LLM Agents Learn Effective Partner Selection for Mutual Cooperation in Social Dilemmas

Xiaoqing Fan
Mathematics of Real World System CDT, University of Warwick
Coventry, United Kingdom
xiaoqing.fan@warwick.ac.uk

Chin-wing Leung
Department of Computer Science, University of Warwick
Coventry, United Kingdom
chin-wing.leung@warwick.ac.uk

Paolo Turrini
Department of Computer Science, University of Warwick
Coventry, United Kingdom
p.turrini@warwick.ac.uk

ABSTRACT

Social dilemmas, where individual incentives conflict with collective welfare, present a fundamental challenge to the emergence of cooperation in multi-agent systems. Partner selection—the capacity for agents to autonomously update their social ties—is a powerful mechanism for promoting cooperation by enabling individuals to ostracise defectors and cluster with cooperators. While prior research has demonstrated that standard reinforcement learning agents can learn cooperation-sustaining rules, it remains unclear how sophisticated, reasoning-capable agents based on Large Language Models (LLMs) navigate these strategic interactions. In this paper, we investigate the co-evolution of strategy and structure among self-interested LLM-based agents. Through experiments using the Prisoner's Dilemma on a network, we find that LLM reasoning alone is insufficient to sustain cooperation in these environments. However, when these agents are able to learn through fine-tuning, they autonomously develop effective partner selection behaviours, such as the "Out-for-Tat" rule. Our results demonstrate that LLMs can dynamically adapt their interaction rules in alignment with their evolving game strategies. Furthermore, we confirm that the relative timescales of link and strategy updates are critical factors in the promotion of cooperation, providing new insights into the stability and self-organisation of LLM-driven societies.

KEYWORDS

Multi-agent Systems, Social Dilemmas, Large Language Model

ACM Reference Format:
Xiaoqing Fan, Chin-wing Leung, and Paolo Turrini. 2026. LLM Agents Learn Effective Partner Selection for Mutual Cooperation in Social Dilemmas. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 6 pages.

1 INTRODUCTION

The tension between self-interest and common good is a ubiquitous feature of strategic social interactions. In many scenarios, while mutual cooperation yields the highest collective benefit, the temptation to defect can lead to the notorious "tragedy of the commons", where everyone is worse off than they would have been if they acted as a collective [8]. To resolve this, researchers have identified various mechanisms to promote socially desirable outcomes. One of the most effective enablers of cooperation is partner selection [5, 17, 23], which introduces a structural dimension to the interaction: agents not only decide how to play, but also whom to play with. By allowing agents to sever ties with uncooperative partners and seek out new associations, partner selection creates "network reciprocity" [16] that can sustain cooperation even in the face of strong incentives to defect.

Previous studies have shown that partner selection rules can be hand-crafted to promote cooperation. More recently, it has been demonstrated that Reinforcement Learning (RL) agents can learn these rules from scratch [2, 7, 14]. In these models, agents often converge on an "Out-for-Tat" (OFT) strategy—maintaining links with cooperators while breaking them with defectors—thereby allowing cooperation to flourish.

The recent rise of Large Language Models (LLMs) as autonomous agents introduces a new paradigm for studying these dilemmas. Unlike traditional RL agents, LLM agents possess advanced reasoning and planning capabilities, enabling more sophisticated social interactions. As LLMs are increasingly deployed in multi-agent environments—from collaborative software development [9] to online market coordination [3]—it becomes critical to understand how they navigate conflicting objectives. While prior work has examined LLM behaviour in fixed interaction structures, it remains unclear whether these models can autonomously learn partner selection rules that reshape their own interaction networks.

Contribution. With this paper, we bring partner selection strategies to the realm of LLM-based agents. Specifically, we investigate a population of LLM agents playing the Prisoner's Dilemma (PD) on a dynamic network, where the agents must decide whether to maintain their existing neighbours or rewire to new ones. Our key contributions are as follows: 1) We show that the inherent reasoning capabilities of LLMs are insufficient to sustain cooperation on their own; without experience-based training, the models fail to coordinate on effective partner selection strategies. 2) We demonstrate that through fine-tuning via Proximal Policy Optimisation (PPO), LLM agents autonomously learn the "Out-for-Tat" rule, effectively using partner selection to purge defectors and promote cooperation. 3) We find that the learned partner selection policies co-evolve with game strategies, resulting in distinct behavioural profiles for cooperators and defectors. 4) We confirm that the relative timescale of updates is crucial: cooperation only flourishes when the network structure (rewiring) evolves at a sufficiently high rate relative to strategy updates.

