Opponent Modelling in the Iterated Prisoner's Dilemma: A Literature Review

Faqih Mahardika
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia
fm.faqihmahardika@gmail.com

***Abstract*—Opponent modelling plays a central role in enabling adaptive behavior in the Iterated Prisoner's Dilemma (IPD), particularly under non-stationary and learning opponents. This systematic literature review examines prior work on opponent modelling in IPD along three dimensions: opponent behavior assumptions, modelling methodologies, and evaluation practices. The reviewed studies reveal a wide range of assumptions regarding opponent adaptivity, including stationary, reactive, learning-based, and population-mediated behaviors. Methodologically, opponent modelling approaches span gradient-based learning, deep reinforcement learning, recursive belief reasoning, evolutionary dynamics, and system identification. Evaluation practices predominantly rely on cooperation rate, payoff-based metrics, and equilibrium-related measures. While these metrics are informative for long-run performance, they are typically applied in settings with unconstrained or extended interaction horizons. As a result, the literature provides limited insight into how effectively opponent modelling strategies identify and adapt to opponents under short or resource-constrained interactions. The findings highlight the need for evaluation frameworks that explicitly consider interaction constraints when assessing opponent modelling effectiveness.**

***Index Terms*—keyword1, keyword2, keyword3**

I. Introduction

Cooperation and conflict are fundamental features of interaction in human societies, biological systems, and artificial-agent environments. Game theory provides a rigorous analytical framework for studying such strategic interdependence, and foundational work on the evolution of cooperation demonstrates how behavioural patterns emerge when agents repeatedly face dilemmas that require balancing self-interest with mutual benefit [2], [3]. Among these models, the Prisoner's Dilemma (PD) has become the canonical representation of situations in which individual rationality conflicts with collective welfare. In the one-shot PD, mutual cooperation yields the highest joint payoff, yet the incentive structure leads rational players to defect, exposing a core tension that recurs in many social, economic, and political contexts.

When the game is repeated as the Iterated Prisoner's Dilemma (IPD), the strategic landscape changes significantly. Repetition enables agents to condition their actions on previous interactions, giving rise to behaviours such as reciprocity, retaliation, forgiveness, and trust-building. Axelrod's influential IPD tournaments demonstrated that simple contingent strategies most notably Tit-for-Tat can foster stable cooperation even in competitive environments, illustrating how long-term interaction and memory contribute to the emergence of cooperative norms without centralised enforcement [2], [3]. These results established IPD as a fundamental tool for analysing adaptive, history-dependent decision-making.

The relevance of IPD dynamics extends across multiple domains. In biology, theories of reciprocal altruism describe how repeated encounters favour contingent cooperation, outlining conditions under which helping behaviour can evolve among self-interested organisms [4]. Evolutionary game theory formalises these mechanisms through the concept of evolutionarily stable strategies, showing how certain behavioural rules can persist under selective pressures [7]. In the social sciences, repeated-game frameworks explain human cooperation, norm enforcement, and punishment systems in public-goods settings [5]. In economics and resource management, problems such as commercial fishing demonstrate how strategic interdependence in shared environments mirrors the structure of repeated dilemmas [6]. Across these fields, the IPD serves as a versatile lens for understanding how cooperation emerges, stabilises, or collapses depending on incentives, information structures, and institutional conditions.

Although classical analyses of the IPD provide essential insights, their simplifying assumptions often diverge from the complexities of real strategic interaction. Real agents either human, biological, or artificial frequently operate under uncertainty, possess limited observability, and adapt their behaviour dynamically. These realities have motivated growing interest in opponent modelling, which equips an agent with the ability to infer an opponent's strategy, behavioural tendencies, or intentions from observed actions. Contemporary surveys highlight a broad spectrum of opponent-modelling techniques across adversarial and multi-agent domains, underscoring their importance for enabling adaptive decision-making in repeated games, negotiation, autonomous systems, and competitive learning environments [1]. Within the IPD, such capabilities are critical: sustaining cooperation depends on recognising cooperative partners, detecting exploiters, and adapting effectively to non-stationary or deceptive strategies.

