# 世界モデルを用いた深層マルチエージェント強化学習における囚人のジレンマ環境下での協力の創発

著者名

November 21, 2025

**Abstract**

本研究では、深層マルチエージェント強化学習 (DMARL) における協力の創発を、囚人のジレンマ (PD) のようなジレンマ構造が明確な環境下で検証する。既往研究では世界モデルを用いた反事実想像により協調行動の獲得が報告されているが、用いられる環境にはゲーム理論的なジレンマ性が弱いものが含まれる。本研究では、Coin Game のような高次元観測を持つ PD 型環境において、世界モデルが協力創発に寄与しうるか、またそのために必要な拡張 (反事実評価・クレジット割当・因果的報酬成形等) は何かを検討する。

## 1 Introduction

独立に行動する AI エージェントが本質的に互恵的な性質を備えることは、社会に破滅的な影響を及ぼさないための重要条件である。一方で、AI エージェントは基本的に自己利益的であり、報酬最大化を目標として動作する。

このギャップを埋めるための手法は、古くからゲーム理論の枠組みで研究され、Reputation、Image Score、シグナリングなどが提案されてきた。2000 年代初頭の研究では、エージェントは単純な戦略を用い、レプリケータダイナミクス等を通じて進化ゲーム論の文脈で協力の創発が論じられてきた。

近年は、深層学習・強化学習の進展により、深層マルチエージェント強化学習 (DMARL) と協力の創発に関する研究が急速に活発化している。DMARL では、複数の自律エージェントが環境内で相互作用しながら、それぞれの目的を達成する方策を学習する。しかし、最適方策の獲得は次の理由から困難である。

1. **非定常性**：各エージェントの方策が同時に更新されるため、単一エージェントの MDP で想定される定常性仮定が破られる。
2. **計算複雑性**：次状態の予測には他エージェントの行動推論が不可欠であり、エージェント数の増加とともに複雑さが増大する [1]。

これらの問題に対処する一つの方向として、**世界モデル (World Model)** の応用が進んでいる。たとえば Chai らは、DreamerV2 を拡張し、世界モデル上で反事実想像 (Counterfactual Imagination) を行うことで、複数エージェントの協調を促し、効率的な方策の獲得を報告している [2]。

しかし、既往研究で用いられる環境にはゲーム理論的なジレンマ性が弱いものが含まれる。たとえば HalfCheetah のように役割分担 (前脚・後脚) がある設定では、両者が前進を選ぶときの利得 $R$ が、一方のみが前進 (相手は別行動) したときの利得 $T$ より大きい ($R > T$) と考えられ、**囚人のジレンマ (PD)** に見られる $T > R$ の関係が成立しない。したがって、**明確なジレンマ構造 (PD 型)** における協力創発の検証としては不十分である。

本研究では、PD のようにジレンマ構造が明確な環境において、世界モデルが協力の創発に寄与しうるか、またそのために必要な**拡張 (反事実評価・クレジット割当・因果的報酬成形等)** は何かを検討する。環境としては、Coin Game のように構造はシンプルだが、高次元観測 (例：$32 \times 32$ のマップ) を持ち、PD 的な利得構造を明示的に設計できる設定を採用する [3]。

# 2 Related Work

# 3 Preliminaries

## 3.1 Multi-Agent Reinforcement Learning

## 3.2 World Models

## 3.3 Game Theory and Prisoner's Dilemma

# 4 Methodology

## 4.1 Environment: Coin Game

### 4.1.1 State Representation and Agent Identification

本研究では、環境として Coin Game [4] を採用する。Coin Game は、任意の数 $M$ のエージェントが $L \times L$ のグリッド上でコインを集めるマルチエージェント環境である。エージェントはそれぞれ異なる色（例：赤と青）で表現される。各ステップにおいて、グリッド上に自身か相手と同じ色のコインがそれぞれ $N$ 個ランダムに出現する。エージェントがコインの位置に移動すると、そのコインを獲得し、報酬を得る。しかし、エージェントが自分と異なる色のコインを獲得した場合、自身は報酬を得、もう一方のエージェントに罰則が与えられる。自身の色のコインを獲得した場合は、相手への罰則はない。

この報酬構造により、Coin Game は囚人のジレンマ（PD）と同様の構造を持つ。利己的なエージェントは、色に関わらず全てのコインを獲得しようとする（裏切り）が、双方がこの戦略を取ると互いに罰則を与え合い、全体としての利得は低下する。双方が自身の色のコインのみを獲得する（協力）ことで、社会的総余剰は最大化される。本研究では、この環境を画像観測（１つのグリッドセル＝ピクセル）としてエージェントに提示する。

