予測可能性を用いたエージェント間の暗黙の社会的規範に基づく協調行動の創発

Emergence of Implicit Social Norm-Based Cooperation among Agents through Predictability

山口　拓巳[1*]　竹内　勇剛[1]
Takumi Yamaguchi[1]　Yugo Takeuchi[1]

[1] 静岡大学大学院総合科学技術研究科
[1] Graduate School of Integrated Science and Technology, Shizuoka University

Abstract: 従来の協調エージェント研究は，自己利益の最大化を前提とし，他者の行動予測や協調行動へのインセンティブ付与を通じて協調を達成してきた．しかし，このような枠組みでは，Ultimatum Game における「不公平な利益分配の拒否」に見られるような，人間特有の非合理的かつ規範的行動を十分に説明することが難しい．本研究では，このような協調行動の背後にある要因として注目される「集団内で共有される行動規範を，ある個体が自らの自己規範として取り入れる過程」である社会的規範の内面化に焦点を当て，その計算的メカニズムを検討する．本稿では，個体が社会的規範を内面化する際，自身の行動の予測可能性を高めたいという動機によって内面化が促進されるという仮説を立てた．この仮説を検証するため，自身の行動の予測可能性を向上させる欲求を報酬関数に組み込んだ強化学習エージェントを設計し，従来のモデルと比較した．すれ違いタスクを題材としたシミュレーション実験において，規範的行動の発生を分析した結果，提案モデルにおいてのみ，エージェントが一貫して片側を選択するという，集団レベルでの通行パターンの固定化が観察された．これは，行動予測可能性を高めたいという動機が社会的規範の内面化を促進することを示唆している．このことから，社会的規範の内面化の過程において予測可能性の向上という動機が重要な役割を果たす可能性が示された．一方で，本研究は限定的なシミュレーション環境を前提としているため，異質なエージェントや複雑な環境設定での再検証が今後の課題である．

1　はじめに

1.1　背景

人間は中央集権的な制度や法的ルールが存在しない状況においても，他者との衝突回避や行動調整が可能である．その背景要因の一つとして「社会的規範」が挙げられる [1]．社会的規範に関する研究は心理学，経済学，神経科学など多様な分野にまたがっており，規範遵守時の脳活動を調べた研究では，報酬処理に関わる脳領域の活性化が確認されている [2]．このことから，規範遵守そのものが効用をもたらし得ることが示唆されており，個人の利益に対して非合理的な協調行動において社会的規範が何らかの役割を果たしていることが示唆されている．そのため計算論的モデルでは「規範的行動に正の報酬を与える」という形で，規範の存在意義や創発を説明する試みが多くなされている [3][4]．

*連絡先：静岡大学大学院総合科学研究科
〒 432-8011 静岡県浜松市中央区城北 3-5-1
E-mail: yamaguchi.takumi.20@shizuoka.ac.jp

しかし，実際の人間の行動の目的は報酬最大化が直接的な目的ではないことが明らかにされている．これは一般的に遵守されるテーブルマナーや服装などの獲得報酬と関係ない行動や，Ultimatum Game において他者と報酬を分割して受け取る際，自身の損失が全くない場面で報酬を享受するだけだとしても，分割の割合が明確に不公平である場合，その受け取りを拒む行動など，他者との公平性を求める行動からも示されている [5]．このような外的な報酬からは非合理的と見做される行為に関して，Wenzel による税の支払い行動に関する研究では，罰則の強化など外部的手段よりも，「規範の内面化」を通じた遵守促進の方が有効であることが示されている [6]．ここで規範の内面化とは「ある集団の規範が自己の個人的規範として受け入れ，それに従って行動をするようになること」としている．また，集団への帰属意識が高まり，特定の集団の規範が内面化されることで，自己の利益が増加するとしても，他の集団で受け入れられている規範に対して反発が生じることも報告されている [7]．これらの知見は，社会

的規範に基づく行動の創発を理解するためには，規範遵守に外的報酬を与えるのではなく，いかにして規範の内面化が進むのかを明らかにする必要があることを示唆している．

規範の内面化に関して，Danescu らはオンラインコミュニティにおける使用語彙を言語的規範とみなし，ユーザの滞在時間との関係を分析した．その結果，個人の語彙が他のメンバーの語彙に近いほど滞在期間が長くなることを示した [8]．このとき，個人の語彙とコミュニティ平均語彙との距離は，他のメンバーが用いる語彙の「予測可能性」として解釈できる．また，人は他者から規範的情報を得る際に，自らの規範と近い他者を優先して参照する傾向があることも報告されている [9]．これらの結果は，人間が能動的に自己の予測可能性を高める方向へ行動を調整することが，規範の内面化を促進する要因である可能性を示唆している．

