論文 / 著書情報
Article / Book Information

Title	Cooperation: A Systematic Review of how to Enable Agent to Circumvent the Prisoner ' s Dilemma
Authors	Jiateng Pan, Atsushi Yoshikawa, Masayuki Yamamura
Citation	SHS Web of Conferences, Volume 178, ,
Pub. date	2023, 10
DOI	https://doi.org/10.1051/shsconf/202317803005
Creative Commons	Information is in the article.

Cooperation: a Systematic Review of how to Enable Agent to Circumvent the Prisoner's Dilemma

Jiateng Pan[1,a] [0000-0001-6187-9459], Atsushi Yoshikawa[1, 2], Masayuki Yamamura[1]

[1]School of Computing, Tokyo Institute of Technology, Japan
[2]College of Science and Engineering, Kanto Gakuin University, Japan

Abstract. It is widely accepted that rational individuals are unable to create cooperation in a prisoner's dilemma. However, in everyday life, cooperation, for example, during a fishing moratorium, can be observed frequently. Additionally, the appearance of cooperation in the prisoner's dilemma can be seen in numerous simulation studies.

This paper reviews 31 simulation studies published between January 2017 and January 2023 in which agents can be observed in the results to improve cooperation in a prisoner's dilemma.

The proposed methodologies were sorted into seven categories, including Bounded Rationality, Memory, Adaptive Strategy, Mood Model, Intrinsic Reward, Network Dynamics, and Altruistic Attribute. Based on their impacts, the effectiveness of these seven approaches was classified into three categories: generating cooperation, maintaining cooperation, and spreading cooperation.

This review is expected to be helpful for scholars conducting future research on multi-agent cooperation and irrational agent modeling.

1 Introduction

1.1 Background and Rationale

The Prisoner's Dilemma is a game that has been analyzed in game theory (Tucker & Straffin, 1983; Felkins, 2001; Chong, et al., 2007)[4, 9, 35]. It presents a dilemma for two entirely rational agents, where they must decide to cooperate with their partner for the common good or to betray their partner for a personal reward (defect).

Table 1. Payoff table for the prisoner's dilemma

	Cooperate	Defect
Cooperate	**R**, **R**	**S**, **T**
Defect	**T**, **S**	**P**, **P**

The classic prisoner's dilemma depicts a scenario in which two rational prisoners, A and B, are held in solitary confinement and are each faced with an either/or situation: either they choose to be silent (Cooperate), or they choose to betray (Defect). This leads to 4 different outcomes:

1. If A and B both remain silent, they will each serve the lesser charge of 2 years in prison.
2. If A betrays B but B remains silent, A will be set free while B serves 10 years in prison.
3. If A remains silent but B betrays A, A will serve 10 years in prison and B will be set free.
4. If A and B both betray the other, they share the sentence and serve 5 years.

Since both Prisoner A and Prisoner B would avoid the worst of 10 years of incarceration by choosing to "Defect", while having a chance of being released outright, they both end up selecting to "Defect" and receive 5 years of imprisonment. However, overall the outcome would have been more favorable if they had chosen to cooperate. A total of 4 years in prison would have been shorter than the 5 years for both.

The definition of the prisoner's dilemma (Hofstadter, 1983) [16] can be standardized, as shown in Table 1, by defining the gain of cooperating with each other as Reward (R), the gain of betraying each other as Punishment (P), the gain of betraying alone as Temptation (T), and the gain of being betrayed alone as Suckers (S). When the four gains satisfy the inequality:

$$T > R > P > S$$
$$2R > T + S$$

the game is a prisoner's dilemma game.

Many studies related to the prisoner's dilemma have utilized agent-based simulations (Gotts, et al., 2003) [13], which use the ABS model to examine possible social problems in the prisoner's dilemma and to suggest possible solutions.

In these studies, researchers have proposed several ways to prevent agents from falling into a prisoner's dilemma, the most famous of which is the tit-for-tat

[a] pan.j.aa@m.titech.ac.jp

strategy (Rapoport, 1989) [28], which allows agents to mimic the choices their opponents made in the previous round. It is widely believed that only repeated prisoner's dilemma games can cause participants to shift from focusing on $T > R > P > S$ to focusing on $2R > T + S$ (Hofstadter, 1983) [16], which leads to cooperation.

Repeated prisoner's dilemma games come in various forms, and the ontology mainly considers the "game → elimination → reproduction → game" model. In this model, all participants play the game, then the participants with the lowest scores are eliminated and the remaining participants are proportionally reproduced to the original number and the game continues. In this way, the strategies and overall gains of the last remaining participants can be observed to determine whether cooperation is generated.

Firstly, the two most basic policies, AllD and AllC, need to be defined.

Where AllD refers to a participant's unconditional choice of "defect" and AllC refers to a participant's unconditional choice of " cooperate".

It is easy to see that AllD can gain a lot by exploiting AllC. In the "game → elimination → reproduction → game" model, AllC will be eliminated quickly. Since the TFT strategy suppresses the return of AllD and protects the return of AllC, AllC will be retained and not eliminated in studies where the TFT strategy is used.

However TFT still has many drawbacks such as no fault tolerance mechanism (Kopelman, 2020) [19], so more methods or better strategies need to be added to achieve more cooperation.

This review will summarize and conclude these methods.

Also with the development of artificial intelligence, reinforcement learning has been introduced into the study of the prisoner's dilemma (Sandholm & Crites, 1996; Fujimoto & Kaneko, 2019; Gill & Rosokha, 2020) [11,12,29].