エージェントの行動は、抽象的には「自分のコインを取る」、「相手のコインを取る」、「コインを取らない」の 3 つに分類できる。それぞれの行動に対する利得構造を表1に示す。

Table 1: Coin Game における行動と利得構造

| 行動 | 自身の利得 | 相手への利得 |
|---|---|---|
| 自分のコインを取る | $+1$ | $0$ |
| 相手のコインを取る | $+2$ | $-3$ |
| コインを取らない | $0$ | $0$ |

Prisoner's Dilemma の利得構造は以下の不等式で表される。

$$T > R > P > S \tag{1}$$

本環境における利得行列（Payoff Matrix）を表2に示す。各パラメータは $R = 1$, $T = 2$, $S = -2$, $P = -1$ となっており、囚人のジレンマの条件を満たしている。つまり、相手の戦略にかかわらず、各エージェントにとって最適な戦略は「相手のコインを取る（裏切り）」であるが、双方がこの戦略を取ると互いに罰則を与え合い、期待利得は低い。一方、双方が「自分のコインを取る（協力）」戦略を取れば、期待利得は罰則の状態に比べて大きい。

Table 2: 利得行列 (Payoff Matrix)

| | 協力 (Cooperate) | 裏切り (Defect) | コインを取らない (No-op) |
|---|---|---|---|
| 協力 (Cooperate) | $1, 1$ | $-2, 2$ | $1, 0$ |
| 裏切り (Defect) | $2, -2$ | $-1, -1$ | $2, -3$ |
| コインを取らない (No-op) | $0, 1$ | $-3, 2$ | $0, 0$ |

### 4.2 Fixed Strategy Agents

深層学習ベースのエージェントに加えて、ゲーム理論で研究されてきた古典的な固定戦略を実装したエージェントを定義し、学習エージェントとの相互作用を検証する。これにより、学習エージェントが異なる戦略パターンに対してどのように適応するかを分析できる。

#### 4.2.1 Always Defect 戦略（ALLD）

ALLD 戦略は、常に裏切り行動を選択する戦略である。Coin Game 環境においては、**最も近いコインを取る**という行動として実装される。この戦略では、コインの色に関係なく、最も近いコインを常に獲得しようとする。この戦略は短期的な利得を最大化しようとするが、相手エージェントに罰則を与え続けるため、長期的な社会的総余剰は低くなる。

#### 4.2.2 Tit-for-Tat 戦略（TFT）

最も有名な戦略は、しっぺ返し戦略 (Tit-for-Tat, TFT) である [5]。この戦略は、**相手が前回やったことと同じことをする**という単純だが効果的な戦略である。もし前回に相手が裏切り (D) を出してきたら、今回は自分も裏切り (D) を出し、相手が協力 (C) を出していたら、自分も協力 (C) を出す。

TFT 戦略には明確な利点がある。相手に裏切られたらやり返すので、**長期的な利得でどんな相手にも負けることがない**という性質がある。また、初手で協力することから「寛容さ」を示し、相互協力の成立を促進する。

しかし、TFT 戦略には明確な弱点もある。それは**エラーに対して弱く協力状態が維持できない**という点である。TFT 戦略を 2 人とも取ったとしよう。この場合、相互協力状態から始めたとしても、誰かがエラーで間違えて裏切ってしまったとする。そうすると、相手が裏切り、その次には自分が裏切り、という**裏切りの連鎖**に陥ってしまうという問題点がある

#### 4.2.3 Win-Stay-Lose-Shift 戦略（WSLS）

他に有名な戦略としては Win-Stay-Lose-Shift (WSLS) という戦略がある [6]。WSLS は、TFT と同じように前回の結果に応じて行動を変える戦略であるが、**相手の行動ではなく自分の利得**に基づいて判断する点が異なる。

具体的には、自分の利得が良い場合（$(C, C)$ または $(D, C)$ の場合）にはその行動を続け (Win-Stay)、良くなかった場合（$(C, D)$ または $(D, D)$ の場合）には行動を変える (Lose-Shift) という戦略である。

この戦略の良いところは、**エラーに対して強い**という点である。TFT とは異なり、エラーがあったとしても数ステップ後には相互協力状態に戻ることになる。これは、WSLS 戦略が結果に基づいて適応的に行動を変更するためである。

