who argue that precision tuning is essential for striking the right exploration–exploitation balance in uncertain environments.

6.4 Epistemic Weighting: Curiosity Versus Stability

Varying epistemic weighting (γ) revealed that some epistemic drive (γ≈0.2–0.4) accelerates the discovery of CC by favoring uncertainty-reducing actions, leading to higher peak cooperation. However, excessive epistemic weight ($\gamma > 0.5$) destabilized the cooperative regime: agents over-prioritized information gain at the expense of exploiting known cooperative payoffs, resulting in renewed oscillation or defection. Conversely, $\gamma = 0$ (our baseline) yielded slower but more stable cooperation.

These results underscore the dual role of EFE: balancing pragmatic risk (alignment with C-preferences) and epistemic value (uncertainty reduction). Real-world agents likely require moderated curiosity; too little and they fail to explore beneficial norms, too much and they never settle.

6.5 Emergence of Classical Heuristics

Across the range of learning rates we examined, a clear pattern emerges in how agents anchor their simple strategies to the dynamics of cooperation and defection. When learning is very slow (η=0.1), agents rarely sustain cooperation, P(C|C) sits well below chance, yet they reliably punish defection and adhere firmly to win-stay/lose-shift rules. As the learning rate increases to a moderate level (η=0.4), we see a marked increase in cooperative reciprocity and continued strength in win-stay behavior, albeit with less predictable punishment after defection and a weakening of lose-shift. With elevated learning rates (η=0.6 and 0.8), agents exhibit increased "stickiness" in cooperation: P(C|C) substantially exceeds chance levels, win-stay remains strong, while P(D|D) drops below chance, leading to the near-complete disappearance of lose-shift. In other words, quicker updates enhance cooperative inertia, sacrificing consistent retaliation or strategy changes following a loss. These results suggest that increasing the speed of belief updating helps agents lock in cooperation. Still, it also makes them overly forgiving, reducing the effectiveness of punitive or corrective responses. For modeling social dilemmas, this implies that tuning the learning rate is crucial: too slow, and cooperation never takes hold; too fast, and agents lack the flexibility to correct misunderstandings, potentially leaving them vulnerable to exploitation or coordination failures.

6.6 Scaling Up: Three-Agent Dynamics

Adding a third agent under the same baseline parameters (η=0.6, α=6, γ=0) reduced sustained cooperation: CCC fractions peaked briefly mid-run but collapsed to DDD. This highlights that larger groups introduce strategic uncertainty that two‑agent pairs can easily overcome.

However, increasing α to 12 reinstated long‑lived CCC coordination: higher decisiveness counteracts the combinatorial explosion of state possibilities in three-player IPD. Learning-rate sweeps in the three-agent setting showed a narrower η_3 band (≈0.5–0.7) that supports cooperation, misaligned η_3 again derails group synchrony. These findings mirror empirical observations that cooperation declines with group size (Martínez-Martínez & Normann, 2022) unless cognitive parameters are finely tuned.

6.7 Broader Implications, Limitations, and Future Directions

Together, our results demonstrate that Active Inference provides a principled mechanism for emergent cooperation: agents need only maintain and update simple generative models, with no need for prescriptive strategy tables or external reward shaping. Identifying "sweet spots" in η, α, and γ offers actionable guidelines for designing artificial agents in cooperative tasks from swarm robotics to decentralized economic models.

Despite the breadth of our simulations, several vital limitations temper the generality of these findings. First, we treated each agent's C-vector—their payoff preferences—as fixed throughout every run. In many real-world dilemmas, however, an individual's valuation of cooperation versus defection can evolve with experience, reputation effects, or changing stakes. Allowing preferences to adapt or become state-dependent would more faithfully capture the dynamic motivational landscapes that shape social behavior.

Second, our experiments remain confined to small, homogeneous groups; only two or three agents, each sharing the same cognitive parameters. Scaling to larger, heterogeneous populations will introduce combinatorial growth in the joint state-action space and more complex interaction patterns, demanding more efficient inference algorithms or structured approximations.

