学位論文 博士（工学）

AIエージェントの挙動に関する共通認識の構築に向けた人とのコミュニケーションモデルの研究

2023年3月

慶應義塾大学 大学院 理工学研究科

福地 庸介

要旨

機械学習によって行動を学習する自律エージェント—**AI エージェント**—が、技術の発展にともなって複雑な挙動を扱えるようになっている。しかし、既存の AI エージェントの設計では、エージェントと共存するユーザや周囲の人とのコミュニケーションが十分に考慮されていない。AI エージェントの多くは事前に与えられた価値基準を最大化することに特化しており、それをもとに学習した行動方策は、実際の場面で人が持っている評価基準や期待とは必ずしも一致しない。真に人に利する存在となるために AI エージェントは、学習した方策に固定的に従うのではなく、ユーザが期待する挙動を適宜認識できるようにすることが重要である。また、AI エージェントが人の期待を認識するだけでなく、人も AI エージェントの挙動を正しく把握できるようにする必要がある。エージェントの挙動を人が把握できなければ、エージェントが予測不能となり、意図せぬ動作や思わぬ事故に繋がる。しかし、特に深層強化学習モデルは大量の数値パラメータで表現され、多くの場合で人が直接理解できる表現を持たない。こうした学習モデルブラックボックス化の問題は、モデルの複雑化や大規模化とともに深刻さを増している。

本論文の目標は、人と AI エージェントの共存に向けて、AI エージェントの挙動に関する相互理解を構築する**挙動アライメント**を達成することである。そして、コミュニケーションを通じて挙動アライメントを達成するための 2 つのモデルを提案する。2 つのモデルに共通するのは、AI エージェントと人との間で、目標や信念といった心的状態を帰属し合う〈心〉の読みあいの中で、両者の間に存在する情報の差異（非対称性）を明示的に組み込むことで、効果的なコミュニケーションを実現している点である。

第 1 に提案する「期待されるエージェント」モデルは、人が AI エージェントに達成を期待する目標を推測しながらコミュニケーションを行う過程をモデル化したものである。このモデルには、人の目標は AI エージェントの目標と必ずしも一致しないという目標の非対称性の存在が組み込まれている。目標の非対称性を考慮することで AI エージェントは、人から与えられる指示を正しく解釈することができる。さらに、解釈した指示の語彙を流用することで、AI エージェントがどのように動こうとしているかを人に伝達できるようになる。評価実験の結果、「期待されるエージェント」モデルが、指示の背後にある目標を正しく推定できること、推定した目標を元に解釈した指示の語彙を AI エージェントの動きの伝達に利用することで、人が AI エージェントの挙動を精度よく予測できるようになることが示された。

第 2 に提案する「推測されるエージェント」モデルは、AI エージェントの動きを見た人がエージェントに対して帰属する目標を、エージェントの側から推測しながらコミュニケーションを行う過程をモデル化している。この時、人と AI エージェントが異なる視界から環境を観測しているという観測の非対称性を考慮することで、人が AI エージェントに帰属する目標を正しく推測できる。さらに、このモデルを応用することで、AI エージェントの目標を伝達する動き (表意動作) を生成できる。評価実験の結果、「推測されるエージェント」モデルが生成する表意動作によって、人が AI エージェントの目標をより早く、正しく推測できるようになることがわかった。

ABSTRACT

With the development of machine learning, AI agents, which autonomously learn actions through machine learning, are becoming capable of making complex decisions. However, their lack of ability to communicate with humans makes human-agent coexistence challenging. As a result of poor communication regarding the agents' behavior, they cannot behave as humans expect them to. The behavior of AI agents with machine learning depends on training datasets or values defined by designers, so it does not always match users' expectations. Through communication, AI agents should be able to understand what users expect of them to align their behavior with user expectations. In addition, AI agents should be able to communicate what they will do because their black-box decision-making modules prevent users from understanding their future behavior, which can lead to unintended behaviors and serious accidents.

This thesis presents two studies that aim at *behavior alignment*, i.e., aligning how AI agents behave and how humans expect them to behave through communication. Each study proposes a communication model to enable an AI agent to achieve behavior alignment. The two studies have in common that they exploit "mind"-reading phenomena between humans and AI agents and take into account an information asymmetry between them.

The first study proposes an *expected-agent* model, which enables an AI agent to interpret human instructions to it and convey what the agent will do. By taking into account an asymmetry between the goals of a human and an agent, the agent can correctly interpret human instructions. Moreover, the expected-agent model learns the vocabulary used in human instructions and diverts it to convey what the agent will do. Experimental results show that explanations of an AI agent's future behavior generated by the expected-agent model can reduce errors in human predictions of the agent's future location.

The second study proposes an *inferred-agent* model, which infers an AI agent's mental state attributed by a human observing the agent's motion to enhance behavior alignment communication. The model effectively handles an observation asymmetry between a human and an agent to infer an attributed mental state and generates legible motion, a motion that conveys the agent's goal. Experimental results show that the inferred-agent model can enable human observers to quickly infer an agent's goal by taking into account an observation asymmetry.

目次

1 人と AI エージェントの共存に向けて **1**

1.1 挙動アライメント 1
1.2 目標志向 Explainable AI 2
1.3 心の理論と AI エージェント 3
1.4 挙動アライメント・コミュニケーションモデル 6
1.5 本論文の構成 7

2 技術的背景 **9**

2.1 強化学習 9
2.2 マルチエージェント・システム 10
2.3 言語による挙動アライメント 10
2.3.1 目標志向 XAI における言語 10
2.3.2 指示遂行 11
2.3.3 インタラクティブ強化学習 11
2.4 表意動作 11
2.5 先行研究の限界 12
2.5.1 目標の非対称性 13
2.5.2 観測の非対称性 13

3 「期待されるエージェント」モデル **15**

3.1 モデルの概要と目標と非対称性 15
3.2 状況設定 16
3.2.1 環境 16
3.2.2 指示者 17
3.3 「期待されるエージェント」モデル 17
3.3.1 構成モジュール 17
3.3.2 損失関数 18
3.4 実装 20
3.4.1 学習の手順 20
3.4.2 モデル 20
3.5 実験概要 22

3.6 数値実験 22
3.6.1 目的 22
3.6.2 方法 22
3.6.3 結果 24
3.7 ユーザスタディ 25
3.7.1 目的 25
3.7.2 方法 26
3.7.3 仮説 27
3.7.4 結果 27
3.8 限界と今後の展望 29
3.9 まとめ 30

4 「推測されるエージェント」モデル 31
4.1 モデルの概要と観測の非対称性 31
4.2 客体的自己認識と予告 31
4.3 Bayesian Theory of Mind (BToM) 32
4.4 「推測されるエージェント」モデル 32
4.5 実装 33
4.5.1 環境と AI エージェント 33
4.5.2 「推測されるエージェント」モデル 35
4.6 ケーススタディによる検証 37
4.6.1 目的 37
4.6.2 方法 37
4.6.3 結果と考察 38
4.6.4 ケーススタディのまとめ 42
4.7 表意動作の生成 42
4.7.1 FalseProjective 表意動作 43
4.8 生成された表意動作 44
4.9 表意動作の評価概要 47
4.10 シミュレーションスタディ 47
4.10.1 目的 47
4.10.2 方法 47
4.10.3 仮説 48
4.10.4 結果 48
4.10.5 シミュレーション実験のまとめ 49
4.11 ユーザスタディ 50
4.11.1 目的 50
4.11.2 方法 50
4.11.3 仮説 51

4.11.4 結果 52
4.11.5 ユーザスタディのまとめ 55
4.12 「推測されるエージェント」モデルの限界と今後の展望 56
4.13 まとめ 57

5 研究全体の議論と制約・今後の展望 **58**

6 まとめ **60**

参考文献 **61**

図目次

1.1 人と AI エージェントをつなぐコミュニケーションモデル 3
1.2 心的状態の推定と言動の解釈 4
1.3 心の理論と推測の次数 4
1.4 AI エージェントによる心的状態の推測 5
1.5 「期待されるエージェント」モデルと「推測されるエージェント」モデル 7

2.1 表意動作 12
2.2 観測の非対称性を考慮すべき例 14

3.1 「期待されるエージェント」モデルによる人の目標の推定と指示語彙の解釈 16
3.2 「期待されるエージェント」モデルの構成モジュール 18
3.3 f_N の実装 21
3.4 比のヒストグラム 25
3.5 「期待されるエージェント」モデルと *ablation* によって獲得された f_N の精度 26
3.6 ユーザスタディに用いたシステムのインタフェース 27
3.7 実験参加者の予測の絶対誤差 28

4.1 「推測されるエージェント」モデルのグラフィカルモデル 33
4.2 環境 A 34
4.3 環境 B 35
4.4 実験に用いたユーザインタフェース 38
4.5 simple エピソードにおいて、「推測されるエージェント」モデルと実験参加者が推測した、エージェントの目標がリンゴである確率 39
4.6 blind エピソードの結果 40
4.7 misleading エピソードの結果 1 41
4.8 misleading エピソードの結果 2 41
4.9 Center シナリオで生成された動き 45
4.10 Side-visible シナリオで生成された動き 45
4.11 Side-invisible シナリオで生成された動き 45
4.12 Blind-inside シナリオで生成された動き 46
4.13 Blind-outside シナリオで生成された動き 46

4.14 3 種類の動きが獲得した評価値の遷移 49
4.15 観測の非対称性が存在する場面における FalseProjective と「推測される
エージェント」モデルの評価値 50
4.16 観測の非対称性が存在する場面における FalseProjective と「推測される
エージェント」モデルの評価値 50
4.17 参加者の推測結果 53
4.18 参加者の主観評価 55