The agent that introduces learning can learn through its own choices and gains, and constantly change its strategy. However, it is impossible to escape the prisoner's dilemma by learning only the immediate gains. In recent years, a plethora of methods exist that enable agent to learn group gains.

This review also summarizes such studies.

1.2 Purpose of the Review

Research on cooperation between multiple agents has focused on how to encourage rational agents to cooperate (Jiang & Lu, 2018; Xuan, et al., 2001, May; Lazaridou, et al., 2016) [18, 20, 44]. Simulations of the prisoner's dilemma show that cooperation can be observed under certain approaches.

However, the absence of a unified framework to consolidate these methods has resulted in numerous studies repeating the same work.

The purpose of this paper is to summarize the methods used in recent years for how to get agents out of the prisoner's dilemma and to categorize the methods used so that subsequent researchers can more easily begin their work. Furthermore, it is anticipated that this review will contribute to further research regarding the modeling of irrational behavior.

The objective of this paper is to reach a general conclusion on how agents can evade the prisoner's dilemma by answering the following research questions:

1. What methods are used in these studies to circumvent the prisoner's dilemma?
2. Which environments are these methods applicable to?
3. How effective are these methods?

2 Methods

2.1 Search Strategy

The literature for this review was searched for the period January 2017 to January 2023,

where the subject terms need to include either variable-sum, non-constant-sum or Prisoner's Dilemma,

and in the abstract, it needs to mention either agents, Artificial Intelligence or AI,

also in the abstract needs to mention one of Cooperation, Cooperate or Win-Win.

The search statement is as follows:

("VARIABLE-SUM" OR "PRISONER'S DILEMMA" OR "NON-CONSTANT-SUM" IN SUBJECT TERMS) AND ("AGENTS" OR "ARTIFICIAL INTELLIGENCE" OR AI IN ABSTRACT) AND ("COOPERATION" OR "COOPERATE" OR "WIN-WIN" IN ABSTRACT)

The databases used in this paper are Web of Science and IEEE Xplore, respectively, and the search results are shown in Table 2.

Table 2. Database, search statements and search results

Source	Search string	Result
Web of Science	TS=("variable-sum" OR "Prisoner's Dilemma" OR "non-constant-sum") AND AB=("agents" OR "Artificial Intelligence" OR AI) AND AB=("Cooperation" OR "cooperate" OR "Win-Win") Filter used:2017.01.01-2023.01.31	124
IEEE Xplore	("All Metadata":"variable-sum" OR "All Metadata":"Prisoner's Dilemma" OR "All Metadata":"non-constant-sum") AND ("Abstract":"agents" OR "Abstract":"Artificial Intelligence" OR "Abstract":AI) AND ("Abstract":"avoid" OR "Abstract":"circumvent" OR "Abstract":"solve") AND ("Abstract":"Cooperation" OR "Abstract":"cooperate" OR "Abstract":"Win-Win") Filters Applied: 2017 - 2023	23

In total, there were 124 Web of Science search results and 23 IEEE Xplore search results, making a total of 147 articles.

2.2 Inclusion/Exclusion Criteria

There were five inclusion criteria for the corresponding studies in this review, see Table 3.

First, it must be research related to the prisoner's dilemma; other research is not included whenever it does not meet the definition of the prisoner's dilemma.

Also, the study must be for surrogate research, and primary work in examining humans or other organisms should be excluded.

Other studies, such as those analyzing the effects of the prisoner's dilemma and the social problems that may result from particular strategies, need to be removed.

Finally, all review articles and articles written in languages other than English were also excluded from inclusion.

Table 3. Inclusion/Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Studies on Prisoner's Dilemma	Not studies on Prisoner's Dilemma
Studies on Artificial Intelligence	Studies on humans or other organisms
An increase in cooperation can be observed in the results	Not increased level of cooperation in the results.
Empirical studies	Literature reviews, commentaries or meta-analysis
Written in English	Written in other languages

2.3 Quality Assessment

Based on the Inclusion/Exclusion Criteria of 2.2, 31 studies were finally screened in this paper. The specific screening process is shown in Figure 1.

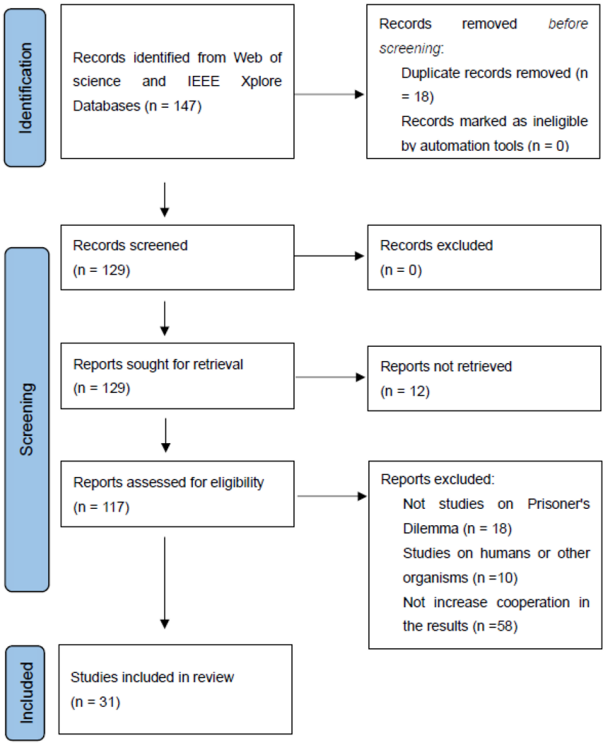

Fig. 1. Screening flow chart

3 Results

3.1 What methods are used in these studies to circumvent the prisoner's dilemma?

3.1.1 Bounded rationality

Among the 31 studies, 5 (Xu, et al., 2018, May; Moriyama, et al., 2017, July; Wang & Jiang， 2019; Otsuka & Sugawara, 2017, August; Otsuka & Sugawara, 2018) [25, 26, 27, 37, 43] used the Bounded rationality method.