Finally, although we swept each cognitive parameter (learning rate, policy precision, epistemic weighting) individually and plotted the results, we did not perform full factorial sweeps that vary all three simultaneously. Such high-dimensional exploration could reveal critical interactions or non-linear regime parameter combinations that encourage cooperation only when tuned together and remain hidden under one-at-a-time analyses. Addressing these gaps represents a clear path forward for future Active Inference models of multi-agent cooperation.

We can simplify our next steps in three ways for future work. First, let agents' preferences change over time instead of staying fixed, so they can learn to value cooperation more after good experiences or be punished for defection. Second, move beyond just two or three identical agents: try larger groups with different learning speeds and decision noise, using more innovative inference methods to keep computation manageable. Third, instead of

tweaking one parameter at a time, run full‑grid sweeps over learning rate, precision, and curiosity; this will show us how these factors interact to produce cooperation. Finally, testing our model against human data would be valuable, linking η, α, and γ to objective measures of how fast people learn, how noisy their choices are, and how much they seek new information.

7. Conclusion

In this thesis, we set out to explore how agents driven purely by Bayesian belief‑updating and free‑energy minimization can learn to cooperate in the Iterated Prisoner's Dilemma. Building a simple Active Inference framework where agents maintain and update a transition model (B‑matrix), hold fixed payoff preferences (C‑vector), and select actions via a softmax over negative expected free energy. We showed that cooperation can emerge without hard‑coded strategies or external reward shaping.

Our main findings can be summed up as follows. First, two‑agent simulations reveal a robust "oscillate–defect–cooperate" progression: early exploration gives way to a temporary surge in unilateral defection before agents lock into sustained mutual cooperation. Second, sweeping the learning rate reveals that matched learning speeds are critical when agents update at similar rates, synchronize their beliefs, and achieve higher cooperation; severe mismatches push them into persistent defection. Third, action‑precision exhibits an inverted‑U relationship, where intermediate precision best balances exploration and exploitation to maximize collaboration. Finally, across our sweeps, Pavlovian Win–Stay, Lose–Shift behavior arises organically from Dirichlet updating plus softmax choice, but it is strongest in the low-to-moderate learning regime. When η is very high, beliefs solidify so quickly that agents stop reacting reliably to losses, and the full heuristic fades. Extending our framework to three agents showed that larger groups can still cooperate, but only under suitably tuned learning rates, and that maintaining belief‑alignment across more players becomes increasingly challenging.

Together, these results establish Active Inference as a principled, unified approach to modeling cooperation under uncertainty. Rather than prescribing behavior, AIF allows cooperative norms to self‑organize through continuous inference and preference‑driven exploration. This opens up promising avenues for studying social decision‑making in more complex or realistic settings by allowing preferences to adapt, scaling to larger and more

diverse populations, and jointly tuning multiple cognitive parameters. Grounding these simulations in human behavioral data would further bridge theory and experiment. In sum, belief‑driven learning under the Free Energy Principle reproduces well‑known strategic patterns and offers new insights into how cooperation can emerge and stabilize in dynamic social environments.

References

Akin, E. (2015). What you gotta know to play good in the iterated prisoner's dilemma. *Games*, *6*(3), 175–190. https://doi.org/10.3390/g6030175

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396. https://doi.org/10.1126/science.7466396

Baek, S. K., & Kim, B. J. (2008). Intelligent tit-for-tat in the iterated prisoner's dilemma game. *Physical Review E*, *78*(1). https://doi.org/10.1103/physreve.78.011125

Barker, J. L. (2017). Robert Axelrod's (1984) The Evolution of Cooperation. In *Springer eBooks* (pp. 1–8). https://doi.org/10.1007/978-3-319-16999-6_1220-1

Claus, C., & Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. *National Conference on Artificial Intelligence*, 746–752. https://www.aaai.org/Papers/AAAI/1998/AAAI98-106.pdf