表目次

3.1 f_N のハイパーパラメータ 21
3.2 *Encoder* のハイパーパラメータ 21
3.3 A2C の学習に用いたハイパーパラメータ 24

4.1 私的・公的自己認識の数式的表現 32
4.2 各シナリオの設定 44
4.3 *rapidity* 指標に関する仮説 52
4.4 *correctness* 指標に関する仮説 52
4.5 *rapidity* 指標に関する結果 55
4.6 *correctness* 指標に関する結果 55

1 章

人と AI エージェントの共存に向けて

1.1 挙動アライメント

環境の状態を認識し、自律的に行動するコンピュータ (自律エージェント) によって、人々の生活は便利で豊かなものになる。身体を持った自律エージェントであるロボットが人の作業を代替することで、人は労力を他の作業に配分できるようになる。またロボットは、重い荷物の運搬や危険な環境での作業といった、人には身体的に難しい作業をこなすことができる。ロボットと人にはそれぞれ強みと弱みがあり、互いが強みを生かして欠点を補い合うことで、単体では成しえない仕事を達成することも可能になる。

自律エージェントは、機械学習技術の導入によって、人手で設計できる限界を超えた複雑な挙動を生み出せるようになってきている。機械学習を組み込んだ自律エージェントのことを、本論文では **AI エージェント**と呼ぶ。機械学習の一手法である深層学習の発展によって、カメラやセンサから取得される実世界の情報をコンピュータが直接処理する能力が向上した。そして深層学習を、報酬をもとに試行錯誤の中で行動を学習する強化学習と組み合わせることで、AI エージェントが様々なタスクに対して高い性能を発揮するようになっている [40, 58]。機械学習技術の発展に伴って、AI エージェントの活躍の場は今後ますます広がっていくと期待される。

しかし、機械学習の活用によって新たに発生する様々な問題も危惧されている。例えば、AI エージェントの自律的な行動が引き起こした失敗の責任を誰がどのように負うのか、トロッコ問題に代表される葛藤場面において AI エージェントはどのように振る舞うべきか、機械学習によって学習された挙動と人の常識・価値観との差異、といった倫理面の問題が議論されている [69, 59]。工学の観点からも、学習された結果と設計者やユーザの期待との適合性を如何に検証するか、学習結果の信頼性を如何に担保するか等、残された問題は多い [52]。

本論文にまとめる研究の目標は、AI エージェントと人とがコミュニケーションを通じて、AI エージェントの挙動についての共通認識を構築する**挙動アライメント**を実現することである。AI エージェントの行動は外部に対して不可逆的な変化をもたらす。特に身体を持ったロボットの場合、予測不能な行動は思わぬ事故を招く。AI エージェントと人とが同じ環境で共存していくためには、AI エージェントの取ろうとしている挙動を人が

正しく把握し、制御できるようにする必要がある。

人がAIエージェントを理解する方向とは逆に、AIエージェントの側が人の望む・望まない行動を認識し、それに合わせて行動を選択できるようにするという点でも挙動アライメントは重要である。機械学習によって獲得される挙動は学習のデータセットや設計者が与えた価値基準に依存しており、実際に人が期待する挙動に合致しているとは限らないためである。さらに、目標の共有や互いの役割に関する共通認識の構築は、AIエージェントと人とが力を合わせて一つの仕事に取り組む協調を実現するためにも不可欠である [24]。

1.2 目標志向 Explainable AI

機械学習モデルのブラックボックス化 [53] の問題は、AIエージェントが周囲の人とコミュニケーションを取れるようにする必要性を高めている。深層学習モデルは複雑な挙動を可能にしている一方で、モデルが大量の数値パラメータで表現され、学習した内容や行動を選択する過程をパラメータから直接読み取ることが困難であることが多い [25]。機械学習モデルのブラックボックス化により、AIエージェントがどういった挙動を学習したのかを人が理解することは難しくなる。特に問題が深刻になるのは、AIエージェントの設計に関する知識の乏しい一般のユーザが相手になった場合である。人とAIエージェントの共存に向けて、AIエージェントの挙動を一般のユーザでも理解できるように説明することは重要である。

ブラックボックス化する機械学習モデルに対し、判断の説明性を備えた知的システムは説明可能AI (XAI: Explainable AI) と呼ばれる [53]。Adadi & BerradaはXAIの意義について、(i) AIの判断の正当化、(ii) AIに対する制御性の向上、(iii) AIの改善、(iv) 学習結果が人に洞察を与えること、の4点を挙げている [1]。XAIのうち、AIエージェントを対象にしたものは目標志向 (Goal-oriented) XAIやExplainable agencyと呼ばれる [4, 31]。目標志向XAIは、人がAIエージェントの挙動を制御したり、意図せぬ挙動を回避するために重要である [14]。

目標指向XAIに向けたアプローチは多様である。Puiutta & Veithは目標指向XAIの手法を2本の軸で4つの象限に分類している [1]。1つ目の軸は、手法を本来的 (intrinsic)/事後的 (post-hoc) に二分する。本来的な手法とは、機械学習モデル自体に人が理解可能な構造を持たせるものである。人にとっての可読性が比較的高い決定木やアテンション機構をモデルに取り込んだり [26]、AIエージェントの目標のような人が解釈可能な変数を持つようネットワーク構造に制約を加えた深層学習モデルがある [7, 57]。事後的な手法は、既に構築されているブラックボックスな機械学種モデルに対して事後的に説明を与える手法である。Hayes et al. が提案する質疑応答システムは、AIエージェントが学習した行動をマルコフ決定過程 (MDP: Markov decision process) モデルによってモデル化し、このモデルをもとにAIエージェントが特定の場面でどの行動を選択するかや、特定の行動を取る理由を説明することができる [25]。もう1つの軸は、生成される説明が大局的 (global) か局所的 (local) かである。大局的説明は、AIエージェントの挙動の概略を説明する [33, 25]。局所的説明は、特定の状況におけるAIエージェントの挙動を説明す

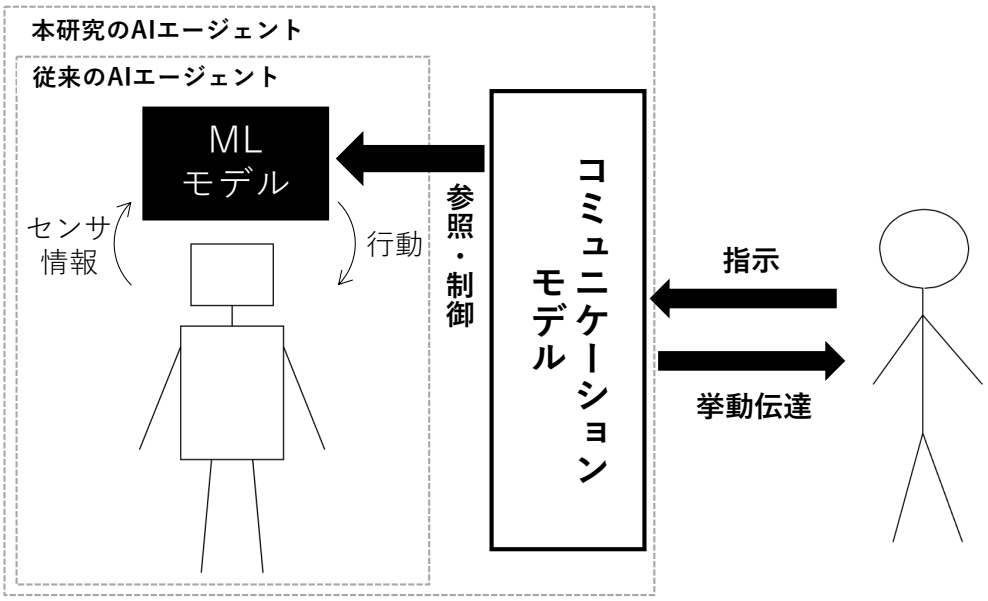


図 1.1: 人と AI エージェントをつなぐコミュニケーションモデル

る。局所的説明の方法として一般的なものの一つが、モデルに入力される情報のうち判断に強い影響を与えた部分を可視化する、サリエンシーマップを説明に用いる手法である [61, 28, 42]。

本論文でも、AI エージェントがブラックボックス化した機械学習器を持っている状況を考える。図 1.1 に枠組みを示す。本論文では、ブラックボックスとなっている機械学習モデルに対して人とのコミュニケーションを担うモデルを事後的に加え、このモデルを人と AI エージェントの接点としてコミュニケーションを実現させる方法を考える。また、コミュニケーションの内容は局所的なものを考える。具体的には、

(i) 人が AI エージェントに取らせたい挙動を示す「指示」

(ii) AI エージェントが選択した挙動を人に示す「伝達」

の交換によって、AI エージェントの取る挙動に関する共通認識を構築することを目指す。

1.3 心の理論と AI エージェント

AI エージェントと人との接点を担うコミュニケーションモデルは、どのように設計すればよいだろうか。コミュニケーションは、人同士の関わりにおいても重要である。人が生きる上で、他者との関わりを完全に断つことは現実的でない。我々はほとんどの場合、社会の一員として、コミュニケーションを通じて同じ社会の他者と協力し、利害の折り合いをつけ、時には欺きながら共存している。

人同士の社会的な関わりの基盤となる能力に、心の理論 (Theory of mind) がある。心の理論とは、他者の言動から信念や欲求、目標、意図といった心的状態を推測する能力である [61]。人は、表面的に観測可能な言動ではなく、背後にある心的状態という抽象化された変数によって人の言動を理由づけ（心的状態を帰属し）、将来の言動の予測に役立て

