能動的推論に基づいた規範遵守行動と規範共有過程の検証

Verification of Norm-Compliance Behaviour and Norm-Sharing Process Based on Active Reasoning

鈴木遥花 [1*]　竹内勇剛 [1]
Haruka Suzuki[1]　Yugo Takeuchi[1]

[1] 静岡大学
[1] Shizuoka University

Abstract:　人間は，自らの習慣や人の流れや信号機，交通標識といった環境中のサインの観察によって，集団内の空気を読んだ行動をとることができる．これらの行為は，環境中の他者の意図の推定やその場の空気や規範に沿うことでの利益や不利益を考えることなく自動的に行われているものであり，エージェント同士の相互作用よりもエージェントと環境の相互作用が行われていると考えられる．このような規範遵守行動の理由として，報酬最大化ではなく，予測誤差最小化の観点から能動的推論が着目されている．本研究の目的は，能動的推論を用いたシミュレーションを行うことで，集団内に共有されている規範に従う過程や規範が共有されていく過程を数理的に明らかにすることである．シミュレーション実験では，特定の観測値と行為の相関を示す項を能動的推論の式に導入することで，環境中の特定のサインの認識によってエージェントの行動が変化することが示された．これにより，環境を仲立ちとして，その場の空気を読んだ行動や規範順守行動が可能になると考えられる．

1　はじめに

人間は、環境中から得られる情報やそれまでに学んだ習慣によって，状況に応じた行動をとることができる．例えば日本では東京と大阪ではエスカレーターで立ち止まる位置が逆であるが，そのルールを明確に把握していなくても，人の流れや他人の行動をみることや,「エスカレーターではどちらか片方に寄る」などの経験から獲得した習慣によって，自分もそれに応じた行動をとることができる。

このように，人間はたとえ法律や条例などのような明文化されたルールが存在しなくとも，観察や習慣によってその場に生じている「空気」に従った行動をとることができる．これによって，その場にいる集団内の「一般化された他者」[1] が望む行動をとることが可能になっている．

このような規範遵守行動や規範の発生，伝播の過程を捉える数理的モデルの枠組みとして，予測誤差最小化や能動的推論が提案されている [2][3][4][5].

能動的推論では，強化学習などにおける効用の最大化だけではなく，現在の状態の曖昧さの解消や新規性の追求なども目標に追加されている．これらは，価値関数として定義されるものではなく，能動的推論を行うエージェント内で自発的に発生するものである [2]. 自由エネルギーを社会規範を捉える枠組みとして用いることについて，Hartwig ら（2021）らは，選択の自由が社会的エージェントにとってどのような価値を持つのか，文脈によってなぜ協力を行い，社会ルールに同意するあるいはしないのかを直接的に示すことができるとしている [3].

また，Constant ら (2019) は，能動的推論の枠組とマルコフ決定過程に基づく計算論的構成概念として，義務的な社会規則を示す選択・行動・行動系列 deontic value とそれを誘発する環境中の手がかりとして deontic cue を導入して，社会適合や意志決定の概念を定式化している [5]．この cue は一度学習されると，ポリシーの選択を制約するため，例えば交差点で誰も見ていない状況下でも，赤信号という観測値を得るだけでエージェントは「停止する」という規範に従った行動が可能となる [5]．環境中の手がかりを参照することで，他者の意図推定を省いた即時的な行動をとることができると考えられる．

本研究の目的は，エージェントが社会的な評価の考慮や他者の意図推定を行わずとも，その場の空気や規範に従った行動を取ることを示すために，価値関数によって算出される報酬や罰が与えられない状況下でも，規範を遵守した行動を取ることを能動的推論の枠組みでモデル化し，その妥当性を検証することである．

*連絡先：静岡大学情報学部
〒 432-8011 静岡県浜松市中央区城北 3-5-1
E-mail: suzuki.haruka.22@shizuoka.ac.jp

能動的推論の枠組みを用いることで，規範遵守行動を従来の曖昧性の解消や新規性の追求などと加えて，予測誤差最小化の観点から一つの枠組みで扱うことができると考えられる．また，Constant らが数理的枠組みで提唱した deontic cue と deontic value の理論を導入することで，即時にその場の空気や規範に沿った行動を取ることができるのは，他者の意図を常時推定しているのではなく，環境中の特定の手がかりを観察しているからであることを示すことができる．これにより，集団内の空気を読んだ行動や，規範遵守行動が人と人の間のインタラクションだけではなく，人と環境のインタラクションによって説明できる可能性が生まれる．

本研究で，規範遵守行動やその強化過程についてモデリングを行うことで，規範遵守行動を行う際の人間の内部状態，および環境との相互作用を捉えることが可能になると考えられる．

2 背景・関連研究

2.1 自由エネルギー原理

自由エネルギー原理は Friston ら (2006) が提案した脳の統一理論である．これは知覚が感覚信号からの外界の構造や状態を推論した結果であるという Helmholz(1860) の考え方をベイズ理論の枠組みで評価関数の最小化としてとらえることを示したものであり，これによって脳機能が設計されているとしている [6].