このような知見は，人間や他のエージェントとの相互作用を踏まえたエージェント設計に対して大きく貢献すると考えられる．近年では人工知能技術の発展に伴い，特にロボティクス領域において，農業や災害救助の現場でのロボット同士の協調行動，道路交通場面における自動運転エージェントと人間の協調設計など，様々な場面でエージェント間の相互作用を踏まえた設計が考察されている [10][11][12]．しかし，自律型エージェントによる協調設計の多くは，前述したように規範的行動や協調行動自体に正の報酬を与えることで，報酬最大化を前提としたエージェントによって実行されるものである．そのため，協調行動によって報酬の増加が見込まれない状況や，設計者が想定していない協調行動が必要な環境では，エージェントが協調的に振る舞うことが困難である．

1.2 本研究の目的

以上の考察を踏まえ，本研究では協調行動を創発可能なエージェントの設計に向けて，社会的規範の内面化が可能なエージェントの設計を検討する．ここで本稿では**エージェント自身の行動に関する予測可能性を高める動機が社会的規範の内面化を促進する**という仮説をたてる．複数のエージェントが上述した動機に基づいて行動する環境をシミュレーションすることで，社会的規範として解釈可能な安定的行動パターンが自発的に形成されるかを観察し，仮説を検証する．ここで社会的規範として解釈可能な行動とは，Henrich and Ensminger の定義に従い，(1) 他者の「行動予測」，(2) 他者の「すべき行動」に関する期待，(3) その期待に沿おうとする内発的動機づけ，の三側面を含むものとする [13]．られた結果を通じて，予測可能性が社会的規範の内面化に寄与しうるかを検証し，ロボティクスなど協調エージェント設計への応用可能性を探る．

本稿の残りの部分は以下のように構成されている．2 章ではシミュレーション実験の環境と，その環境におけるどのような行動を社会的規範の内面化として評価するのかを解説する．3 章では自身の行動に関する予測可能性を高める動機を持つエージェントの数理モデルを提案する．4 章ではシミュレーション実験を通して，提案したモデルと従来の強化学習モデルの振る舞いの差を示す．5 章では得られた実験の結果をもとに考察を述べる．

2 実験設計

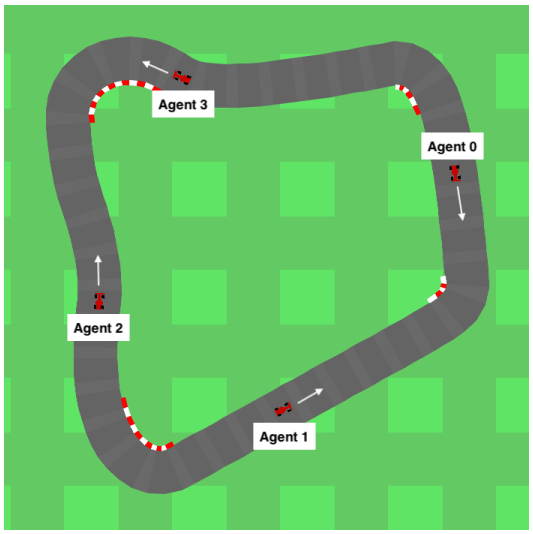
2.1 目的

本実験の目的は，エージェントの行動選択過程に「自身の行動に対する予測可能性」を導入することで，エージェント間における社会的規範の内面化が促進されるかを検証することである．具体的には，すれ違いタスクにおいて各エージェントが路面の一方（左端または右端）に偏って走行する傾向が自発的に生じるかを観察する．また，従来の報酬最大化型モデルとの比較を通じて，(1) 提案する予測可能性を導入したモデル，(2) 従来の強化学習モデル，の間で行動特性にどのような差が生じるかを明らかにする．