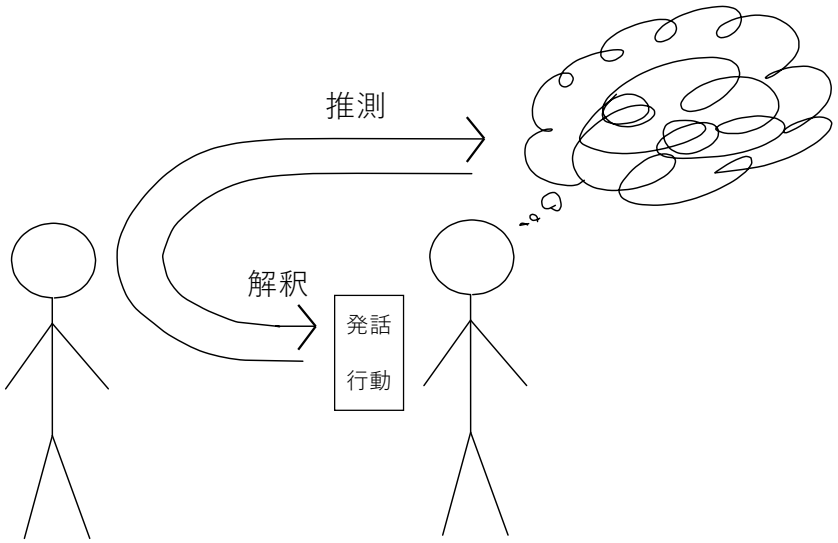


図 1.2: 心的状態の推定と言動の解釈

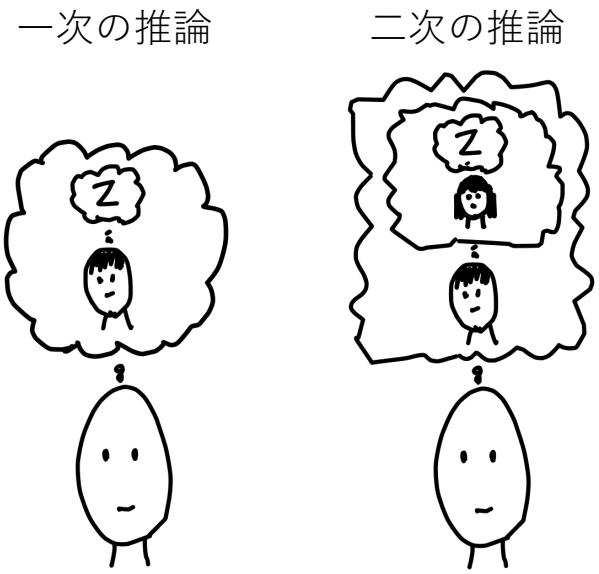


図 1.3: 心の理論と推測の次数

る（図 1.2）。また、心的状態を帰属することは、皮肉や欺瞞といった観測可能な言動と真の意図との不一致を認識することにもつながる。つまり、心的状態を考慮することによって、相手の言動を適切に解釈することができるようになる。

心的状態の推測は、時に再帰的な構造を持つ（図 1.3）。つまり、「アンは、『サリーが人形は箱の中にあると考えている』と思っている」のように、入れ子構造であらわれる心的状態を考えることができる。入れ子の深さは次数と呼ばれる。次数の定義に関しては研究によって相違があり、本論文では、他者の心的状態を推測することを 1 次の推測、上記の例文のように入れ子が 1 段階埋め込まれた推測を 2 次の推測と数えることにする。

人が心の理論を適用する対象は人だけではなく、ロボットや人形、さらには幾何学図形に対しても心的状態を帰属しようとすることがある [22]。心的状態を仮定して他者の言動

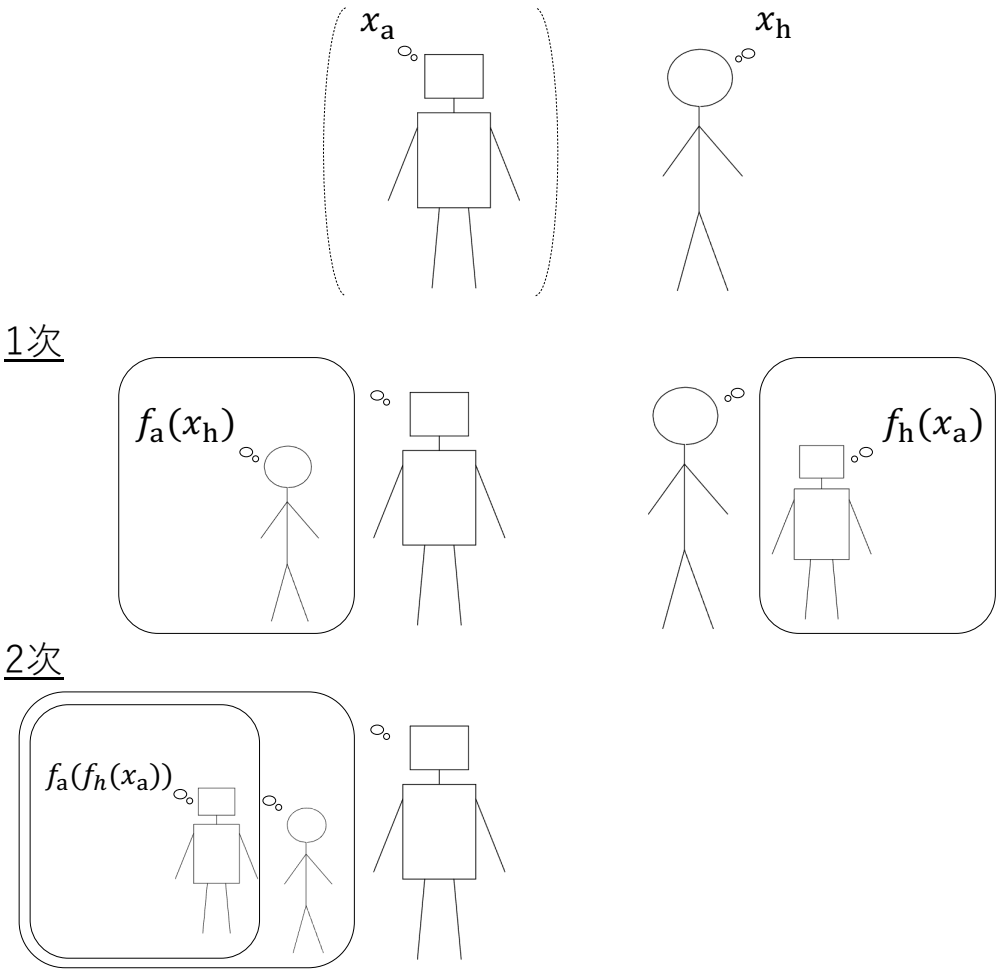


図 1.4: AI エージェントによる心的状態の推測

を理解しようとする人の態度は意図スタンスと呼ばれ、物理法則に従って運動すると考える物理スタンスや、特定の入力に対して定型的な出力をすることで動作していると考える設計スタンスと区別される [10]。これまでの研究で、人工エージェントのどういった要素が人の意図スタンスの採用に影響するかが調査されており、見た目の人らしさといった外見のほかに、動きの目標志向性や合理性、自己推進性、物理法則の違反などが要因として挙げられている [22, 48, 35, 55]。

AI エージェントは、人が意図スタンスを採用する要素を多く備えている。ロボットは外力が働いていない状態であっても、物理法則に抗って自己推進的に動く。合理的なエージェントは自らの効用を最大化しようと行動するものであり [29]、これはまさに強化学習が解こうとする問題である。強化学習の報酬関数は、エージェントが特定の目標を効率的に達成するよう設計される。すなわち、正の報酬がエージェントを特定の目標に向かわせ、負の報酬がより効率的な行動を選択するようエージェントを仕向ける。また、AI エージェントに擬人的な見た目を与えたり、人に意図スタンスを促す要素を付加することは、人と AI エージェントの社会的なインタラクションを促進するために有用である。以上のことから、人が AI エージェントに対して意図スタンスを採用し、言動から信念や意図を帰属しようとする状況は十分に想定される。

そこで、AIエージェントと人の二者間における心的状態の推測の構造を考える（図1.4）。AIエージェントと人が持つ心的状態をそれぞれ、x_a、x_hと記す。x_aとx_hは推測の構造を考えるための仮想的な概念であり、特に、AIエージェントが実際に心的状態x_aを持っているかは問わない。重要なのは、人がAIエージェントに意図スタンスを採用し、AIエージェントの行動がx_aに基づいていると信じていることである。

まず、AIエージェントを主体、人を客体とした一次の推測$f_a(x_h)$が考えられる。

$$f_a(x_h) = P(x_h|o_a).$$

f_aは、確率変数x_hを確率変数として、環境や相手に関してAIエージェントがそれまでに観測してきた事柄の全てo_aをもとに推測されるx_aの確率分布である。また、人がAIエージェントに対し意図スタンスを採用するならば、推測の主体と客体を入れ替えた一次の推測$f_h(x_a)$が生じる。これは人が推測するAIエージェントの心的状態の確率である。

$$f_h(x_a) = P(x_a|o_h).$$

すると2者は、一次の推測に対する推測を考えることができるようになる。すなわち、二次の推測$f_a(f_h(x_a)), f_h(f_a(x_h))$が発生する。

$$f_a(f_h(x_a)) = P(f_h(x_a)|o_a)$$
$$f_h(f_a(x_h)) = P(f_a(x_h)|o_h)$$

推測の入れ子は以降も無限に考えることができる。このうちAIエージェントを主体とする推測は、$f_a(x_h), f_a(f_h(x_a)), f_a(f_h(f_a(x_h))), ...$である。