Therefore, this study aims to systematically review the existing

literature on opponent modelling in the Iterated Prisoner's Dilemma. The review is guided by the following research questions:

- **RQ1:** What opponent behavior assumptions are used in opponent modelling for the Iterated Prisoner's Dilemma?
- **RQ2:** What methodological approaches to opponent modelling have been applied in the Iterated Prisoner's Dilemma?
- **RQ3:** How is the effectiveness of opponent-modelling strategies evaluated in the Iterated Prisoner's Dilemma?

To address these questions, Section 2 outlines the Systematic Literature Review (SLR) methodology used to identify, screen, and synthesise relevant studies. Section 3 presents the results and discussion of the review from three analytical perspectives: (i) Opponent behavior assumption comparison, (ii) methodological approach comparison, and (iii) evaluation-based comparison. Section 4 concludes the literature review.

II. Methodology

A systematic review will be employed to address the research questions. This method applies a transparent and structured procedure to collect relevant studies, appraise their quality, and integrate their findings in order to answer a well-defined research question [8].

A. Inclusion and Exclusion Criteria

Studies were included in the review if they satisfied all of the following conditions:

- **Topic relevance:** The study presents a substantive discussion, method, or empirical analysis of opponent modelling, opponent-aware decision making, or explicit reasoning about other agents' strategies in *repeated strategic interaction settings*. This includes, but is not limited to, the Iterated Prisoner's Dilemma (IPD), repeated general-sum games, repeated zero-sum games, and structurally comparable repeated social dilemmas where agents condition behavior on observed opponent actions.
- **Structural comparability:** The game or interaction setting involves repeated encounters, strategic interdependence, and observable opponent actions, enabling adaptation or reasoning over opponent behavior across time. The IPD is treated as a canonical instance within this broader class of repeated strategic games.
- **Scholarly quality:** The work is published in a peer-reviewed journal or international conference proceedings.
- **Publication window:** The study was published between 2019 and 2025. This window was selected to capture contemporary opponent-modelling approaches developed in the modern learning-based multi-agent systems era, while maintaining sufficient coverage given the limited volume of work explicitly addressing opponent modelling in repeated dilemma settings.
- **Accessibility:** The full text is available in English and accessible for analysis.

Studies were excluded if they met any of the following conditions:

- **Irrelevant focus:** The study does not address opponent modelling, opponent-aware adaptation, or reasoning about other agents, or focuses exclusively on one-shot games without repeated interaction.
- **Generic multi-agent learning:** The study applies multi-agent reinforcement learning or learning-in-games techniques without incorporating opponent-aware adaptation, either explicitly or implicitly, through action conditioning, belief updates, or history-dependent response mechanisms.
- **Purely theoretical without opponent reasoning:** The paper provides only abstract or normative theoretical analysis and does not propose, analyze, or operationalize any mechanism for opponent modelling, opponent inference, or opponent-aware decision making in repeated interaction settings.
- **Outside the publication window:** The study was published outside the defined six-year publication window.
- **Non-English or inaccessible:** The full text is not available in English or cannot be accessed.

B. Information Sources and Search Strategy

The literature search was conducted using four major academic databases selected for their comprehensive coverage of computer science, artificial intelligence, and multi-agent systems research: IEEE Xplore, ScienceDirect, Scopus, and SpringerLink. The search strategy employed a structured set of keywords designed to capture studies related to opponent modelling and the Iterated Prisoner's Dilemma (IPD). The following query was applied across all databases:

```
("opponent modelling" OR "opponent
 modeling" OR "opponent model" OR
        "agent modeling")
                AND
("prisoner's dilemma" OR "iterated
 prisoner's dilemma" OR "IPD" OR
"repeated game" OR "social dilemma")
```

Database-specific filters were applied to ensure the relevance and accessibility of retrieved publications.

1) **IEEE Xplore** (04–11–2025): Search applied to all fields.
2) **ScienceDirect** (04–11–2025): Search applied to all fields; restricted to Open Access publications.
3) **Scopus** (04–11–2025): Search applied to all fields.
4) **SpringerLink** (04–11–2025): Search applied to all fields; restricted to Open Access publications.