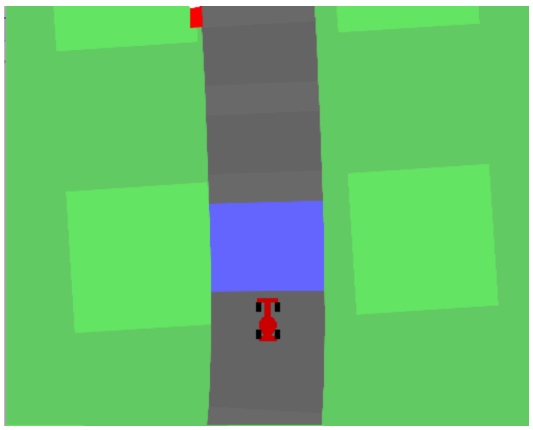
2.2 実験環境の設計

本実験では，OpenAI Gymnasium の CarRacing 環境をマルチエージェントタスクに拡張した独自環境を使用する [14]．CarRacing 環境には，閉ループコースの生成と，連続制御の車両ダイナミクスが実装されていたため，すれ違いタスクの構成に適していると判断した．実際の実験環境には単一の閉ループ型コースが存在し，すれ違い頻度と学習の安定性のバランスを考慮し，4 体の車型エージェントが同一コース上に配置される．各エージェントの目的は，コース逸脱や衝突を回避しつつ，指定された進行方向に沿って可能な限り長く走行を継続することである．進行方向は図 1a のようにエージェントごとに交互に割り当てられ，対向するエージェント同士がすれ違う状況が発生する．

観測値　各エージェントには，図 1b に示すように，自身を中心とした局所的な視界領域を画像情報として与える．本実験では「予測可能性」の導入効果を検証することを主眼とするため，タスク遂行を困難にしすぎないよう，観測画像には進行方向を示す補助ハイライトを追加した．



(a) 環境の全体像．単一のコース上に複数エージェントを配置し，白矢印が示す方向で交互に進行方向を割り当てる．



(b) 各エージェントに与えられる画像観測値．ハイライト部で進行方向を示す．

図 1: 実験環境．

報酬設計　各エージェントは，指定された方向に沿って走行した場合に正の報酬を獲得する．ただし，走行位置が路面端に寄るほど獲得報酬を低下させ（中央走行時の約半分），ランダムに端へ寄った結果を「偶発的な回避行動」と誤学習しないように設計した．これにより，端走行を「意図的な戦略」としてのみ強化できるように制約した．報酬系は環境を更新フレームごとに割り当てられ，その内訳を表 1 に示す．

行動空間　各エージェントの行動は，アクセル，ブレーキ，ステアリングの 3 次元連続値で表される．アクセルおよびブレーキはそれぞれ $[0, 1]$ の範囲で定義され，速度変化を制御する．ステアリングは $[-1, 1]$ の範囲を取り，進行方向を調整する．行動空間の要約を表 2 に示す．

表 1: 各エージェントに与えられる報酬条件

条件	報酬
指定方向への進行	$5 \sim 10$
指定方向に進まない・停止	-0.1
コース逸脱	-0.5
他エージェントとの衝突	-100

表 2: 各エージェントの行動空間

行動要素	値の範囲
アクセル	$[0, 1]$（加速）
ブレーキ	$[0, 1]$（減速）
ステアリング	$[-1, 1]$（方向制御）

2.3　比較条件

本研究では，予測可能性の導入が社会的規範の内面化を促進するかを検証するため，以下の 2 条件でシミュレーションを行う．

1. Control 条件：従来の強化学習モデル同士によるシミュレーション
2. Predictability 条件：予測可能性を考慮した提案モデル同士によるシミュレーション

各条件下でエージェントの路面位置の偏りを比較し，社会的規範の内面化の兆候（右側通行・左側通行などの固定化パターン）が生じるかを分析する．

2.4　評価指標

エージェントの学習状況や社会的規範の内面化状態を定量的に評価するため，本稿では以下の指標を計測する．

- **獲得報酬**：各エージェントがエピソード中に獲得した累積報酬を測定する．この指標は，エージェントが衝突やコース逸脱を回避しつつ進行方向に沿って効率的に走行できているかを示す．
- **走行タイル端での走行率**：エージェントが走行タイルの左端または右端を走行している程度を表す指標である．本研究では，各時刻におけるエージェントの横方向位置を $x \in [-1, 1]$ に正規化して表現し，$x = -1$ はコースを時計回りで見た場合の左端，$x = 1$ は右端を走行している状態に対応するものとする．

特に走行タイル端での走行率がどちらか一方に偏ることは，人間の右側通行のような規範的な行動と対応す

る状態である．そのため，本研究では，この指標をエージェントの社会的規範の内面化状態を直接的に反映するものとみなし，最も着目すべき指標として扱う．

2.5 実験仮説

本実験における仮説は「行動の予測可能性を導入することにより，社会的規範の内面化が促進される」ことである．具体的には，設定した各条件において以下の振る舞いが生じると予想する．

1. Control 条件：報酬最大化のみが優先されるため，報酬が減少する路面端の走行は回避され，エージェントは中央付近を走行する傾向が強まる．その結果，対向車との衝突回避は直前に行われることになり，図 2a に示すような不安定な回避動作が増加すると予想される．