以上の構造に、1.2節に掲げた(i)人からAIエージェントへの指示と、(ii)AIエージェントから人への挙動の伝達を当てはめて考える。(i)の背後には、人の意図や、AIエージェントに関する人の期待をx_hとして想定できそうである。AIエージェントは、与えられた指示やその場の文脈から人の目的や期待を推測し（$f_a(x_h)$）、推測した人からの期待と照らし合わせることで、与えられた指示をより正しく解釈できる。また、(ii)の目的は、人が推測するAIエージェントの目標や意図（$f_h(x_a)$）とAIエージェントの実際x_aを一致させることだと見なせる。AIエージェントの側からは、人からAIエージェントの言動がどのように見られるかを推測することで（$f_a(f_h(x_a))$）、実際との相違を検知し、相違に合わせてより効率的・効果的に挙動を伝達できると期待される。

1.4 挙動アライメント・コミュニケーションモデル

本論文は、人・AIエージェント間のコミュニケーションを通じた挙動アライメントを目指した2つの研究をまとめる（図1.5）。2つの研究では、AIエージェントによる1次・2次の心的状態の推測を行う機構を人とのコミュニケーションモデルに組み込み、人とのコミュニケーションに実際に応用する中で挙動アライメントへの寄与を調査している。

第1の研究は、AIエージェントの人に対する1次の推測の中でも、人がAIエージェントに達成を期待する目標を推測する「期待されるエージェント」モデルを提案する。「期

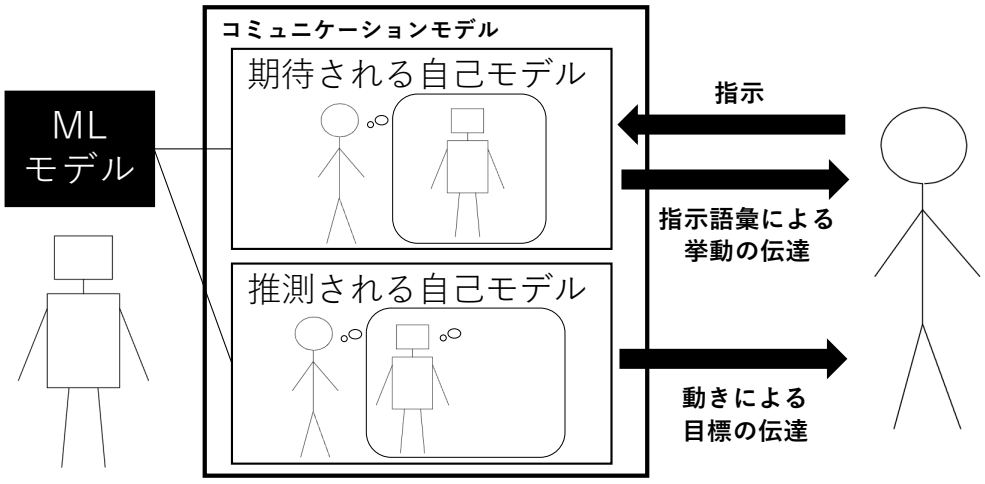


図 1.5: 「期待されるエージェント」モデルと「推測されるエージェント」モデル

待されるエージェント」モデルは、(a) 人から与えられる指示の背後にある目標の推定と、(b) 与えられた指示の解釈、の両者を整合させるように学習を行う。「期待されるエージェント」モデルを検証する実験では、与えられた指示をもとに人の目標を正しく認識できるようになること、人の目標を考慮することで、与えられた指示を正しく解釈できるようになることを示した。さらに、解釈した語彙を、AI エージェントが見せようとしている動きの予告に流用することで、人が AI エージェントの動きをより正確に予測できるようになることがわかった。

第 2 の研究では、AI エージェントの動きを見た人が AI エージェントに対して帰属する目標を、AI エージェントの側から推測する、という 2 次の推測を行う「推測されるエージェント」モデルを提案する。「推測されるエージェント」モデルは、AI エージェントと人が異なる視点に立っていることによる人の認識の遅れや誤解を検知することができる。さらに、「推測されるエージェント」モデルを応用することで、目標を人に伝達する動き (表意動作) を生成する手法を提案する。ユーザスタディの結果、このモデルが生成した表意動作によって、人が AI エージェントの目標を早く正確に推測できるようになることがわかった。

本論文では「期待されるエージェント」モデルと「予測されるエージェント」モデルを挙動アライメント・コミュニケーションモデルと総称する。本論文の趣旨は、2 モデルの実装と検証結果を通じて、挙動アライメントを達成するためのコミュニケーションモデルの設計を議論することである。

1.5 本論文の構成

2 章では、AI エージェントと人とのコミュニケーションに関する先行研究を紹介する中で、挙動アライメント・コミュニケーションモデルを設計するうえでの具体的な課題を示す。3 章では、「期待されるエージェント」モデルを提案し、シミュレーション実験とユーザスタディの結果を報告する。4 章では、「推測されるエージェント」モデルを提案

し、目標を伝達する動きを評価した実験について述べる。5 章で、これまでの検証結果を踏まえて、挙動アライメントにおけるコミュニケーションモデルを議論し、その未来を展望する。最後に、6 章で、本論文を総括する。

2 章

技術的背景

2.1 強化学習

本論文では、強化学習によって行動を学習し、意思決定を行っている AI エージェントを想定する。強化学習は機械学習の一種であり、エージェントが試行錯誤の中で行動 a を決定する方策 π を獲得することを可能にする [60]。強化学習が扱う問題設定を定式化するマルコフ決定過程（MDP: Markov decision process）は、典型的には $\langle \mathcal{S}, \mathcal{A}, T, R \rangle$ の組によって表現される。$\mathcal{S}$ を状態空間、$\mathcal{A}$ を行動空間と呼ぶ。エージェントは、各時刻 t における環境の状態 $s_t \in \mathcal{S}$ をもとに行動 $a_t \in \mathcal{A}$ を選択する。

$$a \sim \pi(s, a) \tag{2.1}$$

行動 a_t により、環境の状態は s_{t+1} に変化する。T は状態遷移関数と呼ばれ、s における a によって環境の状態が s' となる確率を記述する。

$$s' \sim T(s, a, s') = P(s'|s, a) \tag{2.2}$$

また、エージェントには報酬 r が与えられる。報酬を決定する R を報酬関数と呼ぶ。

$$r = R(s, a) \tag{2.3}$$

強化学習の目標は、得られる報酬の累積を最大化する最適方策 π^* を学習することである。

$$\pi^* = \mathbf{argmax}_\pi \mathbf{E}_\pi [\sum_{k=0}^{\infty} \gamma^k r_{t+k}] \tag{2.4}$$

γ は、割引率呼ばれる。

問題設定によっては、エージェントは s に直接アクセスできず、ノイズや観測できる範囲の制限といった不確実性を受けた状態で観測 o として取得する。

$$o \sim O(s, o) = P(o|s)$$

O を観測関数と呼ぶ。こうした条件下での強化学習は、部分観測マルコフ決定過程（POMDP: partially observable Markov decision process）というモデルの下に定式化できる。POMDP は、$\langle \mathcal{S}, \mathcal{A}, T, R, \Omega \rangle$ の組によって表現される。Ω は観測 o の集合であり、観測空間と呼ぶ。

部分観測マルコフ決定過程における強化学習では、エージェントの信念 b を考えることがある。b_t は、時刻 t までの観測 $o_{:t} = (o_0, o_1, ..., o_t)$ を与えられたときに、環境の状態が s である確率を表現する確率分布である。

$$b_t(s) = P(s_t = s|o_{:t})$$

2.2 マルチエージェント・システム

マルチエージェント・システム [64] は、複数のエージェントが互いに影響を与えながら分散的に意思決定を行う過程を、計算機によってモデル化しようとする研究分野である。エージェントが他のエージェントの内部状態や行動を推測・予測し、その結果を意思決定に利用することは、マルチエージェント環境において協調な行動や競争的な行動を選択する上で有効である [50]。こうした発想は、人間が他者の心的状態を推測する心の理論との類似点があり、実際にマルチエージェント・システムによって人間の社会性や心の理論をモデル化しようという取り組みは数多く存在する [54, 49, 47]。Zettlemoyer *et al.* は、1.3 節に挙げた推測の入れ子構造がエージェント同士で生じる場面で、入れ子構造の推測を逐次的な観測をもとに効率的に処理するアルゴリズムである sparse distributions over sequences (SDS) フィルタリングを提案している [68]。

2.3 言語による挙動アライメント

言語は人同士のコミュニケーションにおける主要なメディアであり、人とコンピュータのインタラクションにおいても活用が期待される。本節では、目標志向 XAI 分野において、言語を利用してエージェントの行動を説明する研究を紹介する。さらに AI エージェントと人との間の言語によるコミュニケーションを扱う研究として、指示遂行 (instruction following) とインタラクティブ強化学習の従来研究を紹介する。

2.3.1 目標志向 XAI における言語

Hayes *et al.* は、AI エージェントの行動に関するユーザの疑問に自然言語で応じる質疑応答システムを提案している [25]。このシステムは、“When will you pick up the widget?”、“Why didn’t you inspect the part?”、“What will you do when a human is near you?” といった、特定のテンプレートに従った疑問文を入力とし、AI エージェントの行動をモデル化した MDP モデルをもとに、回答を生成する。Waa *et al.* の手法では、AI エージェントの一回限りの行動ではなく、行動の系列を言語によってで説明する手法を提案している [63]。

しかし、これらの手法では主にグリッド環境 [60] を移動する AI エージェントが対象となっており、それ以外のドメインの AI エージェントに応用するには課題がある。課題の一つが、AI エージェントの行動を表現する語彙を、設計者が定義する必要のある点である。グリッド環境のように単純な環境では、語彙を人手で定義することは比較的簡単である。しかし、例えば AI エージェントがロボットのモータ制御を行動として扱っている場

