立脚点の違いによって相互予測問題を解消する強化学習エージェント
Reinforcement Learning Agents Solving the Mutual Prediction Problem through Different Standpoints

高田 亮介 †，竹内 勇剛 †
Ryosuke Takata & Yugo Takeuchi
† 静岡大学
Shizuoka University
takata.ryosuke.18@shizuoka.ac.jp

概要

人は他者の意図を推定することで円滑に協調できる一方で，自己と他者の意図推定過程が同じである場合は円滑に協調できない “相互予測問題” に陥る．意図推定には再帰のレベルがあり，相互予測問題を解決するためにはこのレベル差を 1 にする必要があることが知られている．これまで意図推定レベルごとに異なるモデル化を行う研究が行われてきたが，意図推定レベルを切り替えるためにモデル化手法を変えなければならないという問題があった．本研究では，全てのエージェントが同一のモデル化手法のもとで円滑な協調を実現することを目的に，エージェントの立脚点が意図推定レベルに相当するという仮説のもとで相互予測問題を解消することを，強化学習を用いたシミュレーション実験によって確認した．本研究の成果は，相互予測問題を解消する意図推定モデルの実現と，それを獲得するプロセスの解明に寄与し得る．

キーワード：立脚点，意図推定レベル，相互予測問題，協調，強化学習

1. はじめに

人は他者と社会的関係を築くうえで，他者の行動を観測し，その行動の背後に隠された意図を推定する．さらに，自分の行動が他者にとってどのような意図として捉えられるのか，といった影響を考えることができる．例えば人混みの中を移動するときには，向かってくる相手と衝突しないように相手の行動をよく観察して相手の移動経路を予測し，その予測結果に合わせて自分の移動経路を変更することがある．このとき，相互に相手の移動経路を推定しながら歩くにもかかわらず，自分が相手との衝突を回避しようと右に移動すると同時に相手も同様の理由で右に移動し，それならばと左に移動すると相手も同時に左に移動する，といった意図推定の衝突が生じることがある．

Premack et al. (1978) は，自己と他者の目的・意図・知識・信念・疑念・推測・嗜好・ふりなどを推定したり理解できる能力を “心の理論（Theory of Mind）” と呼んだ [1]．心の理論における他者の心の状態の推定には，他者が推定した自己の心の状態の推定，といった再帰性があることが知られている [2]．横山ら (2009) は，心の理論に基づく他者の心の状態の推定の再帰性について，特に他者の意図の推定についての再帰の深さを意図推定レベルと呼んだ [3]．自己と他者が相互に相手の意図を予測する相互予測の状況では，自己と他者が同じ意図推定過程に基づく行動決定モデルを有していると円滑な協調が行えないという問題が生じることが知られており [4]，長田ら (2010) はこの問題を意図推定レベルを用いた説明に拡張した [5]．すなわち，自己と他者の行動決定モデルが同じである場合，意図推定レベルが同一だと自己と他者の予測が重複してしまい，結果的に自己と他者の行動が衝突してしまう “相互予測問題” に陥る．相互予測問題を解決するためには，自己と他者の意図推定レベルに 1 レベルの差を持たせる必要があることが明らかになっている（図 1）．以上の知見は，数理モデルによる 2 次元グリッド環境で表された協調課題のシミュレーション実験 [5] だけでなく，進化計算による意図推定レベルの最適化 [6] や，単純な強化学習手法である Q 学習による協調課題のシミュレーション実験 [7]，子どもの遊び行動の中での他者に対する意図推定のモデル化 [8] によって確認されている．