Demekas, D., Heins, C., & Klein, B. (2023). An analytical model of active inference in the iterated prisoner's dilemma. In *Communications in computer and information science* (pp. 145–172). https://doi.org/10.1007/978-3-031-47958-8_10

Foerster, J., Nardelli, N., Farquhar, G., Afouras, T., Torr, P. H. S., Kohli, P., & Whiteson, S. (2017). Stabilising experience replay for deep Multi-Agent Reinforcement learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1702.08887

Fogel, D. B. (1993). Evolving behaviors in the iterated prisoner's dilemma. *Evolutionary Computation*, *1*(1), 77–97. https://doi.org/10.1162/evco.1993.1.1.77

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active Inference: a process Theory. *Neural Computation*, *29*(1), 1–49. https://doi.org/10.1162/neco_a_00912

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*(1–3), 70–87. https://doi.org/10.1016/j.jphysparis.2006.10.001

Friston, K., Lancelot, D. C., Hafner, D., Hesp, C., & Parr, T. (2020). Sophisticated inference. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2006.04120

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214. https://doi.org/10.1080/17588928.2015.1020053

Galesic, M., Barkoczi, D., Berdahl, A. M., Biro, D., Carbone, G., Giannoccaro, I., Goldstone, R. L., Gonzalez, C., Kandler, A., Kao, A. B., Kendal, R., Kline, M., Lee, E., Massari, G. F., Mesoudi, A., Olsson, H., Pescetelli, N., Sloman, S. J., Smaldino, P. E., & Stein, D. L. (2023). Beyond collective intelligence: Collective

adaptation. *Journal of the Royal Society Interface*, *20*(200). https://doi.org/10.1098/rsif.2022.0736

Gergely, M. I. (2022). Finding Cooperation in the N-Player Iterated Prisoner's Dilemma with Deep Reinforcement Learning Over Dynamic Complex Networks. *Procedia Computer Science*, *207*, 465–474. https://doi.org/10.1016/j.procs.2022.09.101

Grujić, J., Cuesta, J. A., & Sánchez, A. (2012). On the coexistence of cooperators, defectors and conditional cooperators in the multiplayer iterated Prisoner's Dilemma. *Journal of Theoretical Biology*, *300*, 299–308. https://doi.org/10.1016/j.jtbi.2012.02.003

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P. A., Strouse, D., Leibo, J. Z., & De Freitas, N. (2018). Social influence as intrinsic motivation for Multi-Agent Deep Reinforcement Learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1810.08647

Kaufmann, R., Gupta, P., & Taylor, J. (2021). An active inference model of collective intelligence. *Entropy*, *23*(7), 830. https://doi.org/10.3390/e23070830

Martinez-Martinez, I., & Normann, H. (2022). Cooperation in multiplayer dilemmas. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4294660

Montero-Porras, E., Grujić, J., Domingos, E. F., & Lenaerts, T. (2022). Inferring strategies from observations in long iterated Prisoner's dilemma experiments. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-11654-2

Nowak, M., & Sigmund, K. (1990). The evolution of stochastic strategies in the Prisoner's Dilemma. *Acta Applicandae Mathematicae*, *20*(3), 247–265. https://doi.org/10.1007/bf00049570

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), 56–58. https://doi.org/10.1038/364056a0

Parr, T., Pezzulo, G., & Friston, K. J. (2022). Active inference. In *The MIT Press eBooks*. https://doi.org/10.7551/mitpress/12441.001.0001

Raihani, N. J., & Bshary, R. (2011). Resolving the iterated prisoner's dilemma: theory and reality. *Journal of Evolutionary Biology*, *24*(8), 1628–1639. https://doi.org/10.1111/j.1420-9101.2011.02307.x

Sandholm, T. W., & Crites, R. H. (1996). Multiagent reinforcement learning in the Iterated Prisoner's Dilemma. *Biosystems*, *37*(1–2), 147–166. https://doi.org/10.1016/0303-2647(95)01551-5