C. Research Procedure

The literature search was conducted following the defined search strategy, with keywords adapted to the specific query capabilities of each database. All identified records were collected and managed using a reference manager to remove duplicates and ineligible entries during the initial screening.

The remaining studies were then screened based on their titles and abstracts to determine relevance. Full-text versions of studies that passed this stage were retrieved and assessed for eligibility as part of the systematic review.

Data extraction focused on gathering information relevant to Opponent Modelling in the Iterated Prisoner's Dilemma, including environmental settings, monitoring methods, modelling approaches, and evaluation metrics. The extracted data were subsequently analyzed and compared to identify patterns, advantages, and limitations of the different approaches. These findings are then used to address the research questions outlined for this review.

III. Results

At each stage of the selection process illustrated in Figure 1, n denotes the number of studies considered. During the identification phase, 10 records were removed as duplicates based on their titles and abstracts. An additional 48 records were excluded for failing to meet the inclusion criteria. At the screening stage, 26 papers were removed because they were not relevant to this study (for example, they did not involve opponent modelling or did not use the IPD as the research context), and 6 papers were excluded because they were non-primary studies, resulting in a total of 32 removals. One record could not be retrieved due to access limitations. Consequently, 20 records were selected for full-text review.

Among the 20 full-text articles obtained, 2 were excluded because they focused on IPD population simulations or did not present any opponent modelling framework. Another 2 were excluded because they were non-primary studies and therefore did not propose their own opponent modelling approach. One additional paper was excluded for focusing solely on theoretical aspects of the IPD without providing an opponent modelling framework.

After applying all exclusion criteria, 15 studies were included in the final systematic review. The titles and characteristics of these studies are presented in Table I.

IV. Synthesis / Discussion

integrates findings across these perspectives to identify persistent limitations and articulate the research gap. While a broad set of attributes was extracted, only methodologically discriminative dimensions are used for direct comparison; environment and evaluation attributes are summarized separately.

A. What opponent behavior assumptions are used in opponent modelling for the Iterated Prisoner's Dilemma?

To operationalize *opponent behavior assumptions* across heterogeneous formulations, we summarize observable behavioral properties rather than internal learning mechanisms. The table below provides a descriptive categorization based on whether an opponent's actions (i) depend on environmental conditions, (ii) are mediated by population-level interactions, (iii) respond directly to the focal agent's actions, and (iv) exhibit action divergence under equivalent recent interaction histories. These

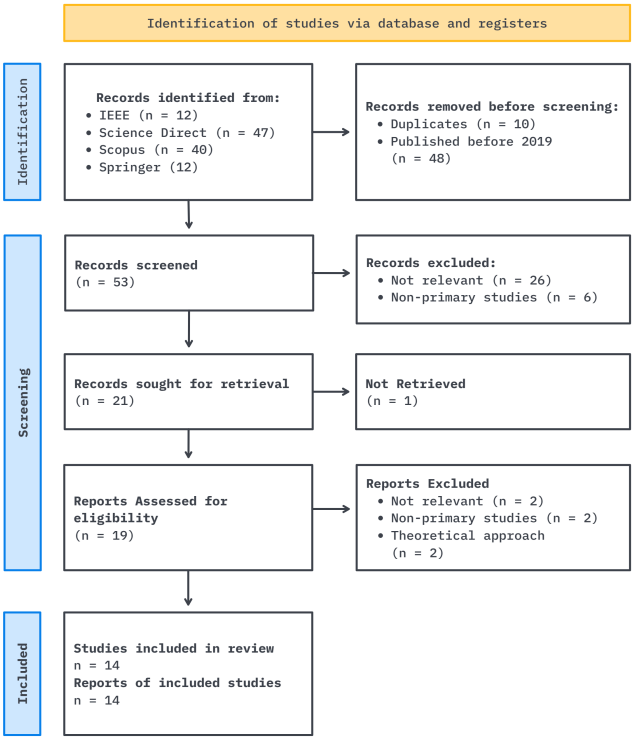

Fig. 1. Flow diagram for study selection.

criteria are intentionally behavior-centric and evaluated at the level of interaction traces, without assuming access to the opponent's internal model, update rules, or optimization objectives.