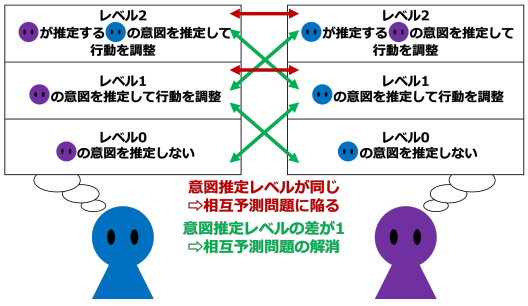


図 1: 意図推定レベルと相互予測問題

これまで行われてきた意図推定レベルの実装によって相互予測問題を解消することを目的とする研究は，複数エージェントが各自の視界を持っており，意図推定レベルごとに内部処理が異なるモデル化を行ったエージェントによる実験であった．そのため，意図推定レベルを切り替えるためにエージェント内部の意思決定モデルのアルゴリズムを変えなければならないという問題があった．この問題によって，複雑な課題を解くことができる深層強化学習を使うことがこれまで不可能であった．なぜなら，深層強化学習はEnd-to-End プロセスであり，モデル内部処理に言及することができないからである．この問題を解決するためには，モデルへの入力段階で意図推定レベルと同質の作用を考える必要がある．

本研究では，“他者の意図を推定する” ということを “他者の立場に視点（立脚点）を置く” ことと同質であると考える．立脚点の変更が意図推定レベルの変更と同質であると仮定すれば，行動モデルの入力の変更によって意図推定レベルを変更することができるため，深層強化学習によって意図推定レベルを表現することが可能となる．深層強化学習のような End-to-End 深層学習は，設計者のドメイン知識を用いることなく，人による設計より優れたモデルの入出力関係を獲得できる [9]．本研究では，深層強化学習を用いることで設計者がエージェント内部の意思決定モデル処理を設計することなく円滑な協調を実現することを目的とする．

本稿では，実験課題として重い箱を複数エージェントで協調的に持ち上げる “箱持ち上げ課題” を用いる．この課題は，2 体のエージェントがタイミングや距離・角度を合わせて行動することで箱が持ち上がるが，エージェント間のタイミングや距離・角度のバランスが合わずに行動すると箱がバランスを崩して落下する．すなわち，2 体のエージェントが時間的・空間的に同期して振る舞う必要があるため，箱持ち上げ課題は相互予測問題に陥る課題であると言える．本研究では，立脚点を変えた条件ごとに深層強化学習を用いて箱持ち上げ課題を解き，その学習結果と振る舞いの変化について分析を行い，構成論的手法によって議論する．本研究の成果は，設計者が環境モデルを保有する必要なくエージェントが円滑に協調する行動決定モデルの実現と，相互予測問題を解消するモデルが構築されるプロセスの解明に寄与し得る．

2. 協調課題

2.1 意図推定レベルと相互予測問題

人は他者と協調するとき，他者の意図を推定し行動を予測している．Dennett (1987) は，心の理論による他者の心の推定には再帰のレベルが存在することを提唱した [2]．横山ら (2009) は，心の理論に基づく他者の意図推定のレベルを表 1 のように定式化した [3]．なお，表 1 ではレベル 2 までしか記述していないが，意図推定レベルは無限に再帰し得る．

表 1: 行動主体が推定する意図推定レベル

レベル	定義
レベル 0	他者の意図を推定せず自己の行動を決定する行動主体
レベル 1	他者をレベル 0 と想定して意図を推定し予測される他者の行動に対応して自己の行動を決定する行動主体
レベル 2	他者をレベル 1 と想定して他者が推定する自己の意図を推定し，その推定から予測される他者の行動に対応して自己の行動を決定する行動主体

2 体の行動主体が相互に相手を予測するとき，同じ行動決定過程を有していると円滑な協調が行えないことが知られている [4, 5]．この問題は相互予測問題と呼ばれる．相互予測問題は，意図推定レベルの差が 1 であるときに解消することが明らかになっている [6]．

2.2 箱持ち上げ課題

実験には，2 体のエージェントが重い箱を持ち上げる “箱持ち上げ課題” を用いる（図 2）．エージェントは，静止，左右移動，ジャンプを行うことができ，バランス良く箱を持ち上げるためには，2 体のエージェントがジャンプのタイミングを合わせ，エージェント間の適切な距離・角度を保つ必要がある．エージェントがジャンプするタイミングが合わなかったり，2 体のエージェント間の距離・角度が適切でない場合，箱はバランスを崩して落下する．このように，時間的・空間的に同期して振る舞う必要があり，2 体のエージェントが相互に相手の行動予測を行うことで相互予測問題に陥る可能性のある課題である．