If rationality is the cause that drives agents into a prisoner's dilemma, then limited rationality can help to alleviate it.

A low rationality level implies that it is possible to choose a random strategy instead of the strategy with the highest payoff (Xu, et al., 2018, May) [43].

Based on this definition, Bounded rationality includes allowing the agent to make mistakes (Moriyama, et al., 2017, July) [25], as well as all artificially set strategies of the game, such as Tit for Tat (TFT) (Wang & Jiang, 2019) [37], Extended Expectation-of-Cooperation (EEoC) (Otsuka & Sugawara, August 2017; Otsuka & Sugawara, 2018) [26,27] and others.

The Tit for Tat (TFT) strategy is a classic approach that enables an agent to imitate the opponent's decision from the previous round, after the initial round of cooperation. On the other hand, the Extended Expectation-of-Cooperation (EEoC) strategy involves cooperating multiple times when agents encounter cooperation, regardless of whether the object changes.

3.1.2 Memory

Among the 31 studies, 6 (Wang, et al., 2016, August; Fernández-Domingos, et al., 2017; Wang, et al., 2017; Heller & Mohlin, 2018; Lotfi & Rodrigues, 2022; Tao, et al., 2022, February) [10, 15, 23, 34, 38, 39] used the Memory method.

With memory, agents can be less concerned with immediate interests. Agents in the population update their strategies not by all the previous interactions but decided by the length of history records (Wang, et al., 2017) [39].

Alternatively, the agent can be allowed to remember the results of previous games between its opponent and itself to choose this game choice (Tao, et al., 2022, February) [34], or to choose whether to interrupt the link with an agent that is less cooperative (Fernández-Domingos, et al., 2017) [10].

Of course, memory here also refers to the fact that the agent can know the choices made by other agents in the past few rounds of the game (whether or not the game was played with itself) and decide its own strategy based on these (Lotfi & Rodrigues, 2022; Wang, et al., 2016, August; Heller & Mohlin, 2018) [15, 23, 38].

3.1.3 Adaptive strategy

Among the 31 studies, 6 (Xue, et al., 2017; Shang, et al., 2021, July; Wang, et al., 2021; Seredyński & Gąsior, 2019; Liu, et al., 2021, August; Xu & Hui, 2019) [22, 30, 31, 36, 42, 45] used the Adaptive strategy method.

In contrast to the humanly set strategy in Bounded rationality, agents can learn and adapt their strategies through reinforcement learning method (Xue, et al., 2017) [45], or adjust their strategy update rates according to the environment they are in (Shang, et al., 2021, July) [31]. For example, if an agent is surrounded by many cooperators, it will lower its strategy update rate to maintain a cooperative state.

In some cases, such as when payoffs are lower than expected (Wang, et al., 2021) [36]; payoffs are lower than those of neighbors (Xu & Hui, 2019; Seredyński & Gąsior, 2019) [30, 42] or communities (Liu, et al., 2021, August) [22], agents learn the strategy of neighbors or game objects that are better than themselves (Wang, et al., 2021; Xu & Hui, 2019; Seredyński & Gąsior, 2019) [30, 36, 42] or the strategy of the best individual in the community (e.g., neighbors of neighbors) (Multi-hop learning) (Liu, et al., 2021, August) [22] to update its strategy.

3.1.4 Mood model

Among the 31 studies, 5 studies (Collenette, et al., 2017, September; Feehan & Fatima, 2022; Collenette, et al., 2019, July; Zeng, et al., 2017; Zeng & Li, 2020)[5, 6, 8, 47, 48] added for agent Mood model.

Unlike Bounded rationality, the degree of rationality of an agent will be determined by an emotion value under the mood model.

Agent has an expectation of payoff, and if the actual payoff is lower than that expectation, its mood value decreases, and a lower mood value means more rationality (Collenette, et al., 2017, September) [5] and more risk seeking (Zeng, et al., 2017; Zeng & Li, 2020) [47, 48], and vice versa.

When the mood is high the agent cooperates with new agents, and when the mood is very high, the agent will always cooperate (Collenette, et al., 2019, July) [6].

The methods used to compare payoffs can be divided into two categories (Zeng, et al., 2017; Zeng & Li, 2020) [47, 48], historical-comparison and social-comparison.

For historical comparison, the expectation of payoffs depends on the agent's past payoffs (Zeng, et al., 2017; Zeng & Li, 2020; Collenette, et al., 2017, September; Feehan & Fatima, 2022) [5, 8, 47, 48].

For social comparison, the expectation of benefits depends on the average of the benefits of the agent's neighbors or the community (Zeng, et al., 2017; Zeng & Li, 2020; Collenette, et al., 2019, July) [5, 6, 47, 48].

3.1.5 Intrinsic reward

Among the 31 studies, 3 different Intrinsic reward approaches (Yuan, et al., 2022; Wu, et al., 2017; Fan, et al., 2022) [7, 40, 46] were proposed to induce cooperation.

An intrinsic reward is an additional reward that is independent of the game, but this reward is involved in the judgment of the game choice. While it seems to be no different than directly modifying the payoff table of the game, they tended to be based on some sociological research that conferred some known psychological need for satisfaction on agents.