2. Predictability 条件：自身の行動の一貫性を維持する（予測可能性を高める）動機づけにより，報酬の減少を許容してでも特定の路面端を走行し続ける「社会的規範」が自発的に形成・内面化される．これにより，図 2b に示すように，遠方からあらかじめ端に寄る安定した回避動作が定着すると予想される．

3 提案モデル

本研究で提案するエージェントは，通常の強化学習手法では考慮されない「自身の行動に関する予測可能性」を評価に取り入れ，これが社会的規範の内面化を促進するかを検証することを目的とする．予測可能性を扱うためには，エージェントが自身の行動の結果を内的に見積もれることが必要である．そこで本研究では，エージェントが環境との相互作用から環境モデルを学習し，将来の状態遷移を予想できるモデルベース強化学習の枠組みを採用する．

モデルベース強化学習では，学習済みの内部モデルを用いて将来の行動系列とその結果を予測できるため，エージェントは自らの行動と環境変化の関係を明示的に評価できる．本研究ではこの評価過程に「行動の予測可能性」を反映する項を新たに報酬関数へ導入し，エージェントが自発的に予測しやすい行動方針を選好するよう設計する．これにより，エージェントが予測可能性を高めることで固定的な行動パターン（社会的規範）が内面化される過程を再現できると考える．

3.1 行動の予測可能性の定義

本研究では，エージェントが「自身の行動を他者からどの程度予測しやすいか」を内的に評価できるようにする．そのために，行動の「予測可能性」を定量的に定義する必要がある．この定量化された指標は，最適制御理論の一分野である情報理論的制御理論に基づき，その一部を行動の予測可能性として既存の強化学習手法に取り入れる [15].

ここでは，エージェントの行動に伴うコスト全体を自由エネルギーと呼ばれる指標で表し，この値を最小化するように制御を行う．自由エネルギーの下限は実際に行った行動によるコスト $S(V)$ と，行動のばらつき（不確実性）を表す KL ダイバージェンス項の和として次のように与えられる：

$$\mathbb{E}_{\mathbb{Q}_{U,\Sigma}}[S(V)] + \lambda \mathbb{D}_{KL}(\mathbb{Q}_{U,\Sigma}||\mathbb{P}) \tag{1}$$

ここで $\mathbb{Q}_{U,\Sigma}$ はエージェントが実際に採択する行動の分布，$\mathbb{P}$ は理想的または想定するべき行動分布である．したがって第 1 項は行動に伴う物理的コストであることから強化学習における報酬と同様の役割を持つもの，第 2 項の KL ダイバージェンスは「理想とする行動分布」と「現実の行動分布」のずれを測る指標であり，このずれが小さいほど行動は一貫していて，他者からも予測しやすいと解釈できる．したがって，本研究ではこの KL 項を自己の行動の予測可能性を表す内部指標として用いる．

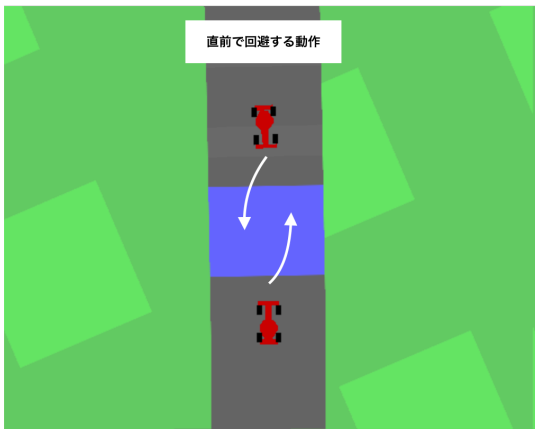
一般的には，理想分布 $\mathbb{P}$ は平均値がゼロの正規分布として定義され，行動の大きさそのものが行動コストに反映される．しかし，この設定だと行動の時間的な変化量を測る指標としては不十分であり，時間的な一貫性を評価しているとはいいがたい．そこで本研究では，「前の自分と似た行動をとり続けるほど予測しやすい」という直感に対応させるため，理想分布を以下のように時間的な滑らかさを考慮して再定義したものを用いる：