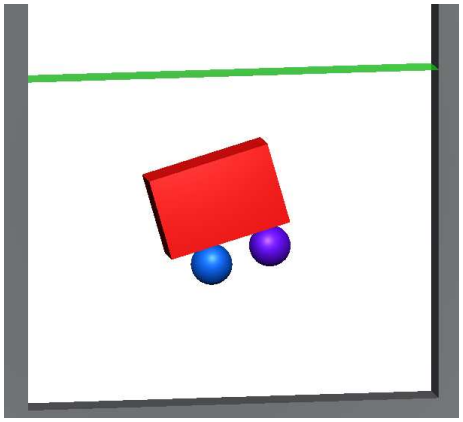


図 2: 箱持ち上げ課題（赤の箱（B），青のエージェント（A1），紫のエージェント（A2），緑の課題達成ラインで構成される）

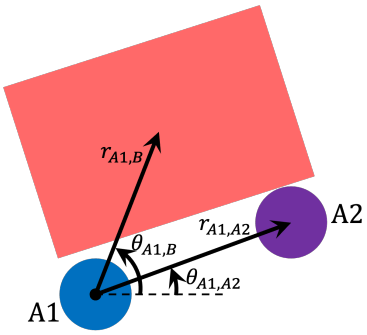


図 3: A1 の立脚点における観測成分

3. 学習実験

3.1 立脚点を変えた実験条件

本研究では，立脚点を変化させることで意図推定レベルが変化するという仮説のもとで，表 2 に示す 4 条件で実験を行った．表 2 において，A1 は左エージェント，A2 は右エージェント，B は箱を表す．2 体のエージェントとも自己の立脚点，他者の立脚点，箱の立脚点，A1 の立脚点，の 4 条件で箱持ち上げ課題を学習させ，相互予測問題を解消するか否かを検証した．

A1 の立脚点から観測する成分を図 3 に示す．A1，A2，B の各立脚点から観測する成分は，立脚点を原点として極座標系を張った際の各対象への距離と角度で表される．

表 2: 実験条件

条件名	立脚点 A1	立脚点 A2	備考
C1	A1	A2	自己の立脚点
C2	A2	A1	他者の立脚点
C3	B	B	箱の立脚点
C4	A1	A1	A1 の立脚点

3.2 学習手法

3.2.1 強化学習

本研究では，エージェントの意思決定モデルをボトムアップに構築する手法として強化学習 [10] を用いた．強化学習は動物の行動決定則の変化をモデル化した手法であり，その仕組みが動物の脳内に存在することを示唆する研究が行われている．Schultz et al. (1993) は，強化学習に用いられる報酬の期待誤差（TD 誤差）が脳神経におけるドーパミン反応と近似していることを明らかにし [11]，Barto (1995) や Schultz et al. (1997) によって，大脳基底核で TD 誤差を用いた強化学習が行われていることが示唆された [12, 13]．さらに Doya (2002) は強化学習における割引率や学習率などのハイパーパラメータがそれぞれセロトニンやアセチルコリンなどの神経修飾物質と対応していることを提唱した [14]．以上のように，脳神経科学の観点から強化学習は人を含む動物の行動学習のモデルとして有効であり，本研究で行うエージェントの意思決定モデルの構築手法として妥当であると考えられる．