This work provides new insights into how AI agents can leverage partner selection to promote and sustain cooperative societies.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus.

	C	D
C	R, R	S, T
D	T, S	P, P

	C	D
C	3, 3	−1, 4
D	4, −1	1, 1

Table 1: Payoff Matrix for the Prisoner's Dilemma, for which inequalities $T > R > P > S$ and $2R > T + S$ hold true. On the left, the general payoff matrix; on the right, our instantiation.

2 RELATED RESEARCH.

Understanding the emergence of cooperation among self-interested individuals remains a central theme in fields ranging from economics and biology to computational social science [4, 12, 16, 22]. In multi-agent systems, partner selection has emerged as a particularly potent mechanism for fostering cooperation by allowing agents to update their social ties autonomously [10, 18]. Recent work has explored how independent reinforcement learning (RL) agents can learn these rules from scratch. For instance, Anastassacos et al. [2] demonstrate that agents can learn to selectively choose partners based on the observed actions of others. Leung et al. [13] showed that cooperation can be sustained even under a minimal degree of freedom, where agents only decide whether to stay with a current partner or be randomly rewired. In social dilemmas under dynamic networks [6, 14], partner selection creates a co-evolutionary dynamic between game strategy and network structure, leading to the formation of cooperative clusters via network reciprocity. Our work extends this lineage to the domain of LLMs, investigating whether the cooperative behaviours observed in traditional RL agents can emerge from agents capable of complex reasoning.

The integration of LLMs as autonomous agents introduces a new dimension to strategic social dilemmas, yet recent findings regarding their inherent pro-sociality remain divided. Lorè et al. [15] found that LLM agents like GPT-4 and Llama-2 exhibit significantly different strategic behaviours depending on the contextual framing of the game, highlighting the sensitivity of their internal reasoning. Similarly, Akata et al. [1] observed that while LLMs perform well in self-interested tasks, they often resort to selfish behaviour in coordination games like the Battle of the Sexes. To address behavioural inconsistencies, Tennant et al. [21] utilise fine-tuning with intrinsic rewards to align LLM agents with human-centric moral values. However, most existing research has examined LLM behaviour within fixed interaction structures. Our study fills this gap by exploring how LLMs navigate social dilemmas when they have the agency to reshape their own interaction networks.

Paper Structure. Section 3 reviews the Prisoner's Dilemma and fine-tuning under Low-Rank Adaptation. Section 4 details the experimental setting. Section 5 presents our findings on cooperation behaviour and the learned partner selection rules. Finally, Section 6 concludes with a discussion of future research.

3 PRELIMINARY

3.1 Prisoner's Dilemma

Social dilemmas are characterised by a conflict between individual incentives and collective welfare, with the Prisoner's Dilemma (PD) being the archetypal model. In this symmetric two-player game, agents choose to either Cooperate (C) or Defect (D). The resulting outcomes yield four payoffs: R (mutual cooperation), P (mutual defection), S (cooperating against a defector), and T (defecting against a cooperator), as shown in Table 1. While mutual cooperation maximises the collective payoff, defection is the strictly dominant strategy, leading to a Nash equilibrium at mutual defection. In our framework, interactions occur on a dynamic network where neighbouring agents are randomly selected to perform link update and play PD, or to imitate others' strategy.

3.2 Fine-tuning with Reinforcement Learning

We adapt LLM agents to these strategic interactions using Proximal Policy Optimisation (PPO) [19], a policy gradient method that utilises a clipped surrogate objective function. This clipping prevents excessively large policy updates, thereby stabilising the training of large language models. To ensure computational efficiency, we employ Low-Rank Adaptation (LoRA) [11]. LoRA freezes the pretrained weights W_0 and optimises trainable low-rank decomposition matrices B and A such that the weight update is $\Delta W = BA$. The modified forward pass is expressed as:

$$h = W_0 x + \frac{\alpha}{r} BAx$$

where r is the rank and α is a scaling hyperparameter. This method significantly reduces the number of trainable parameters while maintaining the model's performance on complex tasks.