$$\tilde{p}(V) = p(V|\alpha\hat{U}, \Sigma) \tag{2}$$

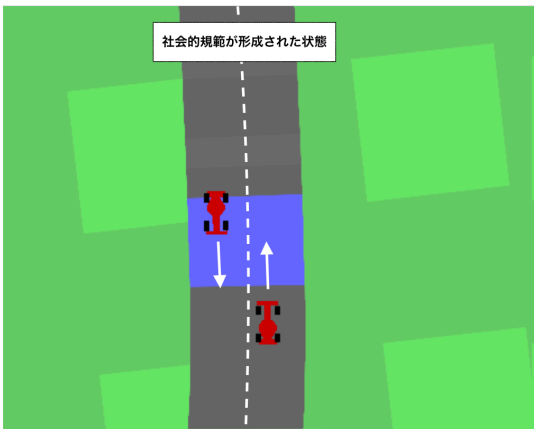
ここで $\hat{U}$ は 1 ステップ前に実際に選択された行動，α はその影響度を調整するパラメータである．これにより，エージェントは「できるだけ動きを大きく変えない」という制約のもとで行動を更新することになる．

以上の定義により自由エネルギーの最小化による行動選択には以下の 2 つの行動原理を表すことになる：

- できるだけ不要な行動をしない（報酬最大化）
- できるだけ前の行動を保つ（予測可能性の維持）



(a) Control 条件で予測される回避動作：互いに直前に急激なステアリング操作を施す.



(b) Predictability 条件で予測される回避動作：互いに事前に対応することで衝突を避ける.

図 2: 回避パターン

3.2 既存モデルへの実装

上記の自身の行動に関する予測可能性は，通常の報酬最大化を目的とする強化学習モデルでは考慮されない．本研究では，この要素を既存のモデルベース強化学習アルゴリズム DreamerV3 に組み込み，行動の予測可能性を考慮する方策学習を実現する [16]．具体的には，DreamerV3 における Actor ネットワークの損失関数へ，次式で定義される「行動コスト」C_{act} を追加する：

$$C_{act} = D_{KL}(p(\mu_{\phi_t}, \sigma_{\phi_t}) || p(\alpha\mu_{\phi_{t-1}}, \sigma_{\phi_{t-1}})) \tag{3}$$

ここで，μ_{ϕ_t} と σ_{ϕ_t} は時刻 t における Actor の出力分布の平均と分散，$\mu_{\phi_{t-1}}$ と $\sigma_{\phi_{t-1}}$ は直前時刻の分布を表す．両者ともガウス分布であるため，KL ダイバージェンスは次式で解析的に計算できる：

$$D_{KL}(\mathcal{N}(\mu_1, \sigma_1) || \mathcal{N}(\mu_2, \sigma_2)) \tag{4}$$

$$= \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{5}$$

この結果，Actor の損失関数 L_{act} は以下のように拡張される：

$$L_{act} = L_{act}^{DreamerV3} + \lambda C_{act} \tag{6}$$

ここで λ は，報酬最大化と行動の予測可能性のトレードオフを調整する重みである．

この修正により，エージェントは「高報酬な行動」を選ぶだけでなく,「過剰に変化しない・他者に予測されやすい行動」を同時に選好するようになる．

4 実験結果

本実験では，各条件において最大学習ステップ数を 11,000,000 ステップとし，学習の進行に伴う各評価指標の時間変化を測定した．最大学習ステップ数は，事前に行った網羅的な探索により，累積報酬が十分に収束することを確認した上で設定した．また Predictability 条件では，学習時の損失関数に行動に関する予測可能性を導入し，その重み係数は 0.001 に固定した．重み係数が大きすぎる場合には，エージェントは報酬の獲得をほとんど無視して常に回転し続けるといった行動が観察され，逆に小さすぎる場合には，通常の強化学習モデルとの差がほとんど生じなかった．そのため，本実験では，予測可能性の追加が報酬最大化を過度に侵害しない範囲で上記の値を選定した．