This categorization is not intended as a formal or exhaustive taxonomy of opponent modeling methods. Instead, it serves as an analytical aid to support comparison across studies that adopt different game settings, learning paradigms, and modeling abstractions. In cases where a paper does not explicitly define opponent assumptions, classifications are derived conservatively from the described experimental setup and observed interaction dynamics.

Several consistent patterns emerge from the behavioral categorization in Table II. Substantial portion of recent studies model opponents whose actions are explicitly agent-responsive and exhibit action divergence [9]–[12], [14]–[18], [20], [22], indicating an assumption of learning or strategic adaptation. Reflects a growing research emphasis on robustness against non-stationary and adaptive opponents rather than optimization against fixed strategies.

Moreover, population-mediated behavior appears primarily in studies of evolutionary or networked environments, where opponent actions are shaped indirectly through aggregate dynamics rather than direct bilateral adaptation [10], [13]–[15], [19]. These settings often decouple individual agent responsiveness from population-level change, leading to behavioral

TABLE I
CHARACTERISTICS OF INCLUDED STUDIES

Ref.	Title	Characteristics
[9]	O2M: Online Opponent Modeling in Online General-Sum Matrix Games	online opponent-modeling for dynamic general-sum games.
[10]	Evolutionary Dynamics of Cooperation and Extortion on Networks With Fitness-Dependent Rules	Analyzes zero-determinant strategies.
[11]	Inducing Coordination in Multi-Agent Repeated Game through Hierarchical Gifting Policies	Proposes a hierarchical gifting mechanism.
[12]	Recursively modeling other agents for decision making: A research perspective	Reviews recursive modeling frameworks.
[13]	Grounded predictions of teamwork as a one-shot game: A multiagent multi-armed bandits approach	Models voluntary teamwork.
[14]	The coupling effect between the environment and strategies drives the emergence of group cooperation	Analyze memory environment coupling.
[15]	Learning to cooperate against ensembles of diverse opponents	Scalable RL approach that fosters cooperation.
[16]	Exploiting a No-Regret Opponent in Repeated Zero-Sum Games	Framework for exploiting no-regret opponents.
[17]	Modeling Theory of Mind in Dyadic Games Using Adaptive Feedback Control	Proposes control-theoretic ToM agents.
[18]	Modeling opponent learning in multi-agent repeated games	Introduces a Stackelberg-based method.
[19]	The effects of population size and information update rates on the emergent patterns of cooperative clusters in a large-scale social particle swarm model	Faster information updates in SNS networks.
[20]	Achieving cooperation through deep multiagent reinforcement learning in sequential prisoner's dilemmas	Introduces a sequential PD and a deep RL method.
[21]	Achieving collective welfare in multi-agent reinforcement learning via suggestion sharing	Introduces a MARL method with sharing action suggestions.
[22]	Higher-order theory of mind is especially useful in unpredictable negotiations	Analyze higher-order ToM in unpredictable environments.

patterns that differ qualitatively from pairwise learning scenarios. Notably, only a subset of works combine environment-dependent, population-mediated, and agent-responsive behaviors [14], suggesting that fully context-conditioned strategic opponents remain comparatively underexplored.

B. What methodological approaches to opponent modelling have been applied in the Iterated Prisoner's Dilemma?

The distribution of opponent behavior assumptions in Table II reflects a broader methodological shift in opponent modeling research, from controlled and stationary assumptions toward behaviorally rich, adaptive, and heterogeneous opponents. While this characterization highlights the nature of the opponents considered, it does not capture how agents are designed

TABLE II
OPPONENT BEHAVIOR ASSUMPTION BASED ON OBSERVABLE BEHAVIORAL DEPENDENCIES.