4 METHODS

4.1 Agents Interactions

We investigate the co-evolution of strategy and structure in a population of $N = 10$ LLM-based agents, initially organised in a regular graph where each agent maintains $k = 3$ neighbours. Agents are assigned an initial binary strategy, $S \in \{C, D\}$ (Cooperate or Defect), chosen uniformly at random. The system evolves through two concurrent processes: structural updates via partner selection and strategy updates via imitation.

Partner Selection: In each discrete time step, a link L_{ij} between agents i and j is randomly selected for potential revision. Both agents are provided with a prompt (see Appendix A) containing the game rules and the opponent's existing strategy s_i or s_j. Agents must then output a structural answer $A \in$ {"stay", "leave"}.

- If both agents choose "stay", the link persists.
- If either agent chooses "leave", the link is severed. One of the agents is then randomly rewired to a new partner in the population to maintain a constant average degree.

Following the structural update, agents play a round of the Prisoner's Dilemma (PD). If the link was severed, the rewired agent interacts with its new partner, while the abandoned agent receives a default payoff of 1 (the payoff under mutual defection), ensuring it is not unfairly penalised for a broken tie.

To ensure network connectivity, we enforce a constraint that each agent must maintain at least one active link. Therefore, if any agent in the pair has only 1 neighbour, the connection remains even though they choose to leave.

Strategy Update: To update strategies, a random agent i and a random neighbour j are selected. Agent i imitates neighbor j's

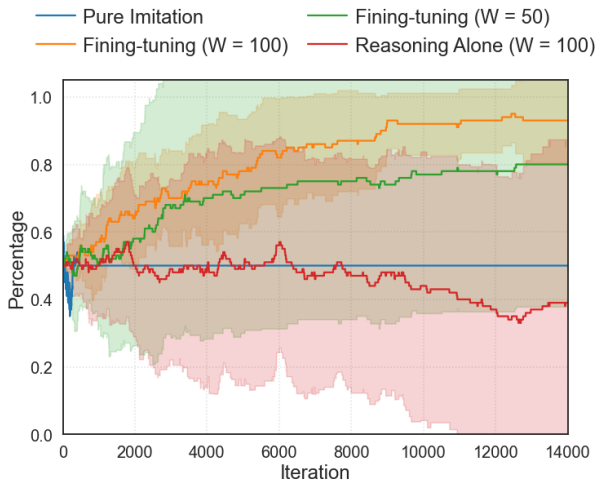


Figure 1: Evolution of cooperation rates under varying structural settings and timescales W. Each curve shows the mean and standard deviation of 10 independent simulations. While baseline imitation (blue) and LLM reasoning (red) fail to sustain cooperation, fine-tuning at a higher timescale ($W = 100$) enables the population to reach near-universal cooperation.

strategy with a probability p determined by the Fermi function:

$$p = \frac{1}{1 + e^{-\beta(f(j)-f(i))}}$$

where $f(i)$ and $f(j)$ represent their respective fitness (accumulated payoffs), and $\beta = 0.005$ is the inverse temperature controlling the propensity of imitation.

Timescale Ratio (W): To analyse the interplay between structural and strategy evolution, we define a timescale ratio, W, which governs the relative frequency of link updates versus strategy updates. In each iteration, a strategy update occurs with probability $p = 1/(1 + W)$, and a link update occurs with probability $1 - p$. As W approaches 0, the network remains static; as W increases, structural flexibility grows, allowing agents to react more promptly to their neighbours' behaviour.

4.2 Implementation Details

We utilise Gemma-2-2b-it [20] (a lightweight and open-source model) as the base model for all agents. Fine-tuning is performed using the PPO trainer from the TRL library for $T = 14,000$ iterations with a batch size of 20. For computational efficiency, we employ LoRA with 4-bit quantisation (rank $r = 64$), training approximately 5% of the total parameters. The decision-making process is mapped to specific tokens: action1 and action2 represent Cooperate and Defect, while the structural decisions "stay" and "leave" are restricted to a single-token generation. Reward scaling and normalisation are applied with gradient accumulation over 5 steps to stabilise the learning process. The learning rate is set to be 1.41×10^{-5} and to prevent the agents' policies from deviating excessively from the base Gemma model, we employed adaptive KL divergence control with an initial coefficient of 0.2.

Table 2: Evolution of partner selection strategies for cooperators and defectors ($W = 100$). Percentages indicate the prevalence of specific heuristics across iterations. While cooperators maintain a stable preference for the OFT strategy, defectors converge toward an Always-Stay heuristic as they become increasingly isolated in the network.