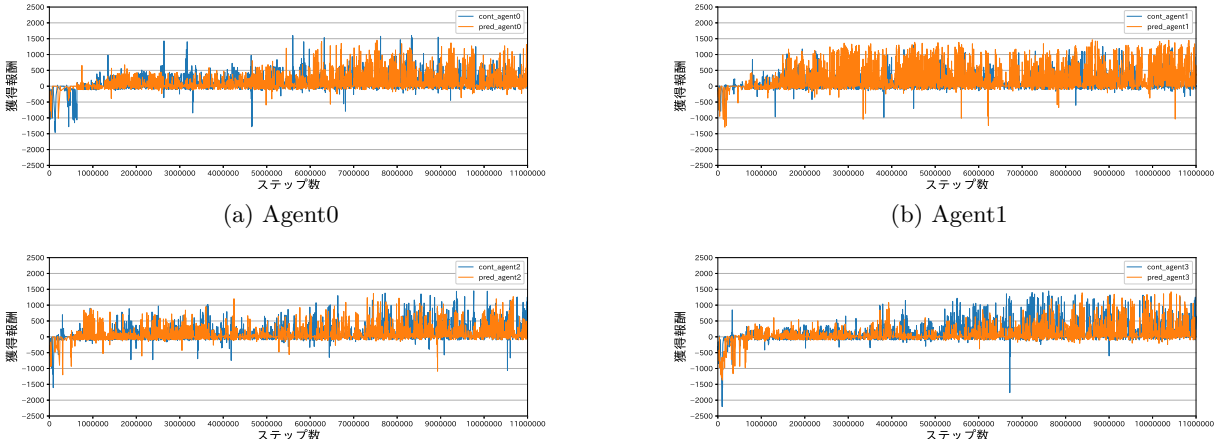
4.1 獲得報酬

各条件におけるエージェント毎の獲得報酬の推移を図 3 に示す．すべてのエージェントにおいて，学習初期には負の報酬が頻発するものの，学習の進行に伴い正の報酬を安定して獲得するようになる傾向が確認された．また，Control 条件と Predictability 条件の間で，学習曲線の収束水準や収束速度に大きな差は見られなかった．

さらに，各エージェントが条件毎に達成したエピソード報酬の最大値を図 4 に示す．いずれのエージェント・条件においても最大値は近い値に収束しており，条件間で顕著な性能差は観察されなかった．このことから，予測可能性の項を導入しても，適切な重みの下では報酬獲得能力が大きく損なわれることはないといえる．

4.2 走行タイル端での走行率

走行タイル端での走行位置の推移を図 5 に示す．各行は Agent0 から Agent3 を，上段から順に表し，左列



(a) Agent0　(b) Agent1

(c) Agent2　(d) Agent3

図 3: 各エージェントの獲得報酬．各図はそれぞれ異なるエージェントに対応しており，青線は Predictability 条件，橙線は Control 条件における獲得報酬を示す．横軸は実行ステップ数，縦軸は獲得報酬を表す．

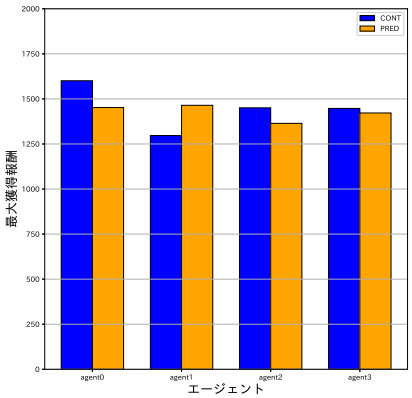


図 4: 獲得報酬の最大値．それぞれのグラフは左から agent0, agent1, agent2, agent3 を表す．

が Predictability 条件，右列が Control 条件に対応する．Predictability 条件では，およそ 1,000,000 ステップ以降から，各エージェントの走行位置が路面の一方（左端または右端）に継続的に偏る傾向が見られた．その後も，偏った側での走行を大きく乱すことなく維持しており，同じ傾向が全エージェントで共通して観察された．

一方，Control 条件では，特に Agent0 と Agent1 において，学習全体を通じて走行位置のばらつきが大きく，特定の端側への安定した偏りは確認されなかった．これに対して Agent2 と Agent3 では，Predictability 条件と同様に，左右いずれか一方への偏りが部分的に見られたものの，全体としては条件間で一貫した傾向は得られなかった．

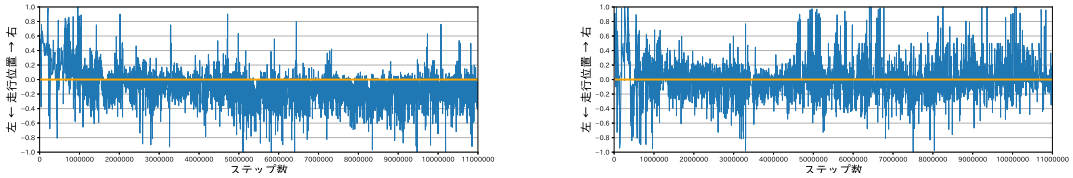
これらの傾向は，走行位置の平均値を集約した図 6 からも観察できる．Predictability 条件では，同じ進行方向を持つエージェント同士（Agent0 と Agent2，Agent1 と Agent3）が同じ側の路面端に偏る一方で，Control 条件ではエージェント毎に偏りの向きが異なり，共通の走行側が形成されにくいことがわかる．これは，予測可能性を考慮した条件において，進行方向毎に通行側が自発的に共有される「規範的な」走行パターンが形成されたことを示唆している．