Env.	Pop.	Agent	Div.	Opponent Behaviour Category	Ref.
–	–	✓	–	Reactive	[21]
–	✓	–	–	Population-Conformist	[13]
–	✓	✓	–	Contextual Reactive	[19]
–	–	✓	✓	Learning Opponent	[9], [11], [12], [16]–[18], [20], [22]
–	✓	✓	✓	Population-Contextual Strategic	[15]
✓	✓	–	✓	Heterogeneous Collective Behavior	[10]
✓	✓	✓	✓	Environment-Conditioned Strategic	[14]

Note:
Env. — behavior varies across environment within the same game;
Pop. — behavior is mediated by population-level interactions;
Agent — behavior responds directly to the focal agent's actions;
Div. — action divergence occurs on equivalent recent interaction histories.
Checkmarks indicate the presence of a dependency.

to cope with such complexity. Therefore, in Table III reorganize prior work from the perspective of agent-side modelling, categorizing approaches by their dominant opponent modelling paradigm and corresponding learning mechanism.

TABLE III
COMPARISON OF OPPONENT MODELLING APPROACHES BY DOMINANT MODELLING PARADIGM AND LEARNING MECHANISM.

Modelling Paradigm	Learning / Update Mechanism	Papers
Reactive reinforcement learning	Value-based RL (DQN)	[11]
Gradient-based opponent shaping	Gradient descent / opponent-aware updates	[9], [18], [20]
Recursive belief reasoning	Bayesian belief update / cognitive hierarchy	[12], [17], [22]
Population-based training	Policy-gradient MARL (population sampling)	[15]
Evolutionary population dynamics	Fitness-based strategy imitation	[10], [14], [19]
Bandit-based learning-in-games	Multi-armed bandit / smooth best response	[13]
System identification	Autoregressive model with exogenous inputs (NARX)	[16]
Communication-driven coordination	Policy-gradient with communication objective	[21]

Table III categorizes prior work by modelling paradigm and learning mechanism, the listed approaches also differ in whether opponent adaptation is handled explicitly or implicitly. Explicit opponent modelling approaches, such as gradient-based opponent shaping [9], [18], [20], recursive belief reasoning [12], [17], [22], and system identification [16], construct internal representations of opponent behavior. In

contrast, implicit approaches including reactive reinforcement learning [11], population-based training [15], evolutionary dynamics [10], [14], [19] and Bandit-based learning-in-games [13] adapt without maintaining explicit opponent model. This highlights alternative design philosophies for addressing non-stationarity opponent.

C. How is the effectiveness of opponent-modelling strategies evaluated in the Iterated Prisoner's Dilemma?

Evaluation choices implicitly define what constitutes success in adaptive multi-agent interaction, such as cooperation outcomes, robustness to exploitation, stability, or predictive accuracy. Table IV provides a descriptive overview of the evaluation environments and metrics used in prior work, with the aim of enabling cross-paper comparison and highlighting recurring evaluation patterns and omissions rather than standardizing or ranking criteria.

TABLE IV
EVALUATION ENVIRONMENTS AND METRICS IN STUDIES

Ref.	Evaluation Environment	Metrics
[9]	Symmetric self-play	MSE during offline training; latent memory accuracy
[10]	Scale-free network with zero-determinant strategies	Frequency of cooperation (C) and exploitation (E)
[11]	Adaptive opponents (averaged over multiple opponents)	Reward value
[12]	No training performed	Not applicable
[13]	Custom teamwork-game simulator (aggregative public good games); synthetic experiments	Aggregate team productivity; χ^2 goodness-of-fit to equilibrium; convergence to Nash equilibrium; individual contribution
[14]	Evolutionary game simulator on structured networks	Cooperation level; fraction of cooperators; phase-transition thresholds
[15]	Repeated matrix games with synthetic opponent populations	Average payoff; cooperation rate; robustness across opponent sets; generalization
[16]	Repeated zero-sum games against Hedge, OMD, and Regret Matching	Prediction error; cumulative payoff; robustness to non-stationarity
[17]	Repeated matrix games; embodied continuous-time robotic simulations	Efficacy; stability; prediction accuracy
[18]	Repeated matrix-game simulations	Average payoff; convergence speed; equilibrium selection
[19]	Large-scale continuous-space simulations (FLAME GPU)	Cooperation rate; cooperative cluster size and count; agent speed; cluster stability
[20]	Custom 2D SPD environments (Fruit Gathering; Apple–Pear games)	Average individual reward; total social welfare; cooperation degree detection accuracy
[21]	MARL benchmarks (Cleanup, Harvest, Sequential PD, Tragedy of the Commons)	Normalized return; cooperation rate; convergence speed; policy discrepancy (MSE)
[22]	Simulated Colored Trails with increasing environmental unpredictability	Allocator score; responder score; total social welfare

Table IV reveals a strong concentration of evaluation practices around aggregate performance outcomes, most notably cooperation rate [14], [19]–[21], average payoff [11], [15], [16], [20], [22], and convergence to equilibrium [13], [18], [21]. These metrics are well aligned with long-horizon repeated interaction settings, where early exploratory losses can be amortized over time and stable behavioral patterns eventually emerge.