合、行動空間の多次元性や行動出力から実際の動きになるまでの時間遅れ、行動を出力する頻度が高周期といった性質を持ち、AI エージェントの行動とそれを表現する語彙を人手で定義することが難しくなる。ユーザとのインタラクションの中で適応的に語彙を獲得できれば、AI エージェントと人の相互理解を構築するにあたって、有益だと期待される。

3 章で提案する「期待されるエージェント」モデルは、人の指示に用いられる語彙を自律的に学習し、学習した語彙をもとに AI エージェントの動きを予告することができる。

2.3.2 指示遂行

指示遂行タスクでは、人から与えられた指示に沿った行動を AI エージェントに学習させることを目指す [9, 38, 3, 37]。指示遂行タスクに対するシンプルなアプローチの一つは、AI エージェントの動きとそのキャプションの対をもとに、対応関係を教師あり学習によって獲得する方法である [36, 2]。

一方、報酬ベース (reward-based) アプローチは指示遂行タスクを強化学習によって解決しようとする方法である [56, 27]。報酬ベースアプローチでは、AI エージェントは環境の状態 s と人からの指示 u を入力され、出力した行動 a によって得られる報酬 r を最大化するように、方策 $\pi_{\text{instruction}}$ を学習する。

$$a \sim \pi_{\text{instruction}}(a_t|s_t, u_t) \propto \mathbf{E}_{\pi_{\text{instruction}}}[\sum_{0 \leq \tau} \gamma^\tau r_{t+\tau}],$$

指示遂行タスクに目標志向 XAI の概念を取り込んだ研究として、Shu *et al.* は階層型強化学習手法を提案している [57]。彼らの手法では、指示遂行タスクをエージェントが学習する過程で、エージェントの方策が指示ごとに分化する。この分化した方策と指示の語彙を対応づけることで、エージェントの行動を説明できる。

2.3.3 インタラクティブ強化学習

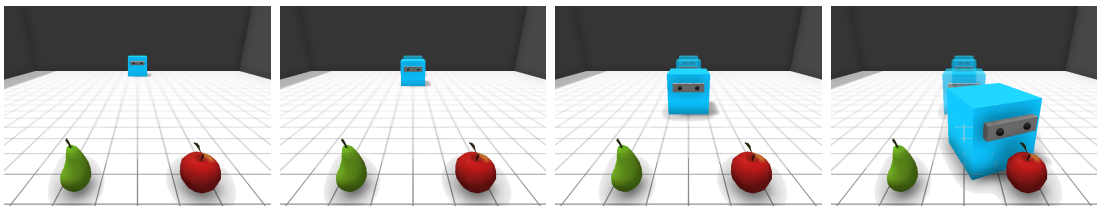
指示遂行は、人の指示にもとづいて AI エージェントの行動を決定するタスクであった。一方、AI エージェントが単独で行動を学習できる場面でも、人からの指示は有用である。

インタラクティブ強化学習は、人からのフィードバックを活用することで強化学習による行動の学習を高速化しようとする枠組である [34]。状態空間や行動空間が大きい複雑な状況では、行動の学習には非常に時間がかかる [30]。問題が特に顕著になるのは、強化学習を実世界に応用した際である。人からのフィードバックによって、学習の際に AI エージェントが探索すべき空間を狭めることができ、学習の高速化につながる。

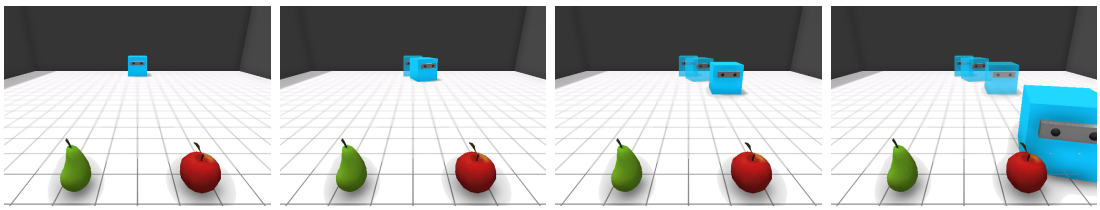
人からのフィードバックの方法は様々あり、例えば人が AI エージェントの行動に対して追加で報酬を与えるという方法が多い一方、言語を活用した方法もある。

2.4 表意動作

人のコミュニケーションチャンネルは言語だけではない。表情やジェスチャといった非言語情報も、コミュニケーションにおける重要な役割を担っている。人と AI エージェン



(a) Original motion ($t = 0, 17, 34, 50$)



(b) 表意動作 ($t = 0, 17, 34, 58$)

図 2.1: 表意動作

トのコミュニケーションにおいても、こうした非言語モダリティを活用することは有用だと考えられる。

Dragan *et al.* は、人が他者の行動から背後の心的状態を推測する性質（1.3 章）を逆手に取り、エージェントの目標を人に伝達する動き (**表意動作** = legible motion) を生成する手法を提案している [13]。この手法では、人がエージェントの目標を推測する過程を確率的にモデル化し、正しい目標を推測する過程を最大化させる行動系列をプランニングする。

図 2.1 に、表意動作の例を示す。図に示しているのは、ある観察者の視界から見たエージェントの動きである。エージェントは固定のスタート地点から、リンゴとナシのどちらかを目標に定めて行動する。図 2.1a の original motion は、エージェントが強化学習によって獲得した方策によって、実際に生成された動きである。エージェントは、より短い時間で目標にたどり着くよう学習している。観察者の視点から、エージェントの目標がリンゴであるかナシであるかを推測することを考えよう。Original motion の場合、エージェントが観察者の側に直進している間は目標をどちらかに決めることができず、観察者の前でリンゴの側に転回した際に初めて、エージェントの目標がリンゴであることに気づくことができる。図 2.1b には表意動作の例を示す。この例では、エージェントはスタート地点ですぐにリンゴの側に転回し、カーブを描いてリンゴにたどり着く。観察者の視点からは、エージェントが転回を始めた段階でエージェントの目標がナシであるという考えが排除され、目標がリンゴであると推測しやすくなっていると考えられる。実際に、表意動作によって、エージェントの動きを観察している人は、エージェントの目標をより早く、正しく推定できるようになる [12]。

2.5 先行研究の限界

先行研究はいずれも、人と AI エージェントのコミュニケーションによる相互理解を実現している。しかし、先行研究にはコミュニケーションを成立させるための暗黙の前提が

ある。具体的には、**目標**と**観測**という 2 種類の心的変数が人と AI エージェントの間で共有されていることを前提としている。本論文の主張は、人と AI エージェントの間に生じる目標と観測の非対称性を無視することがコミュニケーションの障害となる場面が存在するということである。3、4 章で取り上げる「期待されるエージェント」モデルと「推測されるエージェント」モデルは、この 2 つの非対称性の問題を解決して、AI エージェントと人の間でコミュニケーションを成立させることに取り組んでいる。

2.5.1 目標の非対称性

目標の非対称性は、AI エージェントに指示を与える人と AI エージェントの間で目標が共有されていない状況を指し、指示遂行やインタラクティブ強化学習と密接にかかわる課題である。指示遂行の報酬ベースアプローチでは、人と AI エージェントの間で目標 $g \in \mathcal{G}$ が共有されていることが暗黙の前提になっている。つまり、AI エージェントの目標 g_{agent} と人の目標 g_{human} が一致していると仮定している。ここで、g_{agent} とは、エージェントの報酬を決定する変数である。ここからは、式 2.3 に定義した報酬関数を以下に置き換える。

$$r = R(s, a, g_{agent}). \tag{2.5}$$

状態 s で取った行動 a に対する報酬は、g_{agent} によって異なる。そして、g_{human} は人の指示を決定する変数である。エージェントに指示を与える人を、関数 H としてモデル化しよう。

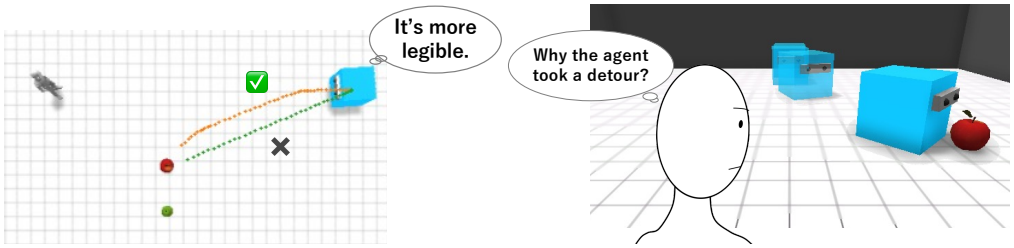
$$u = H(s, g_{human}). \tag{2.6}$$

指示遂行では、エージェントの行動が人の指示に従っているほどエージェントに大きな報酬が与えられる。これは、$g_{agent} = g_{human}$ だからである。しかし、エージェントの設計をよく知らない一般のユーザと AI エージェントのインタラクションを考えれば、g_{agent} と異なる g_{human} に関して人が言及する場合や、エージェントが達成しようとしているタスク以外を実行させようとする場合は容易に想定できる。3 章ではこうした状況を、人とエージェントの間で目標の非対称性が発生している場面と定義し、目標の非対称性がある中でもエージェントが人の指示を理解したり、エージェントの行動を人に説明することを考える。

2.5.2 観測の非対称性

観測の非対称性とは、人とエージェントがそれぞれの視点から異なる形で世界を観測していることであり、具体的には環境の中で両者の見える範囲が異なることを意味する。

表意動作の従来研究のほとんどはシンプルな環境を想定しており、人とエージェントは環境の状態を完全に把握できることを前提としている。しかし、一般に実世界の人やエージェントは不確実性にさらされており、一方が持っている情報を他方が持っていないという観測の非対称性が存在する。異なる観測によって構築される信念もまた非対称性が生じる。こうした観測や信念の非対称性は、人の心の理論を考える上でも重要な要素と考えら