3.2.2 PPO

本研究では，設計者がモデル内部のプロセスに言及しない深層強化学習手法である PPO（Proximal Policy Optimization）[15] を用いた．PPO は，環境からの情報取得と目的関数の最適化を交互に繰り返すアルゴリズムであり，ゲーム課題や物理演算シミュレーション等で成果を挙げている [15, 16]．PPO の特徴は，式 (1) を目的関数として勾配法を用いる点である．式 (1) 中の clip 関数によって，方策を更新する際にその変化量が大きくなり過ぎないようにクリッピングされる．clip 関数では，式 (2) に示す方策の変化量比が $1-\epsilon$ より小さい場合，および $1+\epsilon$ より大きい場合に変化量を一定の値にする処理が行われる．なお，式 1 中の $\hat{A}_t$ は時点 t における Advantage（状態に依らない行動自体の価値）の推定値を表している．以上の処理によって，エージェントは自身の方策に対して急激な変化を行わないため，安定した学習が期待される．また，方策勾配は再帰型ニューラルネットワーク（Recurrent Neural Network, RNN）によって近似的に

求められる．これにより，意思決定モデルの中で時間的な行動系列を扱うことができる．

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \tag{1}$$

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \tag{2}$$

3.2.3 学習パラメータ

エージェントの強化学習に用いる状態空間を表 3 に，行動空間を表 4 に示す．表 3 における r，θ は，図 3 に示すようにそれぞれ相対距離と相対角度を表し，状態空間は立脚点によって異なるものとする．なお，エージェントの直径は 1，箱の縦サイズは 2，箱の横サイズは 3，フィールドの横サイズは 10，課題達成ラインまでの高さは 8 である．

表 3: 立脚点と状態空間

立脚点	状態変数	範囲
A1	$r_{A1,A2}$	$[1.0, 11.7]$
	$\theta_{A1,A2}$	$[-\pi, \pi]$
	$r_{A1,B}$	$[1.5, 11.4]$
	$\theta_{A1,B}$	$[-\pi, \pi]$
A2	$r_{A2,A1}$	$[1.0, 11.7]$
	$\theta_{A2,A1}$	$[-\pi, \pi]$
	$r_{A2,B}$	$[1.5, 11.4]$
	$\theta_{A2,B}$	$[-\pi, \pi]$
B	$r_{B,A1}$	$[1.5, 11.4]$
	$\theta_{B,A1}$	$[-\pi, \pi]$
	$r_{B,A2}$	$[1.5, 11.4]$
	$\theta_{B,A2}$	$[-\pi, \pi]$

表 4: 行動空間

行動名	値
ジャンプ	(静止, ジャンプ)
左右移動	(静止, 左移動, 右移動)

2 体のエージェントに毎ステップ与える報酬は式 (3) によって計算される．式 (3) において，$height^t$ は時点 t において箱を持ち上げた場合の箱の高さであり，式 (4) によって表される．また，$FieldHeight$ は地面から課題達成ラインまでの高さを表す．式 (4) において，y_B^t は時点 t における箱の Y 座標を表す．

$$Reward^t = \frac{height^t}{FieldHeight} \tag{3}$$

$$height^t = \begin{cases} 0 & (y_B^{t-1} > y_B^t) \\ y_B^t & (otherwise) \end{cases} \tag{4}$$

PPO におけるハイパーパラメータは表 5 のように設定した．今回の設定は Unity ML-Agents[1] のデフォルト設定を用いた．また，学習ステップ数は 10,000,000（以降，10M と表記）ステップとした．

表 5: PPO のハイパーパラメータ

パラメータ名	値
バッチサイズ	2048
バッファサイズ	20480
方策変化量の閾値 ϵ	0.2
エントロピー正規化率 β	0.005
正規化パラメータ λ	0.95
学習率 η	0.0003
割引率 γ	0.99
エポック数	3
隠れ層のニューロン数	256
隠れ層の数	2