However, comparatively little attention is given to interaction constraints, such as limited game horizon, finite learning budget, or the opportunity cost incurred during opponent probing. Even when non-stationary or adaptive opponents are considered, evaluation is typically conducted under conditions where prolonged interaction is assumed [9], and performance is assessed after convergence rather than during the learning phase itself.

As a result, existing evaluation protocols tend to underrepresent scenarios in which exploration is costly, miscoordination is irreversible, or early decisions dominate total return. This gap is particularly salient for online opponent modelling, where identifying opponent behavior necessarily requires intervention through action, yet the cost of such intervention is rarely isolated as an evaluation dimension. Consequently, current metrics provide limited insight into how efficiently an agent balances opponent identification against immediate performance under constrained interaction settings.

V. CONCLUSION

This review synthesized the opponent-modelling literature in repeated strategic interaction by examining how opponent behavior is assumed, how it is modelled, and how modelling effectiveness is evaluated, with a primary focus on canonical social dilemma settings commonly instantiated through the Iterated Prisoner's Dilemma. The analysis reveals substantial diversity in opponent behavior assumptions, ranging from stationary and reactive opponents to adaptive and population-mediated dynamics. While many studies implicitly account for non-stationarity arising from learning dynamics or environmental coupling, these assumptions are frequently embedded within experimental designs rather than stated explicitly. As a result, modelling approaches that appear methodologically similar may operate under fundamentally different opponent behavior assumptions, complicating direct comparison across studies.

Across the reviewed literature, a broad range of modelling methodologies has been employed, reflecting differing views on the appropriate abstraction for representing adaptive opponents. Evaluation practices predominantly emphasize outcome-based metrics such as cooperation rate, average payoff, and equilibrium convergence. These metrics are effective for assessing long-horizon performance and stability, but they are typically applied in settings where interaction horizons are unconstrained or sufficiently long to amortize exploration costs. Consequently, although opponent identification and adaptation are central motivations of opponent modelling, comparatively limited attention is paid to the efficiency and timeliness of such identification under constrained interaction horizons. This observation highlights an opportunity for

future research to examine opponent-modelling effectiveness in online settings where interaction opportunities are limited, exploration is costly, and early-stage adaptation is critical.

REFERENCES

[1] S. B. Nashed and S. Zilberstein, "A Survey on Opponent Modeling in Adversarial Domains" unpublished.

[2] R. Axelrod, "The Evolution of Cooperation" unpublished.

[3] R. Axelrod and W. D. Hamilton, "The Evolution of Cooperation" *Science*, 1981.

[4] R. L. Trivers, "The Evolution of Reciprocal Altruism" *The Quarterly Review of Biology*, vol. 46, no. 1, pp. 35–57, 1971. Available: https://www.journals.uchicago.edu/doi/10.1086/406755

[5] Y. Chen, "Cooperation and punishment in public goods experiments (by Ernst Fehr and Simon Gächter)," in *The Art of Experimental Economics*, G. Charness and M. Pingle, Eds., London: Routledge, 2021, pp. 134–143. Available: taylorfrancis.com/books/9781003019121/chapters/10.4324/9781003019121-12

[6] A. Leung and A.-Y. Wang, "Analysis of Models for Commercial Fishing: Mathematical and Economical Aspects," *Econometrica*, vol. 44, no. 2, p. 295, 1976. Available: https://www.jstor.org/stable/1912725

[7] J. M. Smith, *Evolution and the Theory of Games*. (Publication year not provided.)

[8] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, Mar. 2021.