(a) 観測の非対称性を考慮しない表意動作　　(b) 人の視界から見た動き

図 2.2: 観測の非対称性を考慮すべき例

れていて、例えば心の理論の能力を測定する心理課題であるサリーとアン課題は、観測の非対称性によって生じる誤った信念を理解できるかを問う [66]。

表意動作の生成においても、観測の非対称性が存在する場面ではエージェントの行動の解釈に相違が生じる。図 2.2 に具体例を示す。環境にはリンゴとナシが並んでいて、青色のエージェントはリンゴに向かって移動する（図 2.2a）。この際、エージェントの動きを見ている人に、エージェントの目標がリンゴであることを表現する表意動作を考える。リンゴへの最短経路は、リンゴに対して直線的に向かう動き（緑）である。一方、あえてナシを避けてリンゴの側にカーブを描く動き（オレンジ）を示すと、人から見てエージェントが目標をナシだと思う可能性が低下させられており、エージェントの目標をより効果的にエージェントに伝達する羽後になることが期待できる。

この動きは、図 2.2b の視界からどのように見えるだろうか。限定された視界からエージェントの動きを観測する人からは、リンゴの隣にあるナシがあることがわからない。すると、特に序盤の直進する動きは、エージェントが人の方向に向かっていると推測することもできる。人の視界が限られているこの例では、リンゴとナシの両方が観測できている場合と比べてエージェントの目標がナシと誤解されるリスクが減っている。そのため、オレンジの動きよりも、むしろリンゴに直線的に向かう緑の動きのほうが、目標を伝える動きとして適切である可能性もある。

Nikolaidis *et al.*[43] は、従来の表意動作の生成手法を拡張し、観察者の視点を考慮した手法を提案しており、しかし、彼らの手法で考慮しているのは観察者の視点からエージェントの動きを見た際の距離感の問題や、オクルージョンによってエージェントの動きが見えなくなる問題であり、観測の非対称性によって、一方が知っている環境中の物体の存在を他方がそもそも知らない、という状況は扱われていない。

4 章では、観測の非対称性が存在する場面で生じる人の誤解や、観測の非対称性を考慮した表意動作の生成を考える。

3 章

「期待されるエージェント」モデル

3.1 モデルの概要と目標と非対称性

本章では、「期待されるエージェント」モデルを提案する [20, 21][1]。「期待されるエージェント」モデルは、人がエージェントに求める挙動を推測する、エージェントから人への 1 次の心の推測の過程を組み込んだ挙動アライメント・コミュニケーションを行う。

「期待されるエージェント」モデルの特筆すべき特徴として、モデルが人とエージェントの間に存在する**目標の非対称性**を扱うことができる点が挙げられる。「期待されるエージェント」モデルでは、エージェント自体の目標 g_{agent} と人の目標 g_{human} を切り離し、人の指示から背後にある g_{human} を推定しようとする。そして、推定した g_{human} をもとに指示を解釈する。これにより、人が指示に用いた語彙の意味を適切に理解することができる。さらに、人の目標を推定しながら解釈した語彙を、エージェントの動きの予告に流用することで、エージェントが見せようとしている動きを人に説明できるようになる。

「期待されるエージェント」モデルでは、人がエージェントに達成させたい目標をもとに、AI エージェントに対して指示を与える状況を考える。図 3.1 に、本研究で扱う状況と、モデルによる推測の概要を示す。

「期待されるエージェント」モデルでは、人から与えられた指示語彙の解釈（vocabulary learning）と、人の指示の背後にある目標の推定（human-model learning）を、相互依存関係とみなし、両者の整合性が取れるように学習する。具体的には、まず人の指示からその背後にある目標を暫定的に推定する。人の目標が定まれば、指示遂行と同様の方法でエージェントの行動と人の指示の対応づけを学習することができる。つまり、「(A) 人の目標をもとに計算された報酬が多いほど、エージェントの行動が人の指示と対応している可能性が高い」という考えのもと、行動と指示語彙を対応づけることができる。これを、指示語彙の解釈と呼ぶ。次に、学習した指示語彙と行動の対応関係にもとづいて、人の目標を推定するモジュールを更新する。これは、「(B) エージェントの行動の指示語彙による表現が人からの指示と一致しているなら、エージェントの行動でより多くの報酬が獲得できる目標が人の目標である」という考えのもと、人の目標を推定していく。「期待されるエージェント」モデルは、(A) と (B) のそれぞれを表現する損失関数をもとにモデルを

[1]「期待されるエージェント」モデルの実装は、オンラインで公開されている (`https://github.com/fuku5/Manifestor`)。

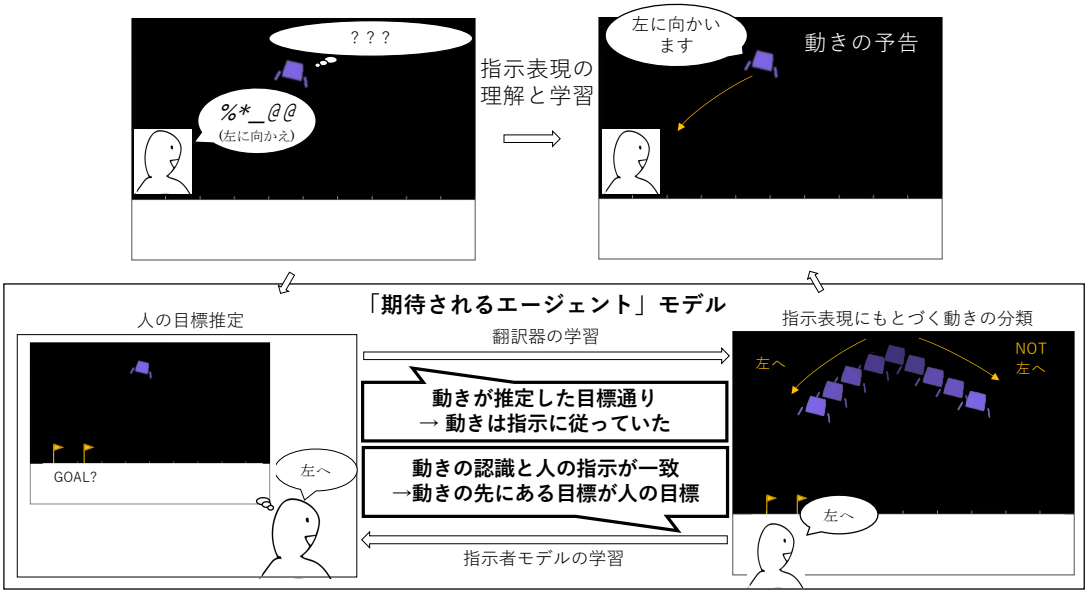


図 3.1: 「期待されるエージェント」モデルによる人の目標の推定と指示語彙の解釈

更新することで、人の目標を推定しながら、その推定結果をもとに指示の語彙を解釈することができる。

3.2 状況設定

3.2.1 環境

「期待されるエージェント」モデルを検証する環境として、Open AI Gym [8] が提供する LunarLander-v2 を用いた。LunarLander-v2 では、AI エージェントはロケットの操作を行う。ロケットの下部にあるメインスラスターと左右のサブスラスターを着火することで、ロケットを着陸地点に着地させることがタスクの目標である。エージェントが選択できる行動 $a \in \mathcal{A}$ は、3 つのスラスターを着火させるか、いずれも着火させないかの 4 通りであり、行動によってロケットが加減速する。スラスターを着火させないと、ロケットは重力と慣性に従って動く。環境の状態 $s \in \mathcal{S}$ は、ロケットの位置、速度、傾きからなる。エージェントに与えられる報酬は、着陸地点からの距離、減速量、傾きの減少量から計算される。また、ロケットが着陸に成功/失敗した時点で、単発の報酬が与えられ、エピソードが終了する。エピソードは、ロケットが着陸を初めてから月面で完全に静止するか、月面に衝突するまでの間を指す。

本研究において目標は、着陸地点の位置を指す。元来の LunarLander-v2 は着陸地点が画面の中央で固定であったため、試行毎に着陸地点を変えられるよう著者が改造を加えた。[2]

LunarLander-v2 で選択される行動は人にとって直観的でなく [51]、人の指示語彙とロケットの動きを対応づけることはグリッド環境に比べて難しい。例えば、同じ a でも、そ

[2]実装は、オンラインで公開されている (`https://github.com/fuku5/multi_lunar_lander`)。

の結果はロケットの速度や傾きによって異なり、左のスラスターを着火することが即座に右への移動に結びつくわけではない。また、行動は 20ms という高頻度で選択され、人の目に見える粒度の動きが単発の行動でなくエージェントの行動の系列によって生じる。そのため、行動と語彙の間には時間遅れも存在し、対応関係を複雑にしている。

3.2.2 指示者

人が AI エージェントに与える指示は、先行研究 [18, 17] をもとに単純なルールによって定めた。

$$H(s,g) = \begin{cases} \text{Go left. (if } s.x > g.x_{right}) \\ \text{Go right. (if } s.x < g.x_{left}) \\ \text{Fall straight down. (else),} \end{cases} \tag{3.1}$$

ここで、$s.x$ はロケットの水平方向の位置である。また、$g.x_{left}$ と $g.x_{right}$ はそれぞれ、着陸地点 g の左右の端の位置を表す。指示はロケットの位置と人の目標のみに依存し、ロケットの速度や慣性は考慮されない。

指示者が与えるのは指示のみで、指示遂行にあるような、エージェントの行動が指示に従っていたかというフィードバックは与えない。指示者からのフィードバックは学習を加速させるので、実際の応用ではこうしたフィードバックを考慮することは重要である。しかし、本研究では、指示の解釈とその背後にある目標の推定の整合性を取るというアイデアの有効性を検証するため、敢えてフィードバックが得られない設定を採用した。