Iteration	Strategy	Stay	OFT	R_OFT	Leave	Percentage
4000	C	36.5%	47.3%	1.4%	14.9%	74%
	D	50%	15.4%	0%	34.6%	26%
8000	C	46.0%	42.5%	1.1%	10.3%	87%
	D	38.5%	38.5%	23.1%	0%	13%
12000	C	37.6%	45.2%	0%	17.2%	93%
	D	85.7%	0%	0%	14.3%	7%

5 EXPERIMENTAL RESULTS

In this section, we present our experimental results. By comparing with pure imitation and partner selection with reasoning alone, we show how allowing agents learn through fine-tuning can promote cooperation. We look at the learned partner selection policies of agents under different timescale settings, as well as the behavioural differences between Cooperators and Defectors.

5.1 Cooperation Emerges under Fine-tuning

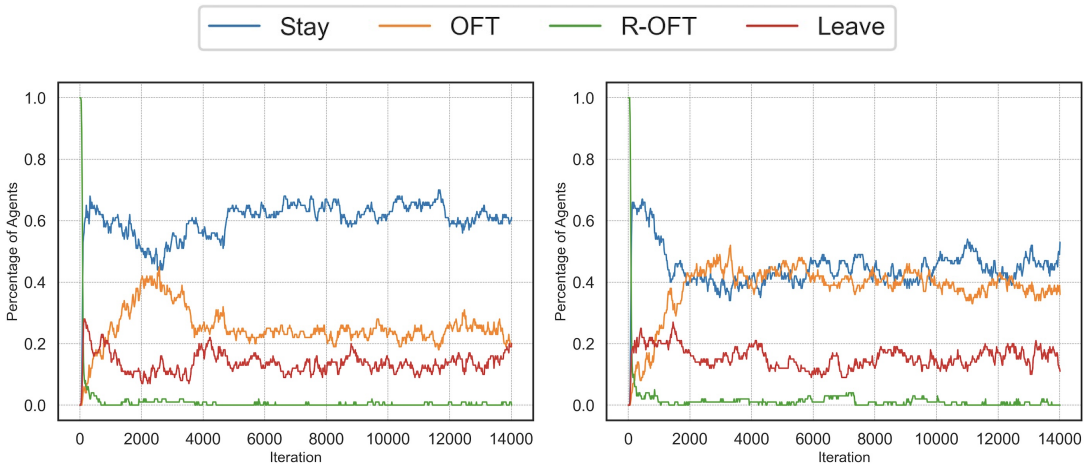
Figure 1 illustrates the evolution of cooperation rates across 14,000 iterations for various structural configurations and timescale ratios W. Our results, averaged over 10 independent simulations, reveal distinct regimes of social coordination:

Pure imitation (baseline): In environments without partner selection (blue line), where agents only adapt their strategies, the cooperation rate remains stagnant at approximately 50%. This confirms that imitation alone, in the absence of structural flexibility, is insufficient to catalyse a transition toward collective cooperation.

Partner selection with reasoning alone: Surprisingly, when LLM agents perform partner selection based on their initial reasoning without fine-tuning (red line), cooperation rates decline slightly over time. This suggests that the inherent reasoning capabilities of LLMs are insufficient for navigating the complexities of dynamic social dilemmas out-of-the-box.

Partner selection with fine-tuning: A critical shift occurs when agents are permitted to fine-tune their policies based on interaction history. For $W = 50$ (green line), the population reaches a stable cooperation rate of approximately 80%. When the timescale is extended to $W = 100$ (orange line), the society undergoes a rapid transformation, converging toward a near-universal cooperative state (97%). This underscores the necessity of high structural flexibility, which allows cooperative agents to promptly isolate defectors and leverage network reciprocity.

In summary, the results demonstrate that as the timescale is sufficiently high, fine-tuned LLM agents can successfully learn to optimise their local network environments. By autonomously purging uncooperative partners, it promotes and sustains a cooperative social environment in the long run.



(a) Learning Imitation with timescale W=50

(b) Learning Imitation with timescale W=100

Figure 2: The Learnt policy types over iteration for different timescales (W) for the PD game. (a) At $W = 50$, agents fail to stabilise the optimal OFT strategy. (b) High structural flexibility ($W = 100$) facilitates the adoption of the Out-for-Tat rule, enabling cooperators to successfully isolate defectors and reach high levels of social coordination.