この自由エネルギー原理は知覚機能に関係するダイバージェンス項と，知覚とは関係しないサプライズ項の 2 つの項の和として表され，サプライズ項を最小化することが運動や行為に関する機能に対応するとされている [7].

自由エネルギー原理を導入することで，知覚と運動の両方の目的が，個々の生物に存在する生成モデルと環境のずれを最小化することであるとして，定式化することができる．これを能動的推論という．

表 1: 記号定義

記号名	データ
o	観測値
s	隠れ状態
p	真の確率
q	信念
τ	時刻
π	ポリシー
P	ポリシー事前確率分布
Q	ポリシー事後確率分布
E	ポリシー事前選好確率分布
G	期待自由エネルギー

2.2 期待自由エネルギー

能動的推論の議論では，行為系列であるポリシー (π) を考えることで，複数ステップの行為によって未来の目標を達成するような行動計画をたてる問題を取り扱える．このポリシーは，未来の自由エネルギー (以下，期待自由エネルギー) を最小化するようなものが選択される．

期待自由エネルギーは，新しい情報を求める認識的価値と好ましい観測データを求める実利価値から構成される [8]．認識的価値は，ある観測値 (o) を得たときに，環境の隠れ状態 (s) に対する信念 ($q(s)$) がどれだけ変化したのかを表している．実利的価値は観測値 (o) に対する事前の選好を示しており，この価値が最大化されるようなポリシー（π）が高く評価される．これにより，探索と利用のトレードオフを一つの枠組みで扱うことができる．

2.3 Regime of expectations

Regimes of Expectations は特定の文化集団に特徴的な世界の状態に関する暗黙的に共有された期待の集合である．

他者の期待に沿うためには，他者の意図を推定しなければならないが，これには多くの計算コストがかかる．しかし，多くの場合，エージェントは ROEs にのっとった行動を自動的に選択することが可能である．これは，評価する他者が実際に環境中に存在しなくても同様のことがいえる.Constant らは最適な行動をとるための学習は，環境に依存するとして，deontic cue, deontic value という概念を定義している [5].

deontic cue は環境中に存在するサインであり，例えば信号機が挙げられる．交差点の信号機が赤を示しているときに立ち止まることは、環境中のエージェント自身や他のエージェントが最も選択しやすい行動である．

deontic value は，deontic cue から直接推論できる行動の価値を指し，2.2 節の期待自由エネルギーの認識的価値と実利的価値と合わせて ROEs のアーキテクチャを構成している．

deontic value の数理的な定義は以下の通りである．

$$\ln P(\pi|o_\tau) \propto [\ln P(o_\tau|\pi) + \ln P(\pi)] \tag{1}$$

式（1）は，特定の結果や手がかり (o) が与えられた場合のポリシーの直接的な確率を表している．

deontic cue の数理的な定義は以下の通りである．

$$P(o_i|s_\tau) = \frac{\alpha_i}{\sum_k \alpha_k} \tag{2}$$

式 (2) は任意の状態 (s) における結果（o）の確率であり，環境が関数として学習するディリクレ分布の濃

度パラメータ α に依存する．エージェントの環境に対する行動数が多くなるほどこの cue はより頑健となり，規範が環境に組み込まれやすくなる．このパラメータの変化によって方策選択の偏りが生じることを規範の生成と捉えることができる．

3　採餌行動シミュレーション

3.1　目的

前章の 2.3 節では，他者の意図推定を行わなくとも，環境中に残された特定の観測値を観測することで空気や規範に従った行動をとることを説明するために，ROEs を学習し，行動するための手がかりとして deontic cue という項を期待自由エネルギーの式へ導入した．Constant ら (2019) の中では定式化にとどまっていたが，本研究ではシミュレーション実験による実証実験を行うことを目的とする．ここでは実際に，環境中に存在する deontic cue をエージェントが観測することで，deontic value に従った行動選択を行うことを，既に存在している期待自由エネルギーの認識的価値と実利的価値のトレードオフとともにシミュレーションすることを目標とする．

具体的には，エージェントが探索を行う環境中に deontic cue である赤信号を置き，それを観測したときのみ「赤信号で停止する」という deontic value に従った行動をとり，それ以外のときには探索行動を行うことを検証する．

3.2　実験環境

本実験は，能動的推論を行う python ライブラリである pymdp[9] で定義されている関数とサンプルとしてある Active Inference Demo: Epistemic Chaining[10] のコードを改変し，jupyter lab で実行して行った．

3.3　実験方法

図 1は，Epistemic Chaining における 2 次元のグリッド平面に信号機マス（R1,R2）を追加したものである．Epistemic Chainig では，エージェントは 2 つの cue を手がかりとして，報酬（緑色マス）がどこにあるかを探索する．エージェントが cue1 の場所に移動すると，L1, L2, L3, L4 のどの場所に cue2 があるかが示され，cue2 に移動すると隠された報酬がどこにあるか示される．これらの cue や報酬や罰はそれぞれ観測値 o に当たり，これは事前の選好として，エージェントの内部モデルで設定されている．