3.3 「期待されるエージェント」モデル

3.3.1 構成モジュール

図 3.2 に、「期待されるエージェント」モデルを構成するモジュールを示す。「期待されるエージェント」モデルは、方策 π_g、予測器 T_N、翻訳器 f_N、指示者モデル M、評価器 E の 5 つのモジュールで構成される。π_g は、強化学習で一般的に定式化される方策と同じで、状態 s においてエージェントの行動 a を決定する。g は方策を学習する際の報酬関数を規定する変数である。T_N は、π_g による行動で生じる行動の系列 $\boldsymbol{a}_{t,N} = (a_t, a_{t+1}, ..., a_{t+N})$ と、それによる環境の遷移 $\boldsymbol{s}_{t,N} = (s_t, s_{t+1}, ..., s_{t+N})$ を N ステップ先の時刻まで予測する。

$$T_N(\pi_g, s_t) = (\boldsymbol{s}_{t,N}, \boldsymbol{a}_{t,N}).$$

f_N は、エージェントの行動が引き起こした環境の遷移 $\boldsymbol{s}_{t,N}$ と語彙 u を対応づける確率分布である。

$$u \sim f_N(u|\boldsymbol{s}_{t,N})$$

エージェントの動きの予告は、T_N で推測される $\boldsymbol{s}_{t,N}$ を f_N で語彙に変換することで生成できる。M は、状態と人の指示の系列からその背後にある目標を推定する。

$$g_{human} \sim M(g_{human}|\boldsymbol{s}_{0,\tau}, \boldsymbol{u}_{0,\tau}),$$

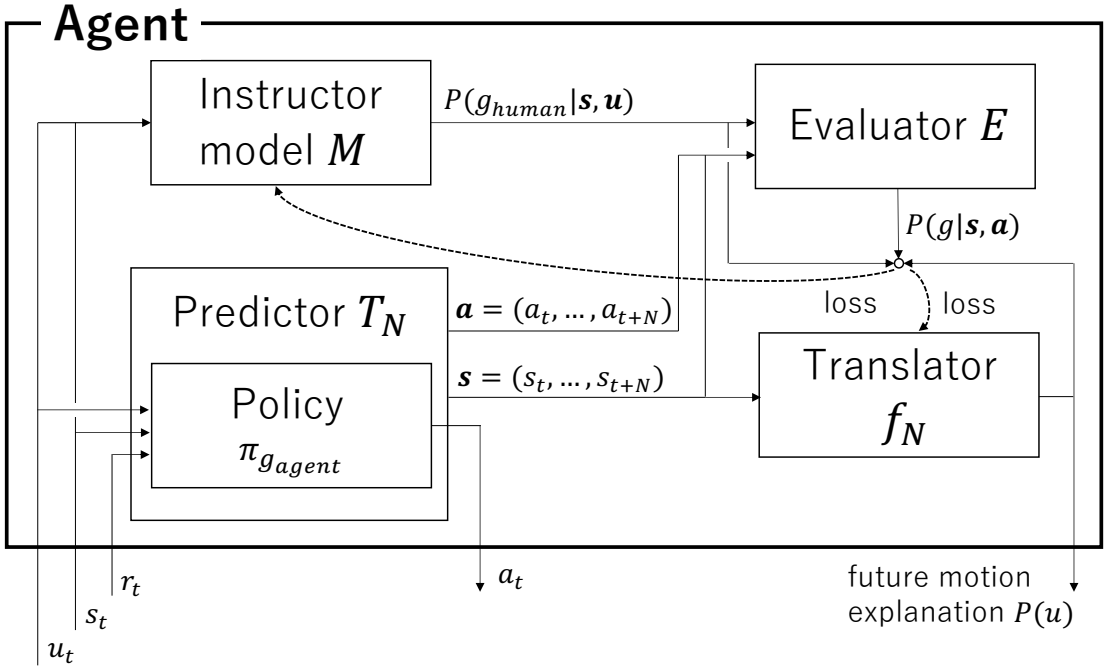


図 3.2: 「期待されるエージェント」モデルの構成モジュール

τ は、エピソードの長さで、$\boldsymbol{u}_{0,\tau} = (u_0, u_1, ..., u_\tau)$ はあるエピソードにおける人の指示の系列である。この定式化は、人の目標がエピソード内で一貫しているという前提に基づいている。最後に、E は、エージェントの行動系列とそれによる環境遷移 $(\boldsymbol{s}_{t,N}, \boldsymbol{a}_{t,N})$ が目標 g に適合している程度を計算する。

$$g \sim E(g|\boldsymbol{s}_{t,N}, \boldsymbol{a}_{t,N})$$

本研究では、E は $(\boldsymbol{s}_{t,N}, \boldsymbol{a}_{t,N})$ によって得られる報酬 $r = R(s, a, g)$ をもとに計算される。

$$E(g|\boldsymbol{s}, \boldsymbol{a}) = \underset{g \in \mathcal{G}}{\text{softmax}}(\sum_{s,a \in \boldsymbol{s},\boldsymbol{a}} R(s, a, g)).$$

本論文では、目標と非対称性が存在する場面で如何に f_N と M を学習させるかに焦点を絞り、T_N や E の学習や、精度等の問題に関しては議論しないこととする。

3.3.2 損失関数

「期待されるエージェント」モデルは、f_N と M の両者の整合性がとれるように損失関数を最小化するように学習することで獲得される。損失関数は、人の背後の目標がわかれば、指示遂行と同様に人の指示語彙とエージェントの動きを対応づけることができ（f_N の学習）、人の指示語彙とエージェントの動きを対応関係がわかれば、人の指示の背後にある目標が推定可能である（M の学習）、という図 3.1 に示した 2 つのアイデアを表現する。

f_N の損失関数

まず、エージェントと人の間で目標が共有されている場合、すなわち、$g_{agent} = g_{human}$ の場合を考えよう。人は状態 s_t と目標 g_{human} のもとに、エージェントに指示 u_t を与える

(式2.6を参照)。エージェントはそこからNステップにわたって行動し($\boldsymbol{a}_{t,N}$)、環境の遷移$\boldsymbol{s}_{t,N}$が得られる。f_Nの更新に用いられる損失関数L_{f_N}は、「$\boldsymbol{s}_{t,N}$が目標$g_{agent}(=g_{human})$に適合した動きであれば、$\boldsymbol{s}_{t,N}$がu_tに従っていた可能性が高い」というアイデアを表現する。

$$L_{f_N} = -E(g_{human}|\boldsymbol{s}_{t,N}, \boldsymbol{a}_{t,N}) \cdot \log f_N(u_t|\boldsymbol{s}_{t,N}). \tag{3.2}$$

f_Nは、L_{f_N}を最小化するように更新される。

それでは、目標の非対称性が存在する場面、すなわち$g_{agent} \neq g_{human}$の場合を考える。この場合の損失関数$L^+_{f_N}$を、L_{f_N}を拡張する形で定義する。

$$\begin{aligned} L^+_{f_N} = -\sum_{g\in\mathcal{G}} &(M(g|\boldsymbol{s}_{0,\tau}, \boldsymbol{u}_{0,\tau}) \cdot E(g|\boldsymbol{s}_{t,N}, \boldsymbol{a}_{t,N})) \\ &\cdot \log f_N(u_t|\boldsymbol{s}_{t,N}), \end{aligned} \tag{3.3}$$

$L^+_{f_N}$は、Mによる人の目標の推定結果を利用する。式3.3における$\sum(M \cdot E)$が、式3.2の$E(g_{human}|\boldsymbol{s}_{t,N}, \boldsymbol{a}_{t,N})$に対応している。$g_{human}$はエージェントから観測できないため、$L^+_{f_N}$では$g_{human}$が取りうる全ての目標$g \in \mathcal{G}$を考慮して計算される。$M \cdot E$は、$g_{human}$を確率変数とみなした際の、$E(g_{human}|\boldsymbol{s}_{t,N}, \boldsymbol{a}_{t,N})$の期待値である。

Mの損失関数

式3.4に、Mの訓練に用いられる損失関数L_Mを示す。

$$\begin{aligned} L_M = -\frac{1}{\beta} &f_N(u_t|\boldsymbol{s}_{t,N}) \\ &\cdot \sum_{g\in\mathcal{G}} (E(g|\boldsymbol{s}_{t,N}, \boldsymbol{a}_{t,N}) \cdot \log M(g|\boldsymbol{s}_{0,\tau}, \boldsymbol{u}_{0,\tau})), \end{aligned} \tag{3.4}$$

L_Mは、(a) エージェントの動きのf_Nによる認識結果が人から与えられた指示に適合しているなら、(b)g_{human}はその動きがより多くの報酬を獲得できる目標である可能性が高い、というアイデアを表現している。式3.4のf_Nと$\sum(E \cdot \log M)$が、それぞれ(a)と(b)に対応する。式3.4によって、(a)と(b)が共起するようにMが更新される。すなわち、f_Nが大きいときに$\sum(E \cdot \log M)$が最大化($-\sum(E \cdot \log M)$が最小化)される。