Using adherence as an intrinsic reward enables agents to consider the collective, thus promoting cooperation (Yuan, et al., 2022) [46].

Assume that each agent has an Internal-standard, a behavior that they consider to be the right. If their behavior is consistent with the Internal-standard, they receive an additional reward; if it is not, they deduct the reward (Wu, et al., 2017) [40].

Using social payoff as an intrinsic reward enables agents to consider the collective (Fan, et al., 2022) [7].

3.1.6 Network dynamics

Among the 31 studies, 5 (Takesue, 2018; Takesue, 2021; Ichinose, et al., 2018; Li, et al., 2020; Guo, et al., 2022) [14, 17, 21, 32, 33] used the Network dynamics method.

In the Network dynamics approach, agents are allowed to change their game objects in different ways, so this approach can also be regarded as "partner selection".

Agents can choose to interrupt the link with an uncooperative object (i.e., refuse to game) and create a new link (Takesue, 2018; Takesue, 2021) [32, 33], or agents can choose to move to an empty node in the network to game with a new neighbor (Ichinose, et al., 2018) [17].

There are also some more specific partner selection mechanisms:

The agents will send a signal with strength to each other, and they will play if both sides receive each other's signal, they will not play as long as one side's signal is not delivered to the other side. The weaker the signal strength is, the harder it will be delivered to the other side. At this

point, allowing agents to adjust the strength of the signal based on their own payoffs (Li, et al., 2020) [21] is also one of the approaches of Network dynamics.

A concept of active or inactive can also be introduced to the agents, so that if the agents are inactive, they will stop playing with any neighbor. In each round, agents select the action of being active or inactive via iterative Q-table (Guo, et al., 2022) [14].

3.1.7 Altruistic attribute

One (Wu, et al., 2018) [41] of the 31 studies mentioned Altruistic attribute.

Several altruistic agents are designed for the network. When an altruistic agent cooperates, its neighbors, regardless of their strategies, can gain additional benefits (Wu, et al., 2018) [40].

3.2 Which environments are these methods applicable to?

For narrative convenience, all environments in this paper are divided into only 2 cases: IPD and Network.

IPD stands for Iterative Prisoner's Dilemma, although it usually includes the game of iterative prisoner's dilemma among multiple agents, in this paper, it refers specifically to the game between 2 agents.

Network refers to all prisoner's dilemma games between more than 2 agents, which includes various network forms such as square lattices network, scale-free network, etc., and includes all game models with partner selection. NIPD (The agent plays with all its neighbors one at a time and has the choice of cooperating with all or none) is also among them.

Among the 31 studies that used Bounded rationality, there were 26 applied to Network and 5 applied to IPD.

Details as follows:

3.2.1 Bounded rationality

Among the 5 studies that used Bounded rationality, there were 4 (Xu, et al., 2018, May; Otsuka & Sugawara, 2017, August; Wang & Jiang, 2019; Otsuka & Sugawara, 2018) [26, 27, 37, 43] applied to Network and 1 (Moriyama, et al.,2017, July) [25] applied to IPD.

3.2.2 Memory

Among the 6 studies that used Memory, there were 5 (Tao, et al., 2022, February; Fernández-Domingos, et al., 2017; Lotfi & Rodrigues, 2022; Wang, et al.,2016, August; Wang, et al., 2017) [10, 34, 38, 39] applied to Network and 1 (Heller & Mohlin, 2018) [15,23] applied to IPD.

3.2.3 Adaptive strategy

All of the 6 studies (Xue, et al., 2017; Shang, et al., 2021, July; Wang, et al., 2021; Seredyński & Gąsior, 2019; Liu, et al., 2021, August; Xu & Hui, 2019) [22, 30, 31, 36, 42, 45] that used Adaptive strategy were applied to Network.

3.2.4 Mood model

Among the 5 studies that used Mood model, there were 2 (Feehan & Fatima, 2022; Collenette, et al., 2019, July) [6,8] applied to Network and 3 (Collenette, et al., 2017, September; Zeng, et al., 2017; Zeng & Li, 2020) [5, 47, 48] applied to IPD.

3.2.5 Intrinsic reward

All of the 3 studies (Yuan, et al., 2022; Wu, et al., 2017; Fan, et al., 2022) [7, 40, 46] that used Intrinsic reward were applied to Network.

3.2.6 Network dynamics

All of the 5 studies (Takesue, 2018; Takesue, 2021; Ichinose, et al., 2018; Li, et al., 2020; Guo, et al., 2022) [14, 17, 21, 32, 33] that used Network dynamics were applied to Network.

3.2.7 Altruistic attribute

The only study (Wu, et al., 2018) [41] that used Altruistic attribute was applied to Network.

3.3 How effective are these methods?

3.3.1 Bounded rationality

Bounded rationality only makes it possible that the cooperators are not all destroyed (Xu, et al., 2018, May) [43]. Agent will try to cooperate several times in a short period of time (Moriyama, et al.,2017, July) [25].

EEoC make it possible to spread cooperation rather than generate it (Otsuka, et al.,2017, August) [26] and if the number of ProbD (A strategy of always choosing defects) agents is not large, EEoC agents can spread and maintain mutual cooperation (Otsuka & Sugawara, 2018) [27].

TFT can promote the cooperative behavior of agents in a multi-agent system and improve the overall benefit of the system (Wang & Jiang, 2019) [37].