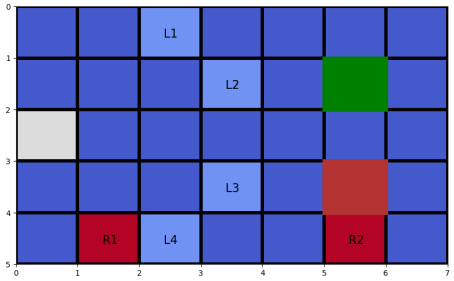

図 1: 信号機マスを追加したグリッド平面

今回は，これらの cue や報酬，罰の観測値 o に，deontic cue を加えることを考える．グリッド平面上の青いマスはエージェントが進むことが可能なマスを指しているが，これに図 1のように，信号機マスを追加することを考える．なお，この信号機マスの位置は，pymdp のデモ [10] でエージェントが通った経路から，エージェントが通る可能性の高いマスに設計している．

3.4　変数定義

以下で実装上の変数の定義を行う．従来の設計に加えて観測値では deontic cue を追加した．また，新たな項として deontic value を加えることで，実際の行動選択にバイアスをかけることを検討した．

3.4.1　deontic value

2.3 節で示した通り，deontic value は特定の観測観測値である deontic cue が観測されたときに，特定のポリシーを引き起こすものである．これは，環境中に存在するエージェントの繰り返し行動によって学習されることから，環境が学ぶ習慣であるといえる．

deontic value を $D(\pi)$，deontic value の重み付けを β とすると，ポリシー事後分布は以下の式のように定式化される，

$$Q(\pi) = \mathrm{softmax}(-\gamma G(\pi) + lnE_{\pi}(o_{\tau}) + \beta * lnD(\pi)) \tag{3}$$

$D(\pi)$ は，E と同じくポリシーに関する確率分布のベクトルであり，pymdp の関数定義部分である control.py[9] で既存の E と同じ機構で設計した．また，ポリシー事後分布の計算を式（3）へ変更して設計した．

本実験では，β を pymdp のデフォルトで定まっている期待自由エネルギーの精度 γ と同じ値にし，「STAY」が選ばれる確率を他のポリシーの 5 倍にして設計を行った．比は以下のように理論上での閾値が出せる．

deontic value($D(\pi)$) で「STAY」が選ばれる確率は，以下の式（4）ように単なる比で表せる．π_{stay} は，次の時刻で「STAY」という行動が選択されるポリシーの数，π_{other} を次の時刻でそれ以外の 4 種類の行動が選択されるときのポリシーの数である．

$$D(\pi) = \frac{\pi_{stay}}{\pi_{stay} + \pi_{other}} \tag{4}$$

エージェントのポリシーの長さは 4 である．また，エージェントの行動数は 3.4.2 より 5 であるのでポリシーの数は合計で $5^4 = 625$ となり，次のステップ選択される行動ごとのポリシーは，$625/5 = 125$ となる．

このとき,「STAY」が選ばれる確率をその他のポリシーの 5 倍とすると，

$$D(\pi) = \frac{5 * 125}{5 * 125 + 1 * (625 - 125)} \approx 0.56 \tag{5}$$

で「STAY」が選択される確率が約 56%となり，deontic value で「STAY」が選ばれる確率にバイアスがかかると閾値だと考えられるため，今回はこのように「STAY」の優先度を設計した．ただし，これは deontic value 内のみのバイアスであり，実際にどのポリシーが選択されるのかは，式 (3) にあるように，期待自由エネルギー（$G(\pi)$）やポリシー事前選好確率分布 (E) からの影響も受ける．

3.5 結果と考察

シミュレーション結果を図 2，図 3に示す．図 2は deontic value が挿入していないエージェント，図 3は deontic value を挿入しているエージェントが，事前選好確率分布（C）のパラメータで設定された報酬マスを目指して 12 ステップで探索行動を行ったときのグリッド上の移動の軌跡と，ポリシーの確率分布を示している．また，グリッド図の下部で，信号機からどのような観測を得たのかを記している．

図 2と図 3の 5 ステップ目，6 ステップ目を比較するとどちらも赤信号（red_light）を観測しているが，deontic value が挿入されている図 3では，式（3）における，deontic value 項の影響が大きくなったため，ポリシー事後確率分布に偏りが生じて「STAY」の確率が高くなったと考えられる．これは，期待自由エネルギーにおける曖昧さを回避するための認識的価値と好ましい結果を求める実利的価値よりも,「何をすべきか」にあたる deontic value が優先されていることを示している．また，赤信号を観測していない時刻では，deontic value が追加されていないエージェントと同じように探索行動をしている．よって，本実験の目標である，エージェントが探索を行う環境中に deontic cue である赤信号を置いたとき，それを観測したときのみ「赤信号で停止する」という deontic value に従った行動をとり，それ以外のときには探索行動を行うことが検証された．

また，図 2においても 11, 12 ステップ目で「STAY」という行動が選択されているが，これは事前選好確率として設定されていた目標の報酬マスにたどり着いた後であるため，それ以上探索行動を行って，期待自由エネルギーにおける実利的価値を求める必要がなくなったためだとと考えられる．