Stewart, A. J., & Plotkin, J. B. (2012). Extortion and cooperation in the Prisoner's Dilemma. *Proceedings of the National Academy of Sciences*, *109*(26), 10134–10135. https://doi.org/10.1073/pnas.1208087109

Tan, M. (1993). Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents. In *Elsevier eBooks* (pp. 330–337). https://doi.org/10.1016/b978-1-55860-307-3.50049-6

Wedekind, C., & Milinski, M. (1996). Human cooperation in the simultaneous and the alternating Prisoner's Dilemma: Pavlov versus Generous Tit-for-Tat. *Proceedings of the National Academy of Sciences*, *93*(7), 2686–2689. https://doi.org/10.1073/pnas.93.7.2686

Appendix

Appendix A: Simulation and Model Details

All code used for simulation, analysis, and figure generation in this thesis is publicly available at: Project files

This appendix provides the core functions and definitions to simulate the Iterated Prisoner's Dilemma (IPD) under the Active Inference framework for the two-agent simulation. The following code snippets define the environment, the agent class, and the helper routines needed to run and analyze simulations. Use the GitHub code to access the full code for both two- and three-agent simulations.

A.1 IPD Environment Definition

```
import numpy as np

class IPDEnvironment:
    """
    Iterated Prisoner's Dilemma Environment:
    - States are encoded as integers:
        0: CC (both cooperate)
        1: CD (agent1 cooperates, agent2 defects)
        2: DC (agent1 defects, agent2 cooperates)
        3: DD (both defect)
    - The step function updates the state based on the two agents' actions.
    """
    def __init__(self):
        self.state = None
        self.reset()

    def reset(self):
        # Initialize to a random state for the first trial
        self.state = np.random.choice(4)
        return self.state

        def step(self, a1, a2):
        # Map action pair (a1,a2) to next state
        if   (a1 == 0 and a2 == 0): self.state = 0
        elif (a1 == 0 and a2 == 1): self.state = 1
```

```
        elif (a1 == 1 and a2 == 0): self.state = 2
        else:                       self.state = 3
        return self.state
```

A.2 Active Inference Agent Class

```
import numpy as np
from scipy.special import rel_entr  # For KL divergence

class ActiveInferenceAgent:
    """
    Active Inference agent for the IPD.
    - Learns transition probabilities (B) using Dirichlet pseudo-counts (pB).
    - Observes states via a deterministic likelihood (A = identity).
    - Selects actions by minimizing full Expected Free Energy:
        EFE = Risk - gamma * EpistemicValue
    """
    def __init__(self, name, C, alpha=4.0, eta=0.3, gamma=1.0):
        self.name = name
        self.alpha = alpha    # Softmax precision (action-selection sharpness)
        self.eta = eta        # Learning rate for Dirichlet updates
        self.gamma = gamma    # Weight on epistemic value

        # Prior preferences over states (softmax-normalized)
        self.C = C
        self.Pref = np.exp(self.C)
        self.Pref /= np.sum(self.Pref)

        # Dirichlet counts and transition model
        self.pB = np.ones((4, 4, 2))  # Prior pseudo-counts
        self.B  = np.zeros((4, 4, 2)) # Normalized transition probabilities
        self._normalize_B()

        # Observation model: perfect mapping from state to observation
        self.A = np.eye(4)

        # Prior over latent states (uniform)
        self.D = np.ones(4) / 4
        self.qs = self.D.copy()  # Posterior belief over states
```

```
def _normalize_B(self):
    # Normalize Dirichlet counts to probabilities
    for old_s in range(4):
        for action in range(2):
            counts = self.pB[:, old_s, action]
            self.B[:, old_s, action] = counts / (np.sum(counts) + 1e-15)

def observe_and_infer_state(self, obs_state):
    # Update beliefs about hidden state given new observation
    likelihood = self.A[obs_state, :]
    posterior = likelihood * self.D
    self.qs = posterior / (np.sum(posterior) + 1e-15)

def compute_negEFE_for_action(self, a):
    """
    Compute negative Expected Free Energy for action a:
    negEFE = -(Risk - gamma * EpistemicValue)
    """
    eps = 1e-8
    # Predict next state distribution
    pred_states = sum(self.qs[s] * self.B[:, s, a] for s in range(4))
    # Risk = KL(pred_states || Pref)
    risk = np.sum(pred_states * (np.log(pred_states + eps) - np.log(self.Pref + eps)))
    # Epistemic value = entropy of predicted observations
    pred_obs = self.A.dot(pred_states)
    entropy = -np.sum(pred_obs * np.log(pred_obs + eps))
    # Return negative EFE for use in softmax
    return -(risk - self.gamma * entropy)

def select_action(self):
    # Softmax decision rule over negative EFE
    negEFE_C = self.compute_negEFE_for_action(0)
    negEFE_D = self.compute_negEFE_for_action(1)
    logits = np.array([negEFE_C, negEFE_D]) * self.alpha
    probs = np.exp(logits) / np.sum(np.exp(logits))
    return np.random.choice([0, 1], p=probs)

def update_transition_model(self, old_state, action, new_state):
    # Apply decay to Dirichlet counts
    self.pB[:, old_state, action] *= (1 - self.eta)
    # Increment count for observed transition
    self.pB[new_state, old_state, action] += self.eta
    # Renormalize to get updated B
```

```
        self._normalize_B()
```