3.3.2 Memory

Memory can increase cooperation frequency substantially (Tao, et al., 2022, February; Wang, et al., 2016, August) [34, 38], but requires the network to be as dense as possible (Fernández-Domingos, et al., 2017) [10], because if game partners are randomly matched and difficult to match again with the same agent, cooperation cannot be induced (Heller & Mohlin, 2018) [15].

For memory length (the number of innings that can be remembered), the longer the memory length, the more it promotes cooperation (Lotfi & Rodrigues, 2022) [23], while there exists an optimal memory length to develop cooperation (Wang, et al., 2017) [39].

3.3.3 Adaptive strategy

By updating the strategy with Adaptive strategy, agents were able to cooperate with their opponents without losing competitiveness (Xue, et al.,2017) [45] and, in some cases, can even make it the only stable state (Shang, et al.,2021, July) [31].

If it is limited to imitating the strategies of other superior agents, Adaptive strategy can promote the formation and maintenance of cooperation, especially when there is a significant payoff difference between cooperators and defectors (Seredyński & Gąsior, 2019; Wang, et al., 2021) [30, 36]. Multi-hop learning can enhance cooperation, and there is an optimal hop number (Liu, et al., 2021, August) [22].

With the extreme tendency to imitate the superior agents, cooperators can dominate a limited population, and the level of cooperation increases with population size (Xu & Hui, 2019) [42].

3.3.4 Mood model

Positive moods can facilitate and achieve a stable cooperation (Collenette, et al., 2017, September; Collenette, et al., 2019, July) [5,6], and even could reach high levels of cooperation (Zeng & Li, 2020) [47]. Negative emotions are detrimental to cooperation, but exploitation can be avoided in extreme environments (Collenette, et al., 2017, September).

The import of mood models can facilitate cooperation (Collenette, et al., 2017, September; Collenette, et al., 2019, July; Zeng & Li, 2020; Feehan & Fatima, 2022; Zeng, et al., 2017) [5, 6, 8, 47, 48], however, highly connected networks will reduce effectiveness (Feehan & Fatima, 2022) [8].

With Historical-comparison a high level of cooperation can be achieved and with Social-comparison the high level of cooperation can only be sustained in a portion of the population (Zeng, et al., 2017) [48].

3.3.5 Intrinsic reward

Intrinsic reward can enhance cooperation (Yuan, et al., 2022) [46], significantly affect the outcome of cooperation evolution (Wu, et al., 2017) [40], and ensure cooperators coexist with defectors even at high temptations to defection (Fan, et al., 2022) [7].

3.3.6 Network dynamics

Network dynamics enhance the cooperation in the prisoner's dilemma (Takesue, 2021; Li, et al., 2020) [21, 33], networks with medium density (a higher number of links between agents playing the game) can increase cooperation (Ichinose, et al., 2018) [17], and cooperation can be the best strategy when the density of the network increases (Takesue, 2018) [32].

For games that introduce the concept of signal strength (Li, et al., 2020) [21], Network dynamics can not only help cooperators escape from the risk of extinction but also can greatly contribute to the population size of cooperators.

For games that introduce active and inactive states (Guo, et al., 2022) [14], Network dynamics can make Inactive agent form a belt to separate cooperative and defective clusters.

3.3.7 Altruistic attribute

Altruistic attribute promoted cooperation (Wu, et al., 2018) [41].

4 Discussion

4.1 Summary

This paper summarizes 31 studies from January 2017 to January 2023 on how to make AGENT cooperate in a prisoner's dilemma and summarizes seven methods: Bounded rationality; Memory; Adaptive strategy; Mood model; Intrinsic reward; Network dynamics; and Altruistic attribute.

The Boundary rationality approach enables agents to engage in certain "prescribed" behaviors, often of a perceptual nature. Numerous strategies have been proposed based on this idea, including the TFT strategy. This perceptual strategy disregards the "benefits of Temptation" that may not provide maximum returns in the short term but can enhance overall benefits in the long run.

In general, not all agents use the same strategy; they are often placed into populations with a certain proportion of ProbD and ProbC, and the proportion of mutual cooperation in the whole population is observed after several rounds of "game → elimination → reproduction → game". Therefore, this approach is more akin to "spreading" cooperation than "generating" cooperation.

The exception to this is allowing agents to make mistakes, which does generate cooperation. If agents are allowed to learn, they will reach cooperation on occasional mutual mistakes and as a result learn to cooperate for a short period of time, somewhat like the operant conditioning chamber (McLeod, 2015) [24], where cooperation is more rewarding and becomes an incentive due to the long-term environment of mutual betrayal.

Unlike the fixed strategy of Bounded rationality, both Adaptive strategies and Memory allow agents to change their strategies through learning. They differ only in the agent's ability to observe the outcome of past rounds of the game.

When an agent holds Memory, it can learn from past gains and learn to cooperate without being tied to immediate benefits, since multiple Reward gains are more than Punishment gains. If agents can also observe other agents' past game choices, they can learn richer strategies to gain more while avoiding exploitation; a simple example is the TFT strategy that can be learned when Memory length is 1 (Anastassacos, et al., 2020, April)[1]. Memory is not only the agent's own, but also can be the whole society's. In fact, in real life, criminal records are an example of social memory means, and employers can judge whether to make employment (cooperation) based on criminal records.

Learning from immediate gains alone cannot produce cooperation (Axelrod, 1980) if Memory is not included[2]. Thus, the best approach is to learn the strategies of "better" performing agents, but this requires that agents observe the total gains of other agents, and agents with the highest gains tend to be able to exploit or cooperate with others over time, and exploitation is not sustainable when agents are able to change their strategies to avoid being exploited. When agents are able to change their strategies to avoid being exploited, the exploitation is not sustainable, and agents end up learning from those who cooperate in the long-term.