5　考察

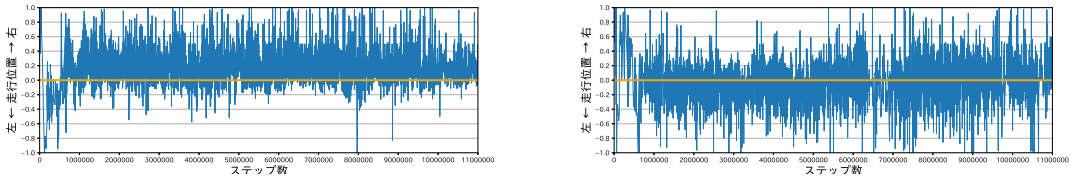
5.1　仮説の検証

実験結果から，自己の行動に関する予測可能性を考慮したエージェント群では，進行方向毎に路面の一方へと走行位置が固定されるパターンが安定して観察された．これは，従来の報酬最大化のみを目的とする強化学習モデルでは明確に見られなかった行動傾向である．このことから，本実験環境において，「行動の予測可能性を導入することにより，社会的規範の内面化が促進される」という仮説が一定程度支持されたといえる．

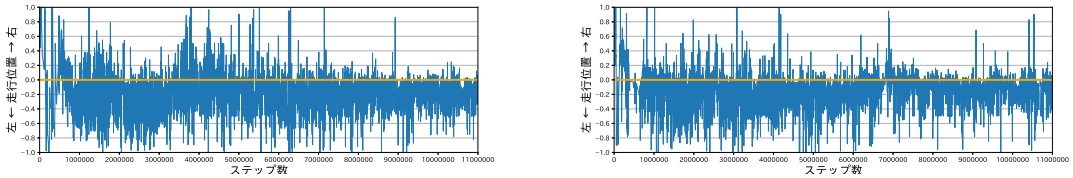
重要なのは，本研究で導入した予測可能性の項が，他者の内部状態や意思決定を明示的に推定するものではなく，あくまで自分自身の行動分布の一貫性を評価する内部指標である点である．それにもかかわらず，集団レベルで固定化された通行パターンが形成されたことは，同質なエージェントから成るマルチエージェント環境において，規範的行動の発生に必ずしも他者モデルや外的インセンティブが必要ではない可能性を示唆している．むしろ，内部報酬として解釈可能な「予



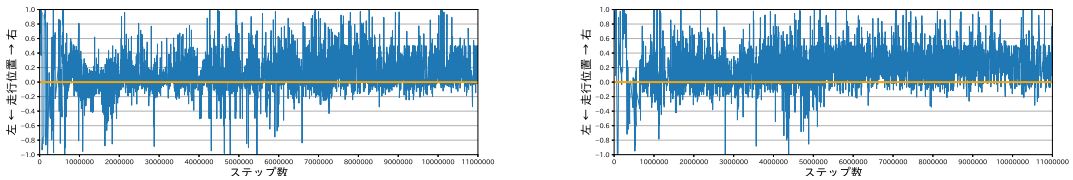
(a) Agent0 (左: Predictability 条件, 右: Control 条件)



(b) Agent1 (左: Predictability 条件, 右: Control 条件)



(c) Agent2 (左: Predictability 条件, 右: Control 条件)



(d) Agent3 (左: Predictability 条件, 右: Control 条件)

図 5: 走行タイル端での走行位置の推移．(a)–(d) はそれぞれ Agent0–3 に対応する．各行において左列は予測可能性を考慮した条件（Predictability 条件），右列は行動コストを導入しない条件（Control 条件）を示す．横軸は実行ステップ数，縦軸は走行タイル上の横方向位置を表し，上方向が右端，下方向が左端に対応する．

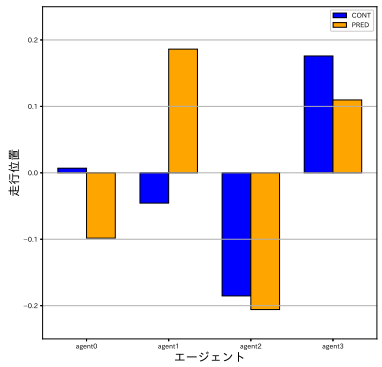


図 6: 走行位置の偏り．各グラフは左から agent0 から agent3 を表し，各色はそれぞれ青色が Predictability 条件，橙色が Control 条件を表す．縦軸は走行タイル上の横方向位置を表し，上方向が右端，下方向が左端に対応する