A.3 Simulation Helper Functions

```
import numpy as np

# Runs one IPD simulation and returns action histories
# Uses global defaults: default_N, default_alpha1/2, default_gamma

def run_simulation_with_history(eta1, eta2, gamma):
    agent1 = ActiveInferenceAgent("A1", C=np.array([3,0.5,4,1]),
                                  alpha=default_alpha1, eta=eta1, gamma=gamma)
    agent2 = ActiveInferenceAgent("A2", C=np.array([3,4,0.5,1]),
                                  alpha=default_alpha2, eta=eta2, gamma=gamma)
    env = IPDEnvironment()

    a1_hist, a2_hist = [], []
    s_old = env.reset()

    # Initial random moves
    for t in range(default_N):
        if t == 0:
            a1, a2 = np.random.choice([0,1]), np.random.choice([0,1])
        else:
            agent1.observe_and_infer_state(s_old)
            agent2.observe_and_infer_state(s_old)
            a1, a2 = agent1.select_action(), agent2.select_action()
        s_new = env.step(a1, a2)
        agent1.update_transition_model(s_old, a1, s_new)
        agent2.update_transition_model(s_old, a2, s_new)
        a1_hist.append(a1); a2_hist.append(a2)
        s_old = s_new

    return np.array(a1_hist), np.array(a2_hist)

# Computes four behavioral metrics from action histories

def compute_tft_wsls(a1, a2):
    payoff = {(0,0):3, (0,1):1, (1,0):4, (1,1):2}
    cc, dd = 0, 0
```

```
cc_tot, dd_tot = 0, 0
win_stay, lose_shift = 0, 0
win_tot, lose_tot = 0, 0

for t in range(1, len(a1)):
    prev = (a1[t-1], a2[t-1])
    curr = a1[t]
    # Pavlovian tit-for-tat counts
    if prev[1] == 0:
        cc_tot += 1
        cc += int(curr == 0)
    else:
        dd_tot += 1
        dd += int(curr == 1)
    # Win-Stay/Lose-Shift counts
    own_payoff = payoff[prev]
    alt_payoff = payoff[(prev[1], prev[0])]
    if own_payoff >= alt_payoff:
        win_tot += 1
        win_stay += int(curr == a1[t-1])
    else:
        lose_tot += 1
        lose_shift += int(curr != a1[t-1])

return {
    "P(C|C)":    cc/cc_tot    if cc_tot    else np.nan,
    "P(D|D)":    dd/dd_tot    if dd_tot    else np.nan,
    "Win-Stay":  win_stay/win_tot  if win_tot  else np.nan,
    "Lose-Shift": lose_shift/lose_tot if lose_tot else np.nan
}
```

Appendix B: Generative Model and Inference Equations

In this appendix, we present the mathematical foundations of the Active Inference simulations used in this thesis, including: 1) the generative POMDP model; 2) Bayesian state inference; 3) learning via Dirichlet updates; and 4) action selection via Expected Free Energy (EFE).