本実験でエージェントが探索を行う環境中に deontic cue である赤信号を置いたとき，それを観測したときのみ「赤信号で停止する」という deontic value に従った行動をとり，それ以外のときには探索行動を行うことが示された．これによって，環境中に存在する規範に従うには，環境中の特定のエージェントの観察や，外部からの報酬や罰がなくても，規範を誘発させるような手がかりである特定の観測値を観測すればよいことが明らかになった．

しかし，式（3）における deontic value($D(\pi)$) やその重みのパラメータである β は設計側で定めたものであるため，トップダウンに設計されたルールとの違いを示すことはできない．

Ramsted ら（2016）は,「文化的アフォーダンス (cultural affordance)」を「自然のアフォーダンス（natural affordances）」「従来型のアフォーダンス (conventional affordances)」の 2 種に区別し，従来型のアフォーダンスは，エージェントが明示的または暗黙的な期待，規範，慣習，協力的な社会実践を活用することに寄与するとしている．赤信号というアフォーダンスが停止という行動の可能性を与えるのは，単に赤信号が停止行動と相関するからではなく，共有された規範，慣習，規則があるからである [11].

本実験での deontic value の設計は，Ramstead らのいう赤信号と停止行動の選択する確率分布の相関にすぎない．また，Constant らの deontic cue の特徴である環境が学習する機構も設計していない．そこで，環境中のエージェントの行動により，環境が特定の観測値と行動の相関を学習し，それがエージェントの行動に反映されることを検証する実験として，次の実験を行った．