3.3 学習結果と考察

PPO を用いて 10M ステップ学習した結果，エージェントが獲得した報酬の推移を図 4 に示す．なお，図 4 では，各条件 10 試行の平均値を描画している．

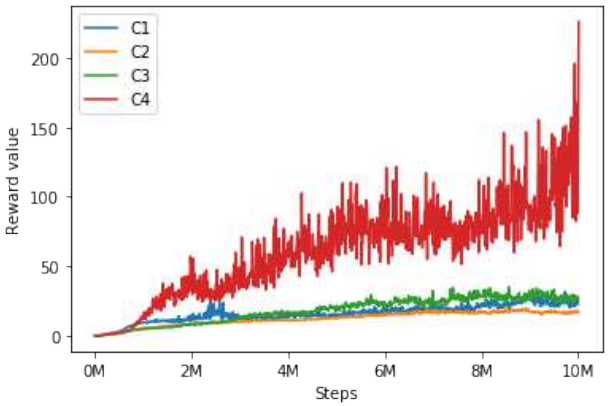


図 4: 学習の結果エージェントが獲得した報酬の推移（各条件 10 試行の平均値）

[1] https://github.com/Unity-Technologies/ml-agents

図 4 より，C4 が他の 3 条件と比べて獲得報酬の水準が高いことがわかる．この結果より，箱持ち上げ課題における相互予測問題を解消し，2 体のエージェントが円滑に協調していることが示唆された．C4 は A1，A2 共に A1 の立脚点から観測し行動を決定する条件である．そのため，A1 にとっては自身の立脚点から観測した相手の状態をもとに行動を決定する “意図推定レベル 1” に相当し，A2 にとっては相手の立脚点から観測した自身の状態をもとに行動を決定する “意図推定レベル 2” に相当すると考えられる．これによって，意図推定レベルに 1 の差が生じ，相互予測問題を解消しているのではないかと推察される．

次に，C4 を除く 3 条件の獲得報酬の推移に注目する．図 5 に 10 試行の平均値の推移を示す．

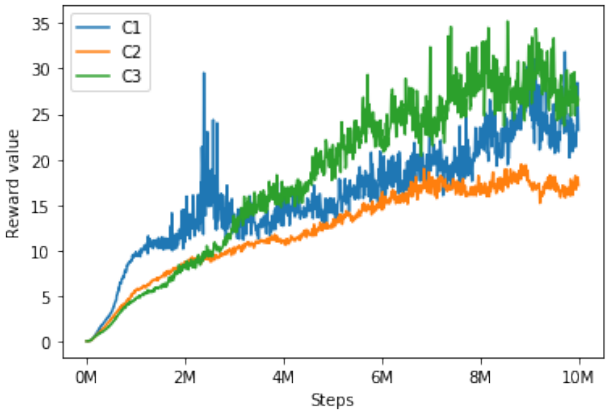


図 5: 学習の結果エージェントが獲得した報酬の推移（C4 を除いた 3 条件，各 10 試行の平均値）

図 5 の C1 に注目すると，2M から 3M ステップにかけて獲得報酬の急激な増加とその後の急激な減少が見てとれる．3M ステップ以降は次第に増加しているが，10M ステップまで学習して最終的に獲得した報酬は，2M から 3M ステップで獲得した報酬と同程度か，それ以下であることがわかる．また，C1 以外の 2 条件では C1 のような急激な増減は見られない．C1 は，2 体のエージェントが共に自身の立脚点から状態を観測し行動を決定する条件であるため，両者とも “意図推定レベル 1” に相当すると考えられる．この結果は，自己の立脚点から状態を観測して行動する場合，学習の初期で円滑に協調する振る舞いを獲得する段階があり，その段階からさらに学習を行うことで過学習による相互予測問題に陥ることを示唆している．