B.1 Generative Model

Two-Agent IPD Generative Model

We model the Iterated Prisoner's Dilemma (IPD) as a discrete-time observable Markov decision process with four latent states $s \in 0, 1, 2, 3$ (CC, CD, DC, DD) and two actions $a \in 0, 1$ (cooperate, defect). The generative model comprises:

- **Observation model** A: a 4×4 likelihood matrix $A_{o,s} = P(o \mid s)$, here taken as identity ($A = I$) to reflect perfect observation of the latent state.
- **Transition model** $B \in \mathbb{R}^{S \times S \times A}$, where $B_{s',s,a} = p(s_{t+1} = s' \mid s_t = s,\, a_t = a)$. Here $S = 4$ and $A = 2$, so B has shape $(4 \times 4 \times 2)$.
- **Preference (C-vector)** $C \in \mathbb{R}^S$, with one raw "log-preference" entry per state. For two agents we set:

  $$C = \big[R,\, S,\, T,\, P\big] \;=\; [3.0,\; 0.5,\; 4.0,\; 1.0]$$

  so that $T > R > P > S$ and $2R = 6 > T + S = 4.5$.

- **Prior over initial states (D-vector)** $D \in \mathbb{R}^S$, here uniform:

  $D = \frac{1}{4}\,[1,\, 1,\, 1,\, 1]^\top$, reflecting complete uncertainty at the start of the simulation.

Three-Agent IPD Generative Model

When we extend to three interacting players, there are now $S = 2^3 = 8$ possible joint outcomes, which we index in order as:

$$\{\,\mathrm{CCC},\ \mathrm{CCD},\ \mathrm{CDC},\ \mathrm{CDD},\ \mathrm{DCC},\ \mathrm{DCD},\ \mathrm{DDC},\ \mathrm{DDD}\}\,.$$

Each agent still has two actions (cooperate = 0 or defect = 1), but the latent-state space grows to size 8. The generative model for each agent i is:

- **Observation model** $A \in \mathbb{R}^{8 \times 8}$, again the identity matrix, so $p(o{=}j \mid s{=}i) = \delta_{ij}$.
- **Transition model** $B \in \mathbb{R}^{8 \times 8 \times 2}$, with entries $B_{s',s,a_i} \;=\; p\big(s_{t+1} = s' \;\big|\; s_t = s,\; a_{i,t} = a_i\big)$, so B has shape $(8 \times 8 \times 2)$.

- **Preference (C-vector)** $C \in \mathbb{R}^8$, one scalar per joint outcome. We assigned $C = [\,3.0,\ 0.5,\ 0.5,\ 0.5,\ 4.0,\ 2.5,\ 2.5,\ 1.0\,]$,

  so that the single-defector payoff is $T = 4$, the sole-sucker S=0.5, mutual cooperation $R = 3$, full defection $P = 1$, and all other mixed profiles receive the intermediate value 2.5. These choices still respect $T > R > P > S$ and $2R > T + S$.

- **Prior over initial states (D-vector)** $D \in \mathbb{R}^8$, uniform:

  $D = \frac{1}{8}\,[1, 1, 1, 1, 1, 1, 1, 1]^\top$.

In both cases, the agent begins with uniform B and D ;maximally uncertain, then refines B via Dirichlet updating (with decay) and selects actions by minimizing Expected Free Energy under its learned generative model.