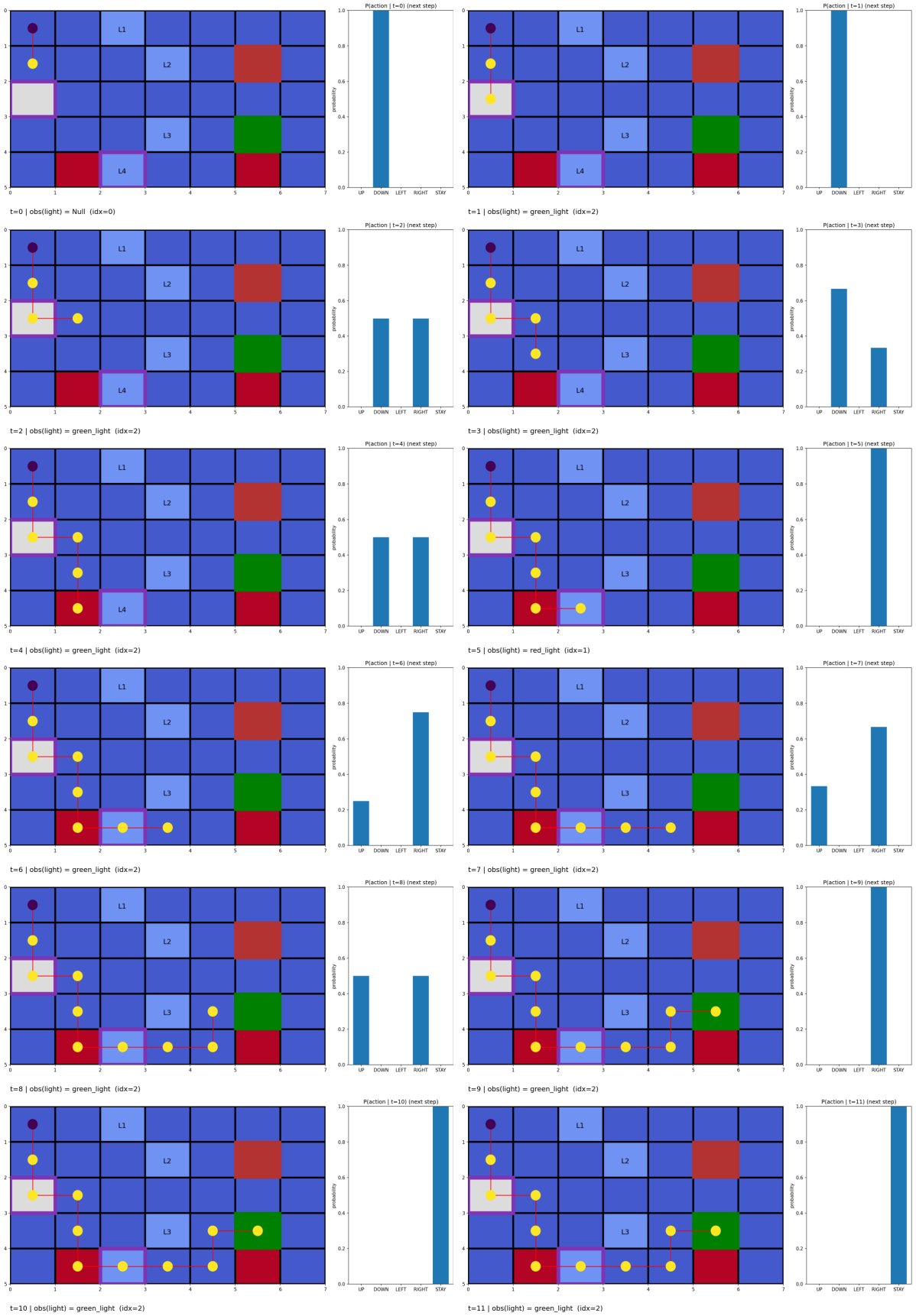

図 2: 実験結果 deontic value なし

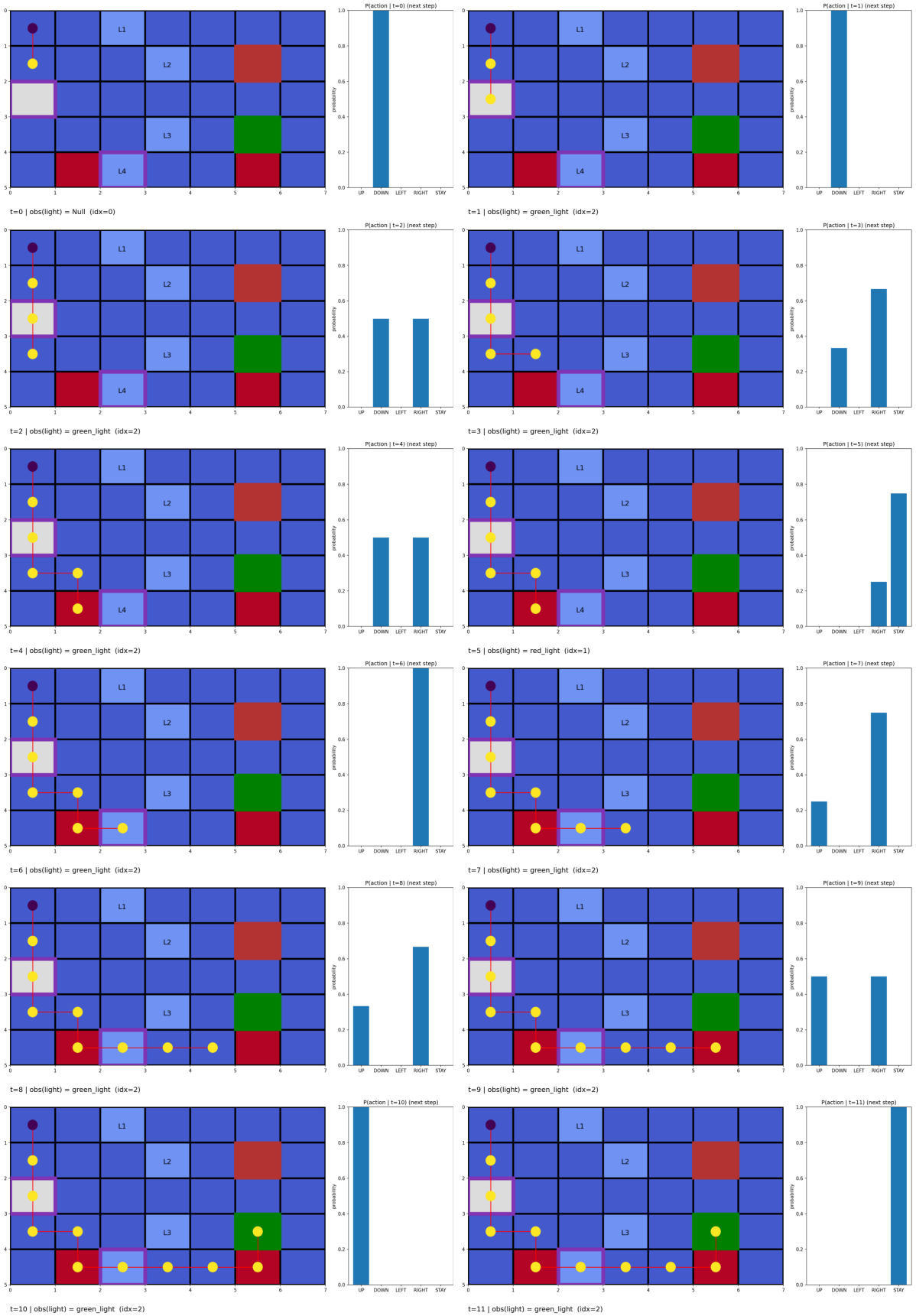

図 3: 実験結果 deontic value あり

4　交通シミュレーション

4.1　目的

3 章の採餌行動シミュレーションでは，deontic value をトップダウンに設計して実験を行った．deontic value によるポリシー選択の偏りは確認できたが，deontic value がエージェントの繰り返しの行動によって生じ，共有される過程を示すことはできていなかった．

そこで本実験では，deontic value が deontic cue として環境中にキャッシュされる過程を示すために，環境側である生成過程のパラメータを設計し，環境が「学習する過程」を検証し，さらにその環境に特定の cue として現れる観測値を観測することで，他のエージェントの振る舞いをみなくても，規範遵守行動が行われることを検証することを目的とする．

また，Constant ら（2019）により，deontic value は孤立した cue ではなく，cue のまとまり（ecology of cues）に依存することが示唆されている [5] ので，本実験では 3 章のような「赤信号」という一つの cue を deontic cue として捉えるのではなく，複数の物理的観測値のまとまりとして定義する．