However, this method needs to ensure that cooperators do not disappear prematurely. If agents can observe the highest global gainers and all turn to exploitation, then cooperators will die out at an early stage and finally become unavailable for learning, which is one of the reasons why all six studies using Adaptive strategies only learn from their neighbors.

The Mood model is actually a kind of bounded rationality, but it is classified as an independent approach due to its dynamic adjustment. It is commonly believed that rationality makes agents choose the optimal solution in a prisoner's dilemma (Campbell & Sowden 1985) [3]- Defection. And the Mood model is set up in such a way that mood becomes low when agents feel their gains decrease and high when agents feel their gains increase. When sentiment is high, agents will be less "calculating" and ignore the optimal solution in favor of cooperation. At the same time, when they are exploited to reduce their gains, they "calm down" and return to rational thinking to avoid further exploitation.

If it is in a two-player IPD, when both parties are in a high mood state, both parties' gains will increase, which in turn will further increase the mood value. In fact, the TFT strategy can also be seen as an extreme Mood model, i.e., if the opponent cooperates, whether they choose to cooperate or defect, the increase in gain will lead to a rapid increase in the mood value and thus the choice to cooperate, and similarly, if they encounter a defection, they will immediately feel frustrated because of the decrease in gain. Thus, the disadvantage of the mood model is very obvious, when the number of participants increases, once someone chooses to defect when the mood value is low, it will quickly break the mutual cooperation and cannot stop the propagation of defections.

This is why sentiment models are asked to be performed in a less connected network. social-comparison performs better than historical-comparison for this same reason, because defections may be passed through the network, so the average gain for the society as a whole is reduced, which makes it less likely to compare with past gains, compared to see a decrease in gains to maintain good mood.

The final method related to the rationality level is Intrinsic reward, which allows agents to use some psychological satisfaction as a reward for participating in the prisoner's dilemma. From a global perspective, these agents act as if they have many "obsessions" in general, such as setting agents to be inclined to cooperate, so that when agents choose to cooperate, they receive an additional reward for whatever their opponent chooses, which will undoubtedly increase cooperation. Although at first glance this appears to be a "cheating" strategy because it directly modifies the payoff table and undermines the definition of the prisoner's dilemma, in reality, all human-set strategies are the same, such as the TFT strategy, which is a strategy that ignores the payoff table. And usually these intrinsic rewards are based on psychological findings or guesses, so they are also considered as a method.

Network dynamics is also a strategy, but it is fundamentally different from the previously mentioned strategies because the strategy is not a decision to cooperate or defect, but a decision to play or not to play with another agent, namely partner selection. Similar to Memory, agents can observe other agents' past choices or changes in their own past gains and choose whether to continue playing with a particular agent.

If the model is "game → elimination → reproduction → game", ProbD will end up being eliminated because no partner is willing to play with him and he cannot gain benefits as a result; if the model is with learning, the agent will know that he can gain more benefits only if he is allowed to play and learn to cooperate. Network dynamics disguises itself as a "self-protection" means, which is also a means of punishment. The example of "criminal record" is mentioned in the discussion of memory, but it is closer to the case of Network dynamics, because the employer does not choose to defect, but refuses to play together.

The last method is Altruistic property, i.e., adding some special agents to the game, and when playing with these special agents, if the special agent chooses to cooperate, the opponent will get an additional reward regardless of what the opponent chooses. Such agents are easy targets for exploitation, so it is necessary for agents to learn self-preservation strategies. Once such a special individual does not choose to cooperate, then his opponent does not receive an additional reward.

When the extra reward is large enough, Reward gains will exceed Suckers gains, so other agents will choose to cooperate. Similar to Intrinsic rewards, it alters the payoff table, but unlike Intrinsic rewards, the Altruism property alters the opponent's payoff, not his own. Even though the additional payoff is not large, other agents may choose to cooperate in order to receive additional payoffs in the long run and avoid special agents turning to defection. Because the method requires the observation of long-term outcomes, it requires a memory function.

In summary, after reviewing the seven methods for generating cooperation among agents in the 31 studies, it is easy to see that there is some interoperability among the seven methods, some of which are even variants of another method (e.g., Intrinsic rewards are complex manifestations of Bounded rationality). However, this paper distinguishes them as different methods for ease of description and also because they are all set up from different human perspectives. The various strategies in Bounded rationality are similar to the various "must follow" rules we were taught as children, Memory is like the records we acquire (criminal records, awards, etc.), Adaptive strategies are learning, Mood models are, as the name implies, emotional, Intrinsic rewards are psychological satisfaction, Network dynamics are the right to choose

partners, and Altruistic attributes are like social welfare policies.

Some of them can generate cooperation (Memory, Adaptive strategy, Intrinsic reward), some can maintain it (Network dynamics, Altruistic attribute), and some can spread it (Bounded rationality, Mood model).

Returning to the three questions that this paper seeks to answer:

1. What methods are used in these studies to circumvent the prisoner's dilemma?

In total, there are seven methods to get agents to produce cooperation from the prisoner's dilemma. They are: Bounded rationality; Memory; Adaptive strategy; Mood model; Intrinsic reward; Network dynamics; Altruistic attribute.

2. Which environments are these methods applicable to?

All seven methods have been used in multi-agent networks.

Where Bounded rationality; Memory and Mood model were also used in IPD with only 2 agents.