しかし、WSLS 戦略にも致命的な弱点がある。相手が常に裏切り (D) を出す ALLD 戦略のようなエージェントに対しては、**一方的に搾取されてしまう**という問題がある。WSLS 戦略は $(C, D)$ の状態で行動を変えて裏切りに転じるが、その後 $(D, D)$ となり再び行動を変えて協力に戻ってしまうため、常に不利な状態から抜け出せない。

### 4.3 Agent Architecture

エージェントのモデルとして、DreamerV3 に基づく深層強化学習エージェントを採用する。

### 4.4 World Model Architecture

## 5 Experiments

本研究では、世界モデルを用いた深層強化学習エージェントが、囚人のジレンマ構造を持つ Coin Game 環境において協力行動を獲得できるかを検証する。特に、異なる固定戦略エージェントとの相互作用を通じて、学習エージェントがどのように適応し、協力的な行動を学習するかを分析する。

## 5.1 Experimental Setup

### 5.1.1 Environment Configuration

Coin Game 環境は以下のパラメータで設定する：

- グリッドサイズ：$L = 32 \times 32$
- エージェント数：$M = 2$（1 体は学習エージェント、もう 1 体は固定戦略エージェント）
- コイン数：各色 $N = 10$ 個
- エピソード長：$T = 100$ ステップ
- 観測空間：$32 \times 32 \times 3$ (RGB 画像)
- 行動空間：5 次元離散（上下左右、静止）

利得構造は表1に示した通り

### 5.1.2 Agent Configuration

**学習エージェント**には、DreamerV3 ベースのアーキテクチャを採用する。**固定戦略エージェント**として、WSLS (Win-Stay-Lose-Shift) 戦略を実装したエージェントを用いる。WSLS 戦略は、前ステップで正の利得を得た場合は同じ行動（協力または裏切り）を継続し、負または 0 の利得の場合は行動を切り替える。Coin Game 環境における実装では、以下のように動作する：

1. 前ステップの報酬が正（$r > 0$）の場合：同じ色のターゲット（自分の色のコインまたは相手の色のコイン）を維持
2. 前ステップの報酬が非正（$r \leq 0$）の場合：ターゲットの色を切り替える
3. 選択したターゲット色の最も近いコインに向かって移動

### 5.1.3 Training Protocol

実験は以下の手順で実施する：

1. **Phase 1: WSLS 戦略との対戦**（0–500k ステップ）
    - 学習エージェントを WSLS 戦略エージェントと対戦させる
    - 学習エージェントは環境との相互作用を通じて方策を更新
    - WSLS 戦略エージェントのパラメータは固定
2. **Phase 2: 自己対戦**（500k–1M ステップ、オプション）
    - Phase 1 で学習したエージェント同士を対戦させる
    - 両エージェントが同時に学習を継続
    - 相互適応と協力の安定性を検証

各フェーズにおいて、以下の指標を記録する：

- 各エージェントの平均エピソード報酬
- 協力率（自分の色のコインを取った割合）
- 裏切り率（相手の色のコインを取った割合）
- 社会的総余剰（両エージェントの報酬の合計）

### 5.2 Results

## 6 Discussion

## 7 Conclusion

## References


[1] Wong, Annie, Thomas Bäck, Anna V. Kononova, and Aske Plaat. *Deep Multiagent Reinforcement Learning: Challenges and Directions.* Artificial Intelligence Review 56, no. 6 (2023): 5023–56. https://doi.org/10.1007/s10462-022-10299-x

[2] Chai, Jiajun. *Aligning Credit for Multi-Agent Cooperation via Model-Based Counterfactual Imagination.* New Zealand, 2024.

[3] Foerster, Jakob, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. *Learning with Opponent-Learning Awareness.* 2018.

[4] Lerer, Adam, and Alexander Peysakhovich. *Maintaining cooperation in complex social dilemmas using deep reinforcement learning.* 2018.

[5] Axelrod, Robert, and William D. Hamilton. *The Evolution of Cooperation.* Science 211, no. 4489 (1981): 1390–96. https://doi.org/10.1126/science.7466396

[6] Nowak, Martin, and Karl Sigmund. *A Strategy of Win-Stay, Lose-Shift That Outperforms Tit-for-Tat in the Prisoner's Dilemma Game.* Nature 364, no. 6432 (1993): 56–58. https://doi.org/10.1038/364056a0