4.2　実験方法

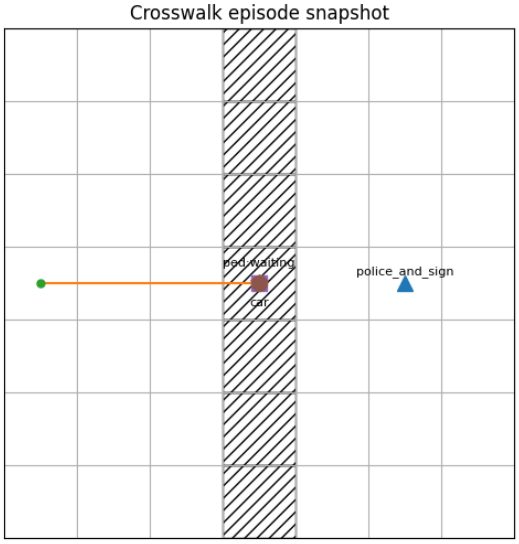

図 4: 交通シミュレーションの概略図
図上では横断歩道列から 2 マスのところに標識と警察官が配置されているが，今回は横断歩道列に入ったら，これら 2 つは観測可能になるものとする．

今回のシミュレーションでは簡略化した横断歩道場面での車の停止行動についてのシミュレーションを行う．3 章の実験結果より，期待自由エネルギーで計算されたポリシー事後確率分布を上回る確率分布を deontic value として挿入することで，deontic value に従った行動を取れることが示された．今回は，deontic cue の発生と共有に着目し，既に規範化されている赤信号で停止する行動よりも判断が揺らぎやすいと考えられる，車エージェントによる横断歩道通過の交通シミュレーションを行う．車エージェントは，横断歩道で歩行者がいるか否か，また横断歩道に標識やペナルティを与えうる警察官などがいるか否かで停止か進むかの判断が揺れると考えられる．

図 4は，実験環境の概略図である．車に乗っているエージェントは横断歩道（図 4の中心部）で横断歩道そのものや，標識，警察官などといった cue 群を観測し，それによって行動が変容するかを検証する．

今回は，環境にキャッシュするエージェント（leader），環境にキャッシュされた情報を得るエージェント（follower）の 2 つに分ける．実環境ではエージェントは複数存在し，それぞれのエージェントが環境にキャッシュしたり，キャッシュされた情報を得てそれに従うことを同時に行っていると考えられるが，今回は簡略化のためこのように定義する．

初期の環境では，cue と停止行動の相関である deontic value は存在しないが，環境中で特定の観測値群（ecology of cues）と leader がとる「STOP」というポリシーの組み合わせをカウントされることで，特定の観測値が停止という deontic value を誘発する deontic cue として刻まれる．このとき leader は警察官による社会罰を嫌がる選好により，環境が観測値を与えたら「STOP」というポリシーを選択しやすくなると設計する．カウントが進んだところでキャッシュが溜まった環境に，follower を入れ，cue を観測したときの振る舞いを見る．

4.3　交通シミュレーションモデル

交通シミュレーションモデルは 3 章のシミュレーションでの用いた python ライブラリである pymdp[9] を用いて設計した．設計方法やパラメータは以下で示す．

4.3.1　行動

本研究では，車エージェントは STOP, GO の 2 つの行動を取れるものとする．車エージェントはグリッド平面の左端から移動して，右端にたどり着きたいという選好を持つものとする．

4.3.2 観測値・norm

ecology of cues($\mathbf{o}$) は deontic value を誘発する手がかりである物理的な観測値の集合であり，これらによって誘発される規範である norm が誘発される．

ecology of cue を式（6）のように定義する．これは，車が観測する物理的な観測値であり，横断歩道，標識，警察の集合である．横断歩道に歩行者がいて，かつこの観測を得たときに norm に従った行動が誘発されることが期待される．

$$\mathbf{o} = \{o^{crosswalk}, o^{sign}, o^{police}\} \tag{6}$$

各物理的観測値は存在しているのか否かでラベリングされている．

norm は以下の式（7）のように定義される．

$$norm = \{None, WeakStop, StrongStop, WeakGo, StrongGo\} \tag{7}$$

今回は，特に横断歩道で停止行動をとることが，deontic value によって可能であることを検証することを目的とする．たとえば，上で定義された ecology of cues を観測し，STOP という停止行動をとったのならば，leader エージェントは，StrongStop という norm をカウントする．

4.3.3 deontic value

前節の norm へのカウントから，その環境下でカウントされた norm が規範である確率 (P_{stop}, P_{go}) は，以下の式 (8)，式 (9) で定義する．α は，カウント数であり，式（8）の $\alpha\_\{WeakStop, StrongStop\}$ は，WeakStop や StrongStop のカウント数，α_norm は norm 全体のカウントの総数を表す．

$$P_{stop} = \alpha\_\{WeakStop, StrongStop\}/\alpha\_norm \tag{8}$$

$$P_{go} = \alpha\_\{WeakGo, StrongGo\}/\alpha\_norm \tag{9}$$

deontic value は，ポリシーの先頭である次状態の行動に P_{stop}, P_{go} の比を取ったものを足して設計している．よって，P_{stop}, P_{go} で確率が高い方が次状態の行動で選ばれやすくなる．