5.2 Learnt Partner Selection Policies

To understand the behavioural foundations of these results, we classify the learned partner selection policies into four types based on the models' next-token preferences. Specifically, given the prompt for partner selection, we extract the model's output logits for predefined action tokens and compare their relative magnitudes to determine the agent's behaviour intent. Based on an agent's response to the opponent's strategy, the agent is categorised as adopting the Always-Stay(Stay) strategy if the probability of choosing "stay" is higher than "leave" regardless of the opponent's behavior, satisfying the conditions $P(\text{stay}|C) > P(\text{leave}|C)$ and $P(\text{stay}|D) > P(\text{leave}|D)$. On the other hand, if the agent maintains the tie given the opponent cooperates but severs it given the opponent defects ($P(\text{stay}|C) > P(\text{leave}|C)$ and $P(\text{leave}|D) > P(\text{stay}|D)$), the agent is categorised as adopting the Out-for-Tat (OFT) strategy. Thus, the four partner selection Policy types are: Always-Stay (Stay), Out-for-Tat (OFT), Reverse-OFT (R-OFT), and Always-Leave (Leave).

We present the emergence of these rules under different timescales (W) in Figure 2. We can observe that the learned partner selection rules are highly sensitive to the timescale. At lower timescales ($W = 50$), agents learn to adopt the OFT strategy to a certain extent in the earlier stage (2000^{th} iteration); however, they fail to commit to the OFT strategy. Instead, a larger proportion eventually settles on an "Always-Stay" approach. This "loyalty" inadvertently limits the network's ability to exclude defectors, resulting in the lower cooperation rates observed in Figure 1. With higher structural flexibility ($W = 100$), the OFT strategy becomes the dominant heuristic. This allows cooperators to effectively purge uncooperative ties and cluster together, stabilising a pro-social environment.

Furthermore, we observe a distinct co-evolution of strategy and interaction rules. As detailed in Table 2, cooperators consistently favour the OFT strategy (around 45% prevalence) as a defensive mechanism across the simulation, followed by Always-Stay and Always-Leave. In contrast, the defectors are switching their stance over time, starting from adopting Always-Stay and Always-Leave at the earlier stage, to Always-Stay and OFT at the 8000^{th} iteration and finally predominantly adopt an Always-Stay heuristic (85.7%) at the end of the simulation. This suggests that in an environment governed by reciprocal cooperators, defectors lose their agency and are effectively forced into static, isolated positions where they can no longer exploit the collective.

6 DISCUSSION

Our results demonstrate that the inherent reasoning of LLM agents is alone insufficient to solve social dilemmas out-of-the-box, and experience-based fine-tuning via PPO enables agents to autonomously learn effective partner selection rules, such as "Out-for-Tat" (OFT). We confirm that high structural flexibility—governed by the timescale ratio W—is critical for stability, as it allows cooperators to proactively purge uncooperative ties and foster collective welfare through network reciprocity. Notably, we observe a distinct co-evolution of strategies during the learning process: while cooperators consistently favour the reciprocal OFT strategy, defectors are forced to shift their stance over time as they become increasingly marginalised. Granting agents the capacity to curate their own social ties, forming and severing links with others, seems crucial to develope autonomous AI systems aligned with prosocial values.

Future research will explore models where agents autonomously decide when to update their strategies and social ties together, moving beyond fixed imitation protocols. Additionally, investigating how natural language communication influences rewiring decisions remains a promising direction for understanding how linguistic cues shape social structure. Also, applying this framework to larger-scale language models with enhanced reasoning capabilities could uncover whether higher level LLM reasoning can lead to more

resilient social norms. This can also be coupled with the analysis of emergence of more sophisticated strategies, where strategy and structure co-evolve. Finally, it is important to understand the repercussion of the design of LLM societies in terms of fine-tuning strategies.

ACKNOWLEDGMENTS

If you wish to include any acknowledgements in your paper (e.g., to people or funding agencies), please do so using the 'acks' environment. Note that the text of your acknowledgements will be omitted if you compile your document with the 'anonymous' option.

REFERENCES

[1] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. Playing repeated games with large language models. *Nature Human Behaviour* (2025), 1–11.

[2] Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. 2020. Partner selection for the emergence of cooperation in multi-agent systems using reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7047–7054.

[3] Devesh Batra, Conor Hamill, John Hartley, Ramin Okhrati, Dale Seddon, Harvey Miller, Raad Khraishi, and Greig Cowan. 2025. A review of llm agent applications in finance and banking. *Available at SSRN 5381584* (2025).