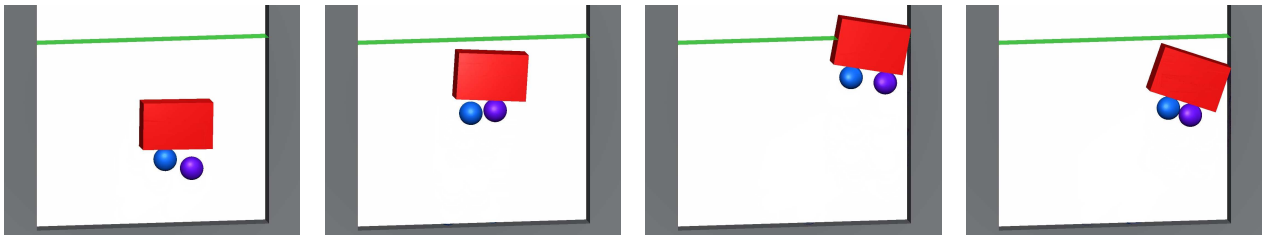
学習の結果獲得したエージェントの行動のうち，特徴的な様子を図 6 に示す．全ての条件で最終的に課題達成ラインまで箱を持ち上げていたが，持ち上げる過程に特徴的な違いが見られた．C1 と C2 はどちらか一方のエージェントが常に箱に接して持ち上げ，もう一方のエージェントは箱が傾いた際に箱を支える行動をとった．このようにリーダーとフォロワーとしての行動が創発した要因としては，各エージェントの立脚点が異なることが考えられる．一方で，C3 と C4 はフィールド右側の壁に箱を寄りかけて持ち上げていた．また，C4 は課題達成ライン付近のフィールド上部で上下させる行動が見られた．C4 に見られた行動は，式 (3) で定義した報酬を最大化するうえで，箱を課題達成ラインまで単調に持ち上げる行動より多く報酬を得られる行動であり，他の条件と比べて高次な戦略を獲得したと言える．

次に，学習の進度に注目する．図 5 より，推移の全体を見ると学習の進度は C2 が最も遅く，最終的に獲得した報酬が最小である．C2 は，2 体のエージェントが共に相手の立脚点から自身の状態を観測し行動を決定する条件であるため，両者とも “意図推定レベル 2” に相当すると考えられる．長田ら (2010) は，意図推定レベル 0 から 2 までのエージェントを設計し，レベルの組み合わせを変えて協調課題のシミュレーションを行っており，意図推定レベル 2 同士では課題達成率は高いが達成までに最もステップ数がかかるという結果を示している [5]．本研究の結果においても，C2 は学習の中で単調に獲得報酬を増加しているが，その進度は他の条件に比べて遅いという特徴が，長田らの結果と合致している．

4. 議論

4.1 立脚点と意図推定レベルの対応

本研究では，相互予測問題が生じることで円滑に協調できない箱持ち上げ課題において，エージェントの立脚点の組み合わせを変化させて学習実験を行った．図 4 に示したように，A1 と A2 の両方が A1 の立脚点から状態を観測して意思決定することで，円滑に協調できることが示された．自身に立脚点を置くことで “自分から観測した相手” の状態をもとに意思決定を行うため，意図推定レベル 1 を表現し，協調する相手に立脚点を置くことで “相手から観測した自分” の状態をもとに意思決定を行うため，意図推定レベル 2 を表現すると考えられる．このように，立脚点をどこに置くかによって異なる意図推定レベルを表現でき，相互予測問題を解消することができることが示唆された．