測可能性」の考慮だけから，協調的な行動配分が導かれ得ることを示唆している．

本研究の結果は，協調や規範遵守を罰則・報酬といった外的インセンティブによって説明する枠組みに対して，行為者自身の行動選択の構造，すなわち「予測可能であろうとする傾向」が集団レベルの整合的行動を支える可能性を示すものである．この観点は，規範的な協調行動が外的制裁によって一方的に強制されるのではなく，行動計画の内面化を通じて成立するという理解と整合的であり，人工エージェントのみならず，人間の社会的協調行動の計算論的理解に対しても一定の示唆を与える．

5.2 限界と今後の展望

一方で，本研究で導入した予測可能性は,「方策の急激な変化を抑制する正則化項」としても解釈可能であり，端走行の固定化が単に行動の平滑化の副産物であ

る可能性は残る．そのため，本稿で示した規範的パターンの解釈をより厳密にするには，環境条件・エージェント構成・タスク構成の三点から追加検証を行う必要がある．具体的には（1）対向車の頻度や視認可能距離など環境条件の変化,（2）一部のエージェントからのみ予測可能性を取り除いた環境,（3）単独走行タスクにおける観察を通して，本稿で見られた端走行が「衝突回避」という社会的文脈に特異的に現れるのかを検証する必要がある．

本研究は，予測可能性に基づく行動制約のみから，外部インセンティブを追加することなく，暗黙的な社会的規範の形成・内面化が起こり得ることを示した点で，協調的人工エージェント設計に向けた新たな理論的・計算論的な基盤を提供するものである．

参考文献

[1] R. Axelrod, “An evolutionary approach to norms,” *American political science review*, vol. 80, no. 4, pp. 1095–1111, 1986.

[2] E. Fehr and C. F. Camerer, “Social neuroeconomics: the neural circuitry of social preferences,” *Trends in Cognitive Sciences*, vol. 11, no. 10, pp. 419–427, 2007.

[3] Y. Shoham and M. Tennenholtz, “On the emergence of social conventions: modeling, analysis, and simulations,” *Artificial Intelligence*, vol. 94, no. 1, pp. 139–166, 1997.

[4] M. J. Gelfand, S. Gavrilets, and N. Nunn, “Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change,” *Annual Review of Psychology*, vol. 75, pp. 341–378, 2024.

[5] D. Kahneman, J. L. Knetsch, and R. H. Thaler, “Fairness and the assumptions of economics,” *The Journal of Business*, vol. 59, no. 4, pp. S285–S300, 1986.

[6] M. Wenzel, “An analysis of norm processes in tax compliance,” *Journal of Economic Psychology*, vol. 25, no. 2, pp. 213–228, 2004.

[7] M. Wenzel and L. Woodyatt, “The power and pitfalls of social norms,” *Annual Review of Psychology*, vol. 76, pp. 583–606, 2025.

[8] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts, “No country for old members: user lifecycle and linguistic change in online communities,” in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW ’13. Association for Computing Machinery, 2013, pp. 307–318.

[9] E. Dimant, F. Galeotti, and M. C. Villeval, “Motivated information acquisition and social norm formation,” *European Economic Review*, vol. 167, p. 104778, 2024.

[10] A. Dutta, S. Roy, O. P. Kreidl, and L. Bölöni, “Multi-robot information gathering for precision agriculture: Current state, scope, and challenges,” *IEEE Access*, vol. 9, pp. 161 416–161 430, 2021.

[11] J. P. Queralta, J. Taipalmaa, B. Can Pullinen, V. K. Sarker, T. Nguyen Gia, H. Tenhunen, M. Gabbouj, J. Raitoharju, and T. Westerlund, “Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision,” *IEEE Access*, vol. 8, pp. 191 617–191 643, 2020.

[12] W. Wang, L. Wang, C. Zhang, C. Liu, and L. Sun, “Social interactions for autonomous driving: A review and perspectives,” *Foundations and Trends® in Robotics*, vol. 10, no. 3–4, p. 198–377, 2022.

[13] J. Henrich and J. Ensminger, “Theoretical foundations: The coevolution of social norms, intrinsic motivation, markets, and the institutions of complex societies,” in *Experimenting with Social Norms: Fairness and Punishment in Cross-Cultural Perspective*. Russell Sage Foundation, 2014, pp. 19–44.

[14] M. Towers, A. Kwiatkowski, J. Terry *et al.*, “Gymnasium: A standard interface for reinforcement learning environments,” 2025, arXiv:2407.17032 [cs.LG].

[15] G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou, “Information theoretic model predictive control: Theory and applications to autonomous driving,” 2017, arXiv:1707.02342 [cs.RO].

[16] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, “Mastering diverse domains through world models,” 2024, arXiv:2301.04104 [cs.AI].