[4] Samuel Bowles and Herbert Gintis. 2013. *A Cooperative Species*. Princeton University Press.

[5] Ilan Eshel and Luigi Luca Cavalli-Sforza. 1982. Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences* 79, 4 (1982), 1331–1335.

[6] Xiaoqing Fan, Chin-wing Leung, and Paolo Turrini. 2025. Co-learning of strategy and structure achieves full cooperation in complex networks with dynamical linking. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*. 72–80.

[7] Zachary Fulker, Patrick Forber, Rory Smead, and Christoph Riedl. 2021. Spite is contagious in dynamic networks. *Nature communications* 12, 1 (2021), 260.

[8] Garrett Hardin. 1968. The Tragedy of the Commons. *Science* 162, 3859 (1968), 1243–1248. http://www.jstor.org/stable/1724745

[9] Junda He, Christoph Treude, and David Lo. 2025. LLM-Based Multi-Agent Systems for Software Engineering: Literature Review, Vision, and the Road Ahead. *ACM Transactions on Software Engineering and Methodology* 34, 5 (2025), 1–30.

[10] David Hirshleifer and Eric Rasmusen. 1989. Cooperation in a repeated prisoners' dilemma with ostracism. *Journal of Economic Behavior & Organization* 12, 1 (1989), 87–106. https://doi.org/10.1016/0167-2681(89)90078-4

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[12] Peter Kollock. 1998. Social dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology* 24, 1 (1998), 183–214.

[13] Chin-wing Leung and Paolo Turrini. 2024. Learning Partner Selection Rules that Sustain Cooperation in Social Dilemmas with the Option of Opting Out *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1110–1118.

[14] Chin-wing Leung, Tom Lenaerts, and Paolo Turrini. 2024. To promote full cooperation in social dilemmas, agents need to unlearn loyalty. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 111–119.

[15] Nunzio Lorè and Babak Heydari. 2024. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports* 14, 1 (2024), 18490.

[16] Martin A Nowak. 2006. Five rules for the evolution of cooperation. *science* 314, 5805 (2006), 1560–1563.

[17] David G Rand, Samuel Arbesman, and Nicholas A Christakis. 2011. Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences* 108, 48 (2011), 19193–19198.

[18] Francisco C Santos, Jorge M Pacheco, and Tom Lenaerts. 2006. Cooperation prevails when individuals adjust their social ties. *PLoS computational biology* 2, 10 (2006), e140.

[19] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[20] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).

[21] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2024. Moral alignment for llm agents. *arXiv preprint arXiv:2410.01639* (2024).

[22] Paul AM Van Lange, Jeff Joireman, Craig D Parks, and Eric Van Dijk. 2013. The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes* 120, 2 (2013), 125–141.

[23] Xiu-Deng Zheng, Cong Li, Jie-Ru Yu, Shi-Chang Wang, Song-Jia Fan, Bo-Yu Zhang, and Yi Tao. 2017. A simple rule of direct reciprocity leads to the stable coexistence of cooperation and defection in the Prisoner's Dilemma game. *Journal of Theoretical Biology* 420 (2017), 12–17.

A PROMPTS FOR DECISION-MAKING

The prompt utilised during the training process is detailed below, which includes the rules of the game, opponent strategy and specifies the available decision space. This is provided to LLM-agents to facilitate partner selection decisions. Action 1 denotes cooperation (C), while Action 2 denotes defection (D).

Training partner selection prompt (*action1 = Cooperate, action2 = Defect*)

<bos><start_of_turn>user
You are playing a game with another agent A. You must decide whether to continue playing the game with agent A by choosing either action stay or action leave. If you choose stay, you will keep playing with agent A. If you choose leave, you will stop playing with A and start playing with a random agent B. Depending on your strategy and A's strategy, you each get a certain number of points. The points are awarded as follows (you are the row player, A is the column player):

```
          |  action1  |  action2  |
|---------|-----------|-----------|
| action1 |    3, 3   |    0, 4   |
| action2 |    4, 0   |    1, 1   |
```

You have played with this opponent before. Last time, they played <opponent_action>. To achieve the highest possible score in points, decide whether to stay with agent A or leave and play with another agent. Your answer must follow this format exactly: either stay or leave. Do not explain your reasoning. Your answer:<end_of_turn>
<start_of_turn>model