4.3.4 状態遷移確率

状態遷移確率は，エージェントが行動をとることにより，状態が変化することを指す．今回の実験では，歩行者が横断歩道で渡っているときに，車エージェントが GO したら事故が起こり，歩行者が横断歩道にいるのにもかかわらず GO したら警察によるペナルティを受けるように設計した．

歩行者は，以下の確率パラメータでで横断歩道に現れ，横断歩道を渡る．

1. ped_appear (横断歩道に現れる確率)
2. p_cross_after_yield（車が STOP して譲った時に渡る確率）
3. p_finish_cross (歩行者が横断歩道を渡り終える確率)

車エージェントは，歩行者が横断歩道で待っていたり，渡っているときに GO をすると事故や警察によるペナルティを受ける．以下はその確率である．

歩行者が待っているにも関わらず，GO を選んだとき，車エージェントは，以下の確立で Ticket というペナルティを受け，渡っているときに GO を選ぶと以下の確率で事故にあう．

1. p_ticket_if_police
   (警察がいるときに違反になる確率)
2. p_ticket_no_police
   (警察がいないときに違反になる確率)
3. p_crash_if_go_crossing
   (歩行者が渡っているときに GO を選んで事故になる確率)

また，エージェントは STOP したときと GO したとき，以下の確率で遅れが生じるものとする．

1. p_delay_stop
   (STOP したときに遅れる確率)
2. p_delay_go
   (GO したときに遅れる確率)

4.3.5 事前選好確率分布

各エージェントは，キャッシュを刻むきっかけとして採餌行動シミュレーションで定義されていたような事前選好を持つとする．項目は，以下の通りであり，それぞれ w で重みづけされている．

1. goal

2. penalty（Tichket,Crash の 2 値）

3. Time (OnTime, Delayed の 2 値)

期待自由エネルギーの実利的価値により，ゴールや OnTime への選好がエージェントにあり，Delayed を嫌うのであれば，エージェントはなるべく早くゴールへたどり着くために，GO を選び，ペナルティを嫌うのであれば，横断歩道に歩行者がいるときには STOP が選ばれやすくなると考えられる．

4.4 エージェント定義

leader エージェントと follower エージェントを以下の表 2のように定義する．leader エージェントは shared 環境にキャッシュする．また，自身は刻まれた deontic value を参照しない．follower エージェントは，leader エピソードと環境を共有しているもの，deontic value を参照するものの 4 条件にわける．

表 2: エージェント定義

エージェント	環境	deontic value
leader	shared	OFF
follower	fresh	OFF
follower	fresh	ON
follower	shared	OFF
follower	shared	ON

状態遷移の確率を表 3のように定める．

表 3: パラメータ定義（状態遷移）

パラメータ	確率
p_ped_appear	0.50
p_cross_after_yield	0.85
p_finish_cross	0.70
p_ticket_if_police	0.90
p_ticket_no_police	0.05
p_crash_if_go_crossing	1.0
p_delay_stop	0.60
p_delay_go	0.10

また，それぞれの選好の重みも以下の表 4のように定める．好ましい観測は正の値，避けたい観測は負の値で表される．

表 4: パラメータ定義（事前選好）

パラメータ	確率
w_goal	8.0
w_ticket	-6.0
w_crash	-20.0
w_ontime	0.0
w_delayed	-10.0

ただし，follower は社会罰や自分の時間の遅れを考慮しないで，deontic value に従った行動をとることを示したいので，follower エージェントは，ゴールのみ (w_goal) の選好を持つとする．

4.5 結果と考察

図 5は，leader エージェントが刻んだキャッシュがあるかどうか，また deontic value としてそれらを参照するかどうかを各 4 件に分けて行ったときの横断歩道に歩行者がいるときに STOP を取った確率を，図 6は deontic value が参照された瞬間の STOP が取られた確率を示している．

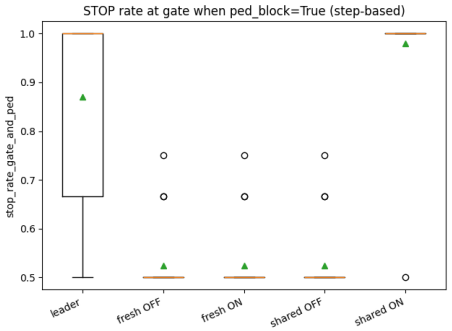

図 5: leader と follower の 4 条件の stop 率比較結果

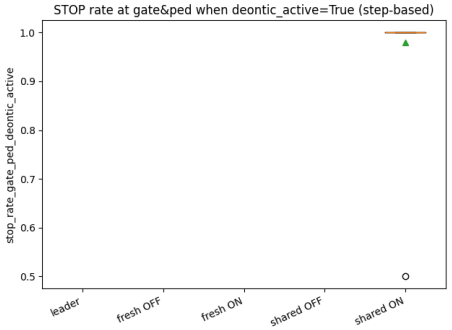

図 6: deontic value ありの leader と follower の 4 条件の stop 率比較結果

図 5からは，leader エージェントと leader エージェントと環境を共有し，かつ deontic value を参照できるエージェントの STOP 行動選択率が高いことがわかる．leader エージェントの停止確率が高いのは，警察からのペナルティや事故を避けたいというエージェントの事前選好が関係していると考えらえる．また，図 6で図 5と同じ follower エージェントの停止確率が高いことから，停止行動を誘発したのは，deontic value であると考えられる．