3. How effective are these methods?

The study showed that Memory, Adaptive strategy, and Intrinsic reward performed well in generating cooperation, Network dynamics, Altruistic attribute was suitable for maintaining the generated cooperation, and Bounded rationality, Mood model could spread the cooperation out.

4.2 Future work

In future work, further research will be conducted to unify the seven methods summarized in this paper, in order to model the generation, maintenance, and propagation of cooperation and simulate real-life human-generated cooperation - such as the development of concepts like fishing moratoriums.

The strengths of these seven approaches will be integrated and explored, with a particular focus on extending the mood modeling component. This will enable agents to spontaneously generate emotions based on their environment and develop corresponding strategies, thereby creating a more realistic social environment.

Additionally, this review will be regularly updated to incorporate any new methods that emerge.

Acknowledgement

This research was conducted with the guidance of several members of the Yoshikawa Research Office, and I have greatly benefited from their valuable advice and opinions.

Additionally, this study received support from both the Shoji Kawashima Memorial Scholarship Fund and the Tokyo Tech Tsubame Scholarship for Doctoral Students.

Reference

1. Anastassacos, N., Hailes, S., & Musolesi, M. (2020, April). 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, No. 05, pp. 7047-7054).
2. Axelrod, R. (1980). Effective choice in the prisoner's dilemma. Journal of conflict resolution, 24(1), 3-25.
3. Campbell, R., & Sowden, L. (Eds.). (1985). Paradoxes of rationality and cooperation: prisoner's dilemma and Newcomb's problem. UBC Press.
4. Chong, S. Y., Humble, J., Kendall, G., Li, J., & Yao, X. (2007). Iterated prisoner's dilemma and evolutionary game theory. In The Iterated Prisoners' Dilemma: 20 Years On (pp. 23-62).
5. Collenette, J., Atkinson, K., Bloembergen, D., & Tuyls, K. (2017, September). Mood modelling within reinforcement learning. In ECAL 2017, the Fourteenth European Conference on Artificial Life (pp. 106-113). MIT Press.
6. Collenette, J., Atkinson, K., Bloembergen, D., & Tuyls, K. (2019, July). Stability of cooperation in societies of emotional and moody agents. In Artificial Life Conference Proceedings (pp. 467-474). One Rogers Street, Cambridge, MA 02142-1209, USA journals-info@ mit. edu: MIT Press.
7. Fan, L., Song, Z., Wang, L., Liu, Y., & Wang, Z. (2022). Incorporating social payoff into reinforcement learning promotes cooperation. Chaos: An Interdisciplinary Journal of Nonlinear Science, 32(12), 123140.
8. Feehan, G., & Fatima, S. (2022). Augmenting Reinforcement Learning to Enhance Cooperation in the Iterated Prisoner's Dilemma. In ICAART (3) (pp. 146-157).
9. Felkins, L. (2001). The Prisoner's Dilemma.
10. Fernández-Domingos, E., Loureiro, M., Alvarez-López, T., Burguillo, J. C., Covelo, J., Peleteiro, A., & Byrski, A. (2017). Emerging Cooperation in N-Person Iterated Prisoner's Dilemma over Dynamic Complex Networks. Computing & Informatics, 36(3).
11. Fujimoto, Y., & Kaneko, K. (2019). Emergence of exploitation as symmetry breaking in iterated prisoner's dilemma. Physical Review Research, 1(3), 033077.
12. Gill, D., & Rosokha, Y. (2020). Beliefs, learning, and personality in the indefinitely repeated prisoner's dilemma. Available at SSRN 3652318.
13. Gotts, N. M., Polhill, J. G., & Law, A. N. R. (2003). Agent-based simulation in the study of social dilemmas. Artificial Intelligence Review, 19, 3-92.
14. Guo, H., Wang, Z., Song, Z., Yuan, Y., Deng, X., & Li, X. (2022). Effect of state transition triggered by reinforcement learning in evolutionary prisoner's dilemma game. Neurocomputing, 511, 187-197.
15. Heller, Y., & Mohlin, E. (2018). Observations on cooperation. The Review of Economic Studies, 85(4), 2253-2282.
16. Hofstadter, D. R. (1983). Metamagical themas. Scientific American, 248(5), 16-E18.