(a) C1: A1 がリーダー，A2 がフォロワーとなって箱を持ち上げる

(b) C2: A1 がフォロワー，A2 がリーダーとなって箱を持ち上げる

(c) C3: A1 と A2 が適切な距離を保ち，箱を壁に寄りかけて持ち上げる

(d) C4: A1 と A2 が密接し，フィールド上方で箱を壁に寄りかけて上下させる

図 6: 学習の結果獲得したエージェントの行動

4.2 人の学習と強化学習との比較

本研究では強化学習を用いてエージェントの協調行動を獲得した．ここで，協調行動の獲得にかかったステップ数に注目する．今回は実験的に 10M ステップの試行であったが，図 4 と図 5 からわかるように，終盤まで右肩上がりの学習曲線になっている．すなわち 10M ステップまで行動獲得のための学習が進行していることを表している．強化学習と人の学習を比較したとき，人の学習は強化学習のように膨大な時間をかけて処理を実行しているとは考え難い．人の学習と強化学習との学習時間・学習ステップの差は，メタ学習の有無と身体的な常時学習の有無から生じると考えられる．人は “どのように学習すべきか” というメタ戦略を学習することができる一方で，今回行った強化学習は箱持ち上げ課題の環境から得られる状態をもとに一元的な学習しか行っていない．そのため，人が行うような転移学習は今回の強化学習では生じないことから学習に時間がかかると考えられる．また，人は成功/失敗体験のフィードバックに基づく認知的な学習だけでなく，皮膚などの膨大な数のセンサーから常に環境の状態を観測し，常に体感フィードバックを得て身体的な学習を行っている．そのため，人が新しい課題に取り組む際には常時行われている身体的な学習によって獲得されている身体能力を用いることができる．一方で，今回行った強化学習では認知的な学習だけでなく身体的な学習を事前学習の無い状態から行っていると考えられるため，人より学習時間がかかったと考えられる．

4.3 立脚点に基づく観測の主観性・客観性

C3 と C4 は，どちらの条件も A1 と A2 が同じ立脚点から状態を観測しているにもかかわらず，学習結果に差が生じた．この要因として，立脚点から得られる情報の質の違いが影響していると推測する．C3 の観測成分は，図 7 に示すように，箱に立脚点を置いた際の A1 と A2 の状態である．このとき，A1 と A2 は共に自身の外部から自身の状態を観測して行動を決定する．すなわち，A1 と A2 はどちらも客観的に（あるいは間主観的に）自身の状態を観測している．それに対して，C4 の観測成分は図 3 に示した通り，A1 にとっては主観的な観測であり，A2 にとっては客観的な観測である．以上より，C4 は相互予測問題を解消し，C3 は相互予測問題によって獲得報酬が C4 ほど増加しなかったと考えると，相互予測問題の解消には立脚点の設置によって主観的な観測と客観的な観測が両方とも実現されることが必要であると予想される．

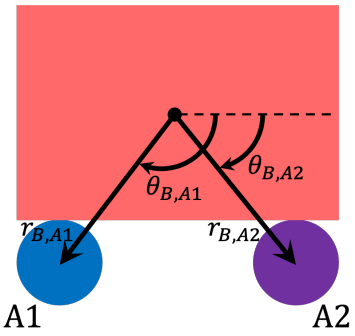


図 7: 箱に立脚点を置いた際の観測成分

5. おわりに

本稿では，時間的・空間的に同期が必要であることから相互予測問題に陥る可能性のある箱持ち上げ課題を題材に，強化学習によってボトムアップに意思決定モデルを構築するエージェントの立脚点の組み合わせが相互予測問題を解消することを検証した．学習の結果，立脚点の設置によって主観的な観測を行うエージェントと客観的な観測を行うエージェントを実現することで円滑に協調できることが認められた．

本研究により，設計者によるモデル設計を伴わない，行動主体の立脚点に基づいたボトムアップなモデル化によって，相互予測問題を解消できることが示唆された．この知見を用いることで，これまで設計者が環境に関するドメイン知識を保有していなければモデル化が困難であった複雑な環境において，相互予測問題を解消して協調的に振る舞うエージェントの設計を行うことができる．また，エージェントが他者とのインタラクションを通して協調行動を学習する過程を観察することができるため，相互予測問題を解消し協調的なインタラクションを獲得するプロセスの解明に寄与し得る．

本研究は，エージェント間の身体的及び認知的な能力差が無いという前提のもとで実験を行った．そのため，本研究では個体間の能力差を考慮した協調については言及できない．人が協調して課題を達成するために他者の意図を推定する場合，少なからず自己と他者の身体能力や認知能力を考慮すると考えられる．このような自己と他者の能力差を考慮して協調するエージェントを実現するためには，エージェントごとに身体的，認知的なパラメータを設定し，それらを観測し相対的に処理する仕組みをモデル内に組み込む必要がある．