B.2 Bayesian State Inference

At each trial, after observing $o_t = s_t$, the agent updates its posterior belief $q(s_t)$ according to Bayes' rule:

$q(s_t) = P(s_t \mid o_t) \propto A_{o_t, s_t} D_{s_t}$

and normalizes so that $\sum_s q(s_t) = 1$

B.3 Learning via Dirichlet Updates with Decay

To learn the transition model, agents maintain a pseudo-count tensor $pB_(s', s, a)$ and update it after each observed transition under action according to:

$pB^{\text{new}}_{s',s,a} = (1 - \eta)\, pB^{\text{old}}_{s',s,a} + \eta\, \mathbf{1}\{s' = s'_{\text{obs}}\}$

Where:

- $\eta \in [0,1]$ is the learning rate (controls decay of past counts and incorporation of new evidence)
- The factor $(1 - \eta)$ implements **decay**, allowing older counts to fade.

Afterwards, the transition probabilities are normalized:

$$B_{s',s,a} = \frac{pB_{s',s,a}}{\sum_u pB_{u,s,a}}$$

Conjugacy ensures the posterior remains Dirichlet after each update.

B.4 Expected Free Energy and Action Selection

Agents select actions by minimizing the Expected Free Energy (EFE), which trades off Risk (preference fulfilment) and Epistemic value (information gain):

1. Risk:

$$\mathrm{Risk}(a) = \mathrm{KL}\big[Q(s_{t+1} \mid a) \,\|\, P_{\mathrm{pref}}(s_{t+1})\big] = \sum_{s'} Q(s' \mid a) \log \frac{Q(s'|a)}{P_{\mathrm{pref}}(s')}$$

2. Epistemic value:

$$\mathrm{EV}(a) = H\big[Q(o' \mid a)\big] = -\sum_{o'} Q(o' \mid a) \log Q(o' \mid a)$$

Where the predicted next-state distribution is:

$$Q(s_{t+1} = s' \mid a) = \sum_s q(s)\, B_{s',s,a}$$

And $Q(o_{t+1} = o' \mid a) = \sum_{s'} A_{o',s'}\, Q(s' \mid a)$

The full EFE is then:

$$\mathrm{EFE}(a) = \mathrm{Risk}(a) - \gamma\, \mathrm{Epi}(a)$$

with epistemic weighting with $\gamma \in [0, 1]$

Finally, a softmax over the negative EFE yields the policy:

$$\pi(a) = \frac{\exp\big(-\alpha\,\mathrm{EFE}(a)\big)}{\sum_{a'} \exp\big(-\alpha\,\mathrm{EFE}(a')\big)}$$

Where policy precision α>0 tunes the determinism of action sampling.

- Higher α: more deterministic (greedy) choice of minimal EFE actions.
- Higher γ: greater drive for reducing uncertainty (exploration).

Appendix C: Use of Artificial Intelligence Tools

In preparing this thesis, AI-based writing assistants were used in accordance with the Rector's Directive No. 2/2024 on the Responsible Use of AI Tools at Comenius University. Below is a summary of the relevant provisions and how they were applied:

1. **Permitted Uses**
    - AI tools were employed exclusively to support drafting and editing text (e.g., rephrasing, grammar checks), in line with Article II.1 of the Directive.
    - No AI tool was used to compose entire substantive passages, formulate original arguments, or replace the author's independent scholarly work (Article II.2).
2. **Verification and Transparency**
    - The author carefully reviewed and, where necessary, corrected all AI-generated suggestions to ensure accuracy and academic integrity, as required by Article II.3.

3. **Ethical and Legal Compliance**

    - Usage complied with all applicable laws, university regulations, and respect for intellectual property (Articles I.4–I.5).

    - The author maintained a critical stance toward AI outputs, recognizing their limitations and potential for error (Directive Preamble).

4. **Educational and Research Context**

    - This engagement with AI tools reflects our commitment to innovation in research methods (Article I.1) while preserving the author's sole responsibility for the thesis's scientific content and conclusions (Article II.2).

    - AI-generated ideas did not substitute for original data analysis, model development, or theoretical derivations, which were carried out entirely by the author.

.