[9] X. Qiao, C. Han, and T. Guo, "O2M: Online Opponent Modeling in Online General-Sum Matrix Games," in *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, Dec. 2024, pp. 358–361. [Online]. Available: https://ieeexplore.ieee.org/document/10899996/. doi: 10.1109/ICAIRC64177.2024.10899996.

[10] L. Zhu, Y. Zhu, and C. Xia, "Evolutionary Dynamics of Cooperation and Extortion on Networks With Fitness-Dependent Rules," in *2025 Joint International Conference on Automation-Intelligence-Safety (ICAIS) & International Symposium on Autonomous Systems (ISAS)*, May 2025, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/11051564/. doi: 10.1109/ICAISISAS64483.2025.11051564.

[11] M. Lv, J. Liu, B. Guo, Y. Ding, Y. Zhang, and Z. Yu, "Inducing Coordination in Multi-Agent Repeated Game through Hierarchical Gifting Policies," in *2023 IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, Sep. 2023, pp. 279–287. [Online]. Available: https://ieeexplore.ieee.org/document/10298392/. doi: 10.1109/MASS58611.2023.00041.

[12] P. Doshi, P. Gmytrasiewicz, and E. Durfee, "Recursively modeling other agents for decision making: A research perspective," *Artificial Intelligence*, vol. 279, p. 103202, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000437021930027X. doi: 10.1016/j.artint.2019.103202.

[13] A. L. de A. Gómez, C. Sierra, and J. Sabater-Mir, "Grounded predictions of teamwork as a one-shot game: A multiagent multi-armed bandits approach," *Artificial Intelligence*, vol. 341, p. 104307, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370225000268. doi: 10.1016/j.artint.2025.104307.

[14] C. Di, Q. Zhou, J. Shen, J. Wang, R. Zhou, and T. Wang, "The coupling effect between the environment and strategies drives the emergence of group cooperation," *Chaos, Solitons & Fractals*, vol. 176, 2023, pp. 114138. doi: 10.1016/j.chaos.2023.114138.

[15] I. Perera, F. de Nijs, and J. Garcia, "Learning to cooperate against ensembles of diverse opponents," *Neural Computing and Applications*, vol. 37, no. 23, 2025, pp. 18835–18849. doi: 10.1007/s00521-024-10511-9.

[16] K. Li, W. Huang, C. Li, and X. Deng, "Exploiting a No-Regret Opponent in Repeated Zero-Sum Games" *Journal of Shanghai Jiaotong University (Science)*, vol. 30, no. 2, 2025, pp. 385–398. doi: 10.1007/s12204-023-2610-2.

[17] I. T. Freire, X. D. Arsiwalla, J. Y. Puigbò, and P. Verschure, "Modeling Theory of Mind in Dyadic Games Using Adaptive Feedback Control," *Information*, vol. 14, no. 8, 2023. doi: 10.3390/info14080441.

[18] Y. Hu, C. Han, H. Li, and T. Guo, "Modeling opponent learning in multiagent repeated games," *Applied Intelligence*, vol. 53, no. 13, 2023, pp. 17194–17210. doi: 10.1007/s10489-022-04249-x.

[19] Z. Elhamer, R. Suzuki, and T. Arita, "The effects of population size and information update rates on the emergent patterns of cooperative clusters in a large-scale social particle swarm model," *Artificial Life and Robotics*, vol. 25, no. 1, 2020, pp. 149–158. doi: 10.1007/s10015-019-00558-6.

[20] W. Wang, Y. Wang, J. Hao, and M. E. Taylor, "Achieving cooperation through deep multiagent reinforcement learning in sequential prisoner's dilemmas," in *ACM International Conference Proceeding Series*, 2019. doi: 10.1145/3356464.3357712.

[21] Y. Jin, S. Wei, and G. Montana, "Achieving collective welfare in multi-agent reinforcement learning via suggestion sharing," *Machine Learning*, vol. 114, no. 8, 2025, p. 190. doi: 10.1007/s10994-025-06823-z.

[22] H. De Weerd, R. Verbrugge, and B. Verheij, "Higher-order theory of mind is especially useful in unpredictable negotiations," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 2, 2022, p. 30. doi: 10.1007/s10458-022-09558-6.