今後の課題として，まずは学習過程の振る舞いの分析が挙げられる．今回実現した 2 体のエージェントがどのように協調行動を獲得していったのか，そのダイナミクスの分析を通して，相互予測問題を解消する協調行動を実現する方法を構成論的に解明していきたい．また，今回行った 4 条件は全てエージェントの移動によって立脚点が変化する条件であった．特に C3 において，立脚点が置かれた箱の位置は A1 と A2 のどちらの行動によっても変化するため，箱の位置情報にエージェント自身の位置情報を含むという解釈が可能であるということから完全に客観的な観測とは言えない点に注意が必要であった．そのため，今回行ったローカル極座標系だけでなく，移動しない実験環境上の原点に立脚点を置いたグローバル座標系での実験を行い，C3 との比較によって立脚点がエージェントの主観性にどのようにかかわるのか考察したい．

文献

[1] Premack, D., & Woodruff, G. (1978) “Does the chimpanzee have a theory of mind?” Behavioral and brain sciences, Vol. 1, No. 4, pp. 515-526.

[2] Dennett, D. (1987) “The Intentional Stance” MIT Press. (若島正, 河田学 訳 (1996) 『志向姿勢の哲学–人は人の行動を読めるのか？』 白揚社.)

[3] 横山絢美, 大森隆司 (2009) “協調課題における意図推定に基づく行動決定過程のモデル的解析” 電子情報通信学会論文誌 A, Vol. 92, No. 11, pp. 734-742.

[4] Nagata, Y., Ishikawa, S., Omori, T., & Morikawa, K. (2007) “Computational model of cooperative behavior: Adaptive regulation of goals and behavior” In Proc. Second European Cognitive Science Conference, pp. 202-207.

[5] 長田悠吾, 石川悟, 大森隆司, 森川幸治 (2010) “意図推定に基づく行動決定戦略の動的選択による協調行動の計算モデル化” 認知科学, Vol. 17, No. 2, pp. 270-286.

[6] 高野雅典, 加藤正浩, 有田隆也 (2005) “心の理論における再帰のレベルの進化に関する構成論的手法に基づく検討” 認知科学, Vol. 12, No. 3, pp. 221-233.

[7] 椿本樹矢, 小林邦和 (2015) “意図推定法を用いたマルチエージェント強化学習システムにおける協調行動の獲得” 電気学会論文誌 C(電子・情報・システム部門誌), Vol. 135, No. 1, pp. 117-122.

[8] 阿部香澄, 岩崎安希子, 中村友昭, 長井隆行, 横山絢美, 下斗米貴之, 岡田浩之, 大森隆司 (2013) “子どもと遊ぶロボット: 心的状態の推定に基づいた行動決定モデルの適用” 日本ロボット学会誌, Vol. 31, No. 3, pp. 263–274.

[9] Bengio, Y., Courville, A., & Vincent, P. (2013) “Representation learning: A review and new perspectives” IEEE transactions on pattern analysis and machine intelligence, Vol. 35, No. 8, pp. 1798-1828.

[10] Sutton, R. S., & Barto, A. G. (2018) “Reinforcement learning: An introduction” MIT press.

[11] Schultz, W., Apicella, P., & Ljungberg, T. (1993) “Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task” Journal of neuroscience, Vol. 13, No. 3, pp. 900-913.

[12] Barto, A.G. (1994) “Adaptive critics and the basal ganglia” Models of Information Processing in the Basal Ganglia, pp. 215-232.

[13] Schultz, W., Dayan, P., & Montague, P. R. (1997) “A neural substrate of prediction and reward” Science, Vol. 275, No. 5306, pp. 1593-1599.

[14] Doya, K. (2002) “Metalearning and neuromodulation” Neural networks, Vol. 15, No. 4-6, pp. 495-506.

[15] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017) “Proximal policy optimization algorithms” arXiv preprint arXiv:1707.06347.

[16] Bøhn, E., Coates, E. M., Moe, S., & Johansen, T. A. (2019) “Deep reinforcement learning attitude control of fixed-wing uavs using proximal policy optimization” In 2019 International Conference on Unmanned Aircraft Systems (ICUAS), pp. 523-533.