17. Ichinose, G., Satotani, Y., & Nagatani, T. (2018). Network flow of mobile agents enhances the evolution of cooperation. Europhysics Letters, 121(2), 28001.
18. Jiang, J., & Lu, Z. (2018). Learning attentional communication for multi-agent cooperation. Advances in neural information processing systems, 31.
19. Kopelman, S. (2020). Tit for tat and beyond: The legendary work of Anatol Rapoport. Negotiation and Conflict Management Research, 13(1), 60-84.
20. Lazaridou, A., Peysakhovich, A., & Baroni, M. (2016). Multi-agent cooperation and the emergence of (natural) language. arXiv preprint arXiv:1612.07182.
21. Li, J., Park, J. H., Zhang, J., Chen, Z., & Dehmer, M. (2020). The networked cooperative dynamics of adjusting signal strength based on information quantity. Nonlinear Dynamics, 100(1), 831-847.
22. Liu, X., Guan, R., Wang, T., Han, L., Qin, Y., & Wang, Y. (2021, August). Multi-hop Learning Promote Cooperation in Multi-agent Systems. In Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Tokyo, Japan, August 14–16, 2021, Proceedings, Part I (pp. 66-77). Cham: Springer International Publishing.
23. Lotfi, N., & Rodrigues, F. A. (2022). On the effect of memory on the Prisoner's Dilemma game in correlated networks. Physica A: Statistical Mechanics and its Applications, 607, 128162.
24. McLeod, S. (2015). Operant Conditioning: What It Is, How It Works, and Examples.
25. Moriyama, K., Nakase, K., Mutoh, A., & Inuzuka, N. (2017, July). The resilience of cooperation in a Dilemma game played by reinforcement learning agents. In 2017 IEEE International Conference on Agents (ICA) (pp. 33-39). IEEE.
26. Otsuka, T., & Sugawara, T. (2017, August). Robust spread of cooperation by expectation-of-cooperation strategy with simple labeling method. In Proceedings of the International Conference on Web Intelligence (pp. 483-490).
27. Otsuka, T., & Sugawara, T. (2018). Promotion of robust cooperation among agents in complex networks by enhanced expectation-of-cooperation strategy. In Complex Networks & Their Applications VI: Proceedings of Complex Networks 2017 (The Sixth International Conference on Complex Networks and Their Applications) (pp. 815-828). Springer International Publishing.
28. Rapoport, A. (1989). Prisoner's dilemma. Game theory, 199-204.
29. Sandholm, T. W., & Crites, R. H. (1996). Multiagent reinforcement learning in the iterated prisoner's dilemma. Biosystems, 37(
30. Seredyński, F., & Gąsior, J. (2019). Emergence of collective behavior in large cellular automata-based multi-agent systems. In Artificial Intelligence and Soft Computing: 18th International Conference, ICAISC 2019, Zakopane, Poland, June 16–20, 2019, Proceedings, Part II 18 (pp. 676-688). Springer International Publishing.
31. Shang, L., & Luo, H. (2021, July). Environmental adaptability promotes cooperation in the evolutionary game. In 2021 40th Chinese Control Conference (CCC) (pp. 7486-7491). IEEE.
32. Takesue, H. (2018). Evolutionary prisoner's dilemma games on the network with punishment and opportunistic partner switching. Europhysics Letters, 121(4), 48005.
33. Takesue, H. (2021). Symmetry breaking in the Prisoner's Dilemma on two-layer dynamic multiplex networks. Applied Mathematics and Computation, 388, 125543.
34. Tao, W., Wei, W., Xin, Y., & Meiqi, H. (2022, February). Strategies to Promote Cooperation in Mobile Networks. In 2022 8th International Conference on Automation, Robotics and Applications (ICARA) (pp. 140-145). IEEE.
35. Tucker, A. W., & Straffin Jr, P. D. (1983). The mathematics of Tucker: A sampler. The Two-Year College Mathematics Journal, 14(3), 228-232.
36. Wang, S. Y., Liu, Y. P., Zhang, F., & Wang, R. W. (2021). Super-rational aspiration induced strategy updating promotes cooperation in the asymmetric prisoner's dilemma game. Applied Mathematics and Computation, 403, 126180.
37. Wang, S., & Jiang, L. (2019). Study of Agent Cooperation Incentive Strategy Based on Game Theory in Multi-Agent System. In Communications, Signal Processing, and Systems: Proceedings of the 2017 International Conference on Communications, Signal Processing, and Systems (pp. 1871-1878). Springer Singapore.
38. Wang, T., Li, L., Zhang, S., Peng, H., Yu, L., & Chen, Z. (2016, August). Memory mechanism enhances cooperation in mobile multi-agent system. In 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) (Vol. 2, pp. 476-479). IEEE.
39. Wang, X., Zhang, L., Du, X., & Sun, Y. (2017). Evolving cooperation in spatial population with punishment by using PSO algorithm. Natural Computing, 16, 99-117.
40. Wu, Y. E., Zhang, Z., & Chang, S. (2017). Effect of self-interaction on the evolution of cooperation in complex topologies. Physica A: Statistical Mechanics and its Applications, 481, 191-197.
41. Wu, Y. E., Zhang, Z., & Chang, S. (2018). Heterogeneous indirect reciprocity promotes the evolution of cooperation in structured populations. Chaos: An Interdisciplinary Journal of Nonlinear Science, 28(12), 123108.
42. Xu, C., & Hui, P. M. (2019). Emergence of cooperation in finite populations under biased selection. Physica A: Statistical Mechanics and its Applications, 535, 122371.

43. Xu, X., Rong, Z., & Tse, C. K. (2018, May). Bounded rationality optimizes the performance of networked systems in prisoner's dilemma game. In 2018 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 1-5). IEEE.
44. Xuan, P., Lesser, V., & Zilberstein, S. (2001, May). Communication decisions in multi-agent cooperation: Model and experiments. In Proceedings of the fifth international conference on Autonomous agents (pp. 616-623).
45. Xue, L., Sun, C., Wunsch, D., Zhou, Y., & Yu, F. (2017). An adaptive strategy via reinforcement learning for the prisoner's dilemma game. IEEE/CAA Journal of Automatica Sinica, 5(1), 301-310.
46. Yuan, Y., Guo, T., Zhao, P., & Jiang, H. (2022). Adherence Improves Cooperation in Sequential Social Dilemmas. Applied Sciences, 12(16), 8004.
47. Zeng, W., & Li, M. (2020). Selective attention to historical comparison or social comparison in the evolutionary iterated prisoner's dilemma game. Artificial Intelligence Review, 53, 6043-6078.
48. Zeng, W., Li, M., & Feng, N. (2017). The effects of heterogeneous interaction and risk attitude adaptation on the evolution of cooperation. Journal of Evolutionary Economics, 27, 435-459.