これにより，最初は leader エージェントの選好から選ばれていた停止行動が，その場に存在する cue と結びついて deontic value となり，follower エージェントは，leader エージェントの行動を観測としてとらえ，推論を行わずとも，環境中に存在している cue を認識することで，leader エージェントと同じ行動をとることが示された．

これにより，環境中のエージェントにより，環境に特定の行動を誘発するような cue が生成されていれば，他のエージェントはその環境に入ってその cue を観察するだけで，特定のエージェントの振る舞いを見ることなしに同じような規範遵守行動がとれることが示された．

5 議論

3 章の採餌行動のシミュレーションでエージェントが探索を行う環境中に deontic cue である赤信号を置いたとき，それを観測したときのみ「赤信号で停止する」という deontic value に従った行動をとり，それ以外のときには探索行動を行うことが示された．しかし，式（3）における deontic value($D(\pi)$) やその重みのパラメータである β は設計側で定めたものであるため，トップダウンに設計されたルールとの違いを示すことはできなかった．

そこで，4 章で deontic value が立ち上がり，共有される過程を検証するために，交通シミュレーションを行った．ここでは，停止行動をとったときに，特定の cue 群を観測することにより，停止行動を誘発させる norm へのキャッシュが溜まることで，cue 群が deontic cue となり，norm が deontic value となり，立ち上がりが検証できた．また，その後 follower エージェントが高い確率で停止行動をとることによって，この deontic value が共有されることが示された．

これによって，規範遵守行動は，特定のエージェントの振る舞いの観察や他者の意図推定を行わずとも，環境中に deontic value としてキャッシュされた特定の観測値である deontic cue を観測することで可能になることが明らかになった．よって，エージェントは他者に直接影響を与えたり，受けたりするだけではなく，環境を仲立ちとして他のエージェントの期待に沿った行動を取ることが可能になると考えられる．

しかし，3 章で行ったシミュレーションのように，ポリシー事後確率分布の計算内での deontic value の寄与を見ていないため，どの程度までポリシーの選択に偏りを生じさせることができるのかは不明である．

また，ここで定義した cue 群は，あらかじめ環境上に存在しているもので，leader エージェントの行動の結果によって生じたものではない．leader エージェントは STOP 行動をとったときにこの cue 群にキャッシュをし，また follower エージェントはそれを直接確率分布として参照している．よってエージェントによってよりキャッシュが刻まれ，deontic value が生成されていく過程を示すには，獣道のような他のエージェントが取った行動の痕跡が環境に刻まれ，それを観測して行動するようなシミュレーションを設計する必要があると考えられる．

参考文献

[1] Mead, H. G. (1934). Mind, Self, and Society. Chicago, IL: University of Chicago Press.

[2] トーマス・パー，ジョバンニ・ペッツーロ，カール・フリストン [著] ; 乾敏郎訳． (2022). 能動的推論 : 心，脳，行動の自由エネルギー原理．ミネルヴァ書房

[3] Hartwig M and Peters A (2021) Cooperation and Social Rules Emerging From the Principle of Surprise Minimization. Front. Psychol. 11:606174. doi: 10.3389/fpsyg.2020.606174

[4] Jordan E. Theriault, Liane Young, Lisa Feldman Barrett, The sense of should: A biologically-based framework for modeling social pressure, Physics of Life Reviews, Volume 36, 2021, Pages 100-136, ISSN 1571-0645, https://doi.org/10.1016/j.plrev.2020.01.004.

[5] Constant, A., Ramstead, M. J. D., Veissière, S. P. L., & Friston, K. (2019). Regimes of expectations: An active inference model of social conformity and human decision making. Frontiers in Psychology, 10, 679. https://doi.org/10.3389/fpsyg.2019.00679

[6] Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. Journal of Physiology-Paris, 100(1–3), 70–87. https://doi.org/10.1016/j.jphysparis.2006.10.001

[7] Friston, K. (2009). The free-energy principle: A rough guide to the brain? Trends in Cognitive Sciences, 13(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

[8] Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. Cognitive neuroscience, 6(4), 187–214. https://doi.org/10.1080/17588928.2015.1020053

[9] Conor Heins, Beren Millidge, Daphne Demekas, Brennan Klein, Karl Friston, Iain Couzin, Alexander Tschantz. (2022). pymdp: A Python library for active inference in discrete state spaces. https://doi.org/10.48550/arXiv.2201.03904

[10] Active Inference Demo: Epistemic Chaining — pymdp 0.0.7.1 documentation. https://pymdp-rtd.readthedocs.io/en/latest/notebooks/cue_chaining_demo.html

[11] Ramstead, M. J. D., Veissière, S. P. L., and Kirmayer, L. J. (2016). Cultural affordances: scaffolding local worlds through shared intentionality and regimes of attention. Front. Psychol. 7:1090. https://doi.org/10.3389/fpsyg.2016.01090