式3.4は(a)と(b)の共起関係を表現しているものの、(a)と(b)の双方の項をそれぞれ減少させることでも値が小さくなるという抜け穴が存在する。そこで、共起関係を捉えるためのペナルティ項としてβが追加されている。

$$\begin{aligned} \beta = &\mathbf{E}_t[f_N(u_t|\boldsymbol{s}_{t,N})] \\ &\cdot \mathbf{E}_t[\sum_{g\in\mathcal{G}} (E(g|\boldsymbol{s}_{t,N}, \boldsymbol{a}_{t,N}) \cdot \log M(g|\boldsymbol{s}_{0,\tau}, \boldsymbol{u}_{0,\tau}))]. \end{aligned}$$

f_N, E, Mの出力にソフトマック関数が用いられ、0にならないために、$\beta = 0$の場合に関しては考慮しない。

3.4 実装

3.4.1 学習の手順

f_N と M は、互いの推論結果をもとに学習しており、学習には相互依存性がある。著者の試みた範囲では、式 3.3、3.4 によって f_N と M を同時に学習させると、学習が不安定になり結果を収束させることができなかった。本研究では $L^+_{f_N}$ と L_M の妥当性の評価に焦点を絞るため、いくつかの仮定と学習の手順の切り分けを行うことで、問題の単純化を行った。手順は 3 段階に分かれる。

最初の段階では、L_M と無関係に教師なし学習によって指示を n 個のグループに類別するモデルを学習させる。教師なし学習モデルには encoder-decoder モデルを採用した。

$$\hat{g} \sim Encoder(\hat{g}|\boldsymbol{s}_{0,\tau}, \boldsymbol{u}_{0,\tau}), \tag{3.5}$$

$$u_t \sim Decoder(u_t|s_t, Encoder(\boldsymbol{s}_{0,\tau}, \boldsymbol{u}_{0,\tau})), \tag{3.6}$$

$\hat{g} \in \hat{\mathcal{G}}$ が、教師なし学習によって類別された結果になる。

次に、$\hat{g}$ をもとに f_N を学習させる。第一段階で得られる教師なし学習の結果からは、$\hat{g} \in \hat{\mathcal{G}}$ と $g \in \mathcal{G}$ の間の関係は得られない。そのため、$\hat{g}$ と g の間の射影 $m : \hat{\mathcal{G}} \to \mathcal{G}$ は複数考えられる。第二段階では、考えられる全ての m に関して f_M の学習を行う。

$$M_m(g|\boldsymbol{s}_{0,\tau}, \boldsymbol{u}_{0,\tau}) = \sum_{\hat{g}\in\hat{\mathcal{G}}} \delta(g, m(\hat{g})) \cdot Encoder(\hat{g}|\boldsymbol{s}_{0,\tau}, \boldsymbol{u}_{0,\tau}), \tag{3.7}$$

ここで、$\delta(a, b)$ は $a = b$ の時に 1 を、それ以外の時に 0 を返す関数である。本論文では、$|\mathcal{G}| = |\hat{\mathcal{G}}| = 3$ であり、$\hat{g}$ と g の間に 1 対 1 の対応関係があると仮定する。すると、m は $3! = 6$ 通り考えられ、結果として 6 通りの f_N が得られる。

第三段階で、それぞれの m から得られる M_m を式 3.4 によって評価する。学習の最終的な結果は、3.4 が最も小さくなる M_m と、それをもとに学習した f_N である。このように、実装では M の学習を、m の選択という問題に単純化することで、学習を安定させている。

3.4.2 モデル

f_N は、Transformer-Encoder model [62] をもとに実装した（図 3.3）。Transformer-Encoder model は時系列データを扱うことのできる深層学習モデルである。モデルは、式 3.3 をもとにした勾配降下法で更新される。Transformer-Encoder model の入力の先頭には [CLS] トークンを入力した [11]。f_N の出力は、[CLS] トークンの位置に出力されるベクトルを多層パーセプトロンとソフトマックス関数によって変換したものである。

式 3.5 の *Encoder* も、f_N と同様のモデルで実装した。ただし、*Encoder* に入力されるのは、s と u を連結したベクトルである。*Encoder* の出力は、指示の背後にある $\hat{g}$ の確率分布を表現する。*Decoder* (式 3.6) は、多層パーセプトロンで実装した。

表 3.1・3.2 に、f_N と *Encoder* のハイパーパラメータを示す。

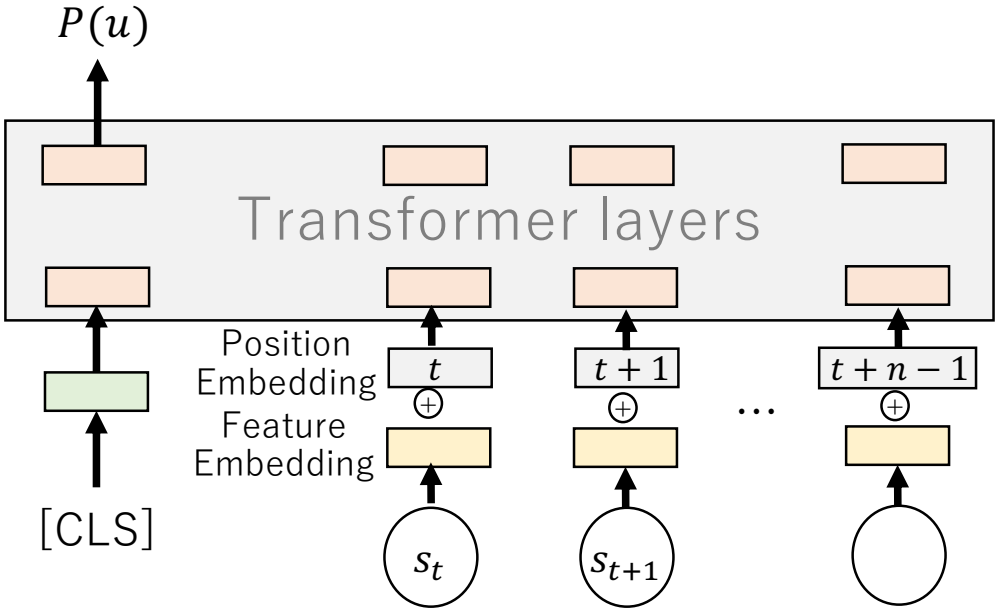


図 3.3: f_N の実装

表 3.1: f_N のハイパーパラメータ

Hyperparameter	Value
The number of Transformer-Encoder layers	2
The number of Transformer-Encoder heads	2
Dimension of the Transformer-Encoder input	32
Dimension of Transformer-Encoder feedforward network model	1024
Dropout rate	0.5

表 3.2: *Encoder* のハイパーパラメータ

Hyperparameter	Value
The number of Transformer-Encoder layers	2
The number of Transformer-Encoder heads	2
Dimension of the Transformer-Encoder input	128
Dimension of Transformer-Encoder feedforward network model	1024
Dropout rate	0.5
Hidden sizes for *Decoder*	[256, 256]

3.5 実験概要

「期待されるエージェント」モデルを評価する 2 つの実験を行った。数値実験では、「期待されるエージェント」モデルの基本的な動作を評価した。また、ユーザスタディでは「期待されるエージェント」モデルによって獲得された f_N をエージェントの動きの説明に流用することで、ユーザがエージェントの見せようとしている動きを予測できるようになるか検証した。

「期待されるエージェント」モデルの比較対象として、*optimal* と *ablation* の 2 種類を用意した。どちらも、「期待されるエージェント」モデルが目標の非対称性を適切に扱うことが出来ているかの検証に利用した。*optimal* は、目標の非対称性が存在しない理想的な状況において g_{agent} と L_{f_N} をもとに学習したもので、「期待されるエージェント」モデルの正解となる結果を出力すると期待される。*ablation* は、目標の非対称性が存在するにも関わらず、人とエージェントの間で目標が共有されているという誤った仮定のもと、g_{agent} と L_{f_N} によって学習された結果である。*optimal* と *ablation* は、既存の指示遂行手法 [56, 27] を目標の非対称性が存在する場面に適用した際の結果を再現するものである。

3.6 数値実験

3.6.1 目的

数値実験では、「期待されるエージェント」モデルの基本的な動作を評価した。具体的には、以下の 2 つの疑問に回答することを目的に実験を行った。

(i) 損失関数 L_M は、M の最適なマッピング m^* を選択できるか？

(ii) 目標の非対称性が存在する場面で $L^+_{f_N}$ をもとに学習した「期待されるエージェント」モデルの学習結果は、*optimal* と一致するか？

これらの疑問は、L_M と L_{f_N} のそれぞれの検証と対応する。

3.6.2 方法

アルゴリズム 1 に、実験に用いるデータを作成する流れを示す。

AI エージェントの方策は、代表的な深層強化学習の手法の 1 つである Advantage Actor-Critic (A2C) によって訓練した。A2C モデルのハイパーパラメータを表 3.3 に示す。得られた方策をもとに、「期待されるエージェント」モデル、*optimal*、*ablation* を訓練・評価するためのデータセットを作成した。データセットは、s, g_{human}, g_{agent} の組からなる。g_{human} は、エピソード毎にランダムに決定した。1.5 億ステップにわたって学習させた方策による skilled データセットのほかに 50 万ステップで学習を打ち切った場合の unskilled データセットを作成し、AI エージェントの方策の性能がモデルの学習結果に与える影響も調べた。データセットはそれぞれ 3,200 のエピソードで構成され、半分をモデルの訓練に、残りの半分を評価に利用した。

Algorithm 1 Training and evaluating 「期待されるエージェント」モデル

$\pi \leftarrow$ a policy of an A2C agent
// Preparation
Build a dataset of tuples $(s, a, g_{agent}, g_{human})$
dataset_training, dataset_evaluation $\leftarrow$ dataset.split()
$\vec{s}, \vec{a}, \vec{g_{agent}}, \vec{g_{human}} \leftarrow$ dataset_training
$\vec{u} \leftarrow H(\vec{s}, \vec{g_{human}})$
$\vec{u}' \leftarrow H(\vec{s}, \vec{g_{agent}})$

for $i = 1, 2, ...,$ to NUM_SEED **do**
    // For drawing histogram
    Train *Encoder* with $(\vec{s}, \vec{u})$
    $m^* \leftarrow \mathbf{argmax}_m$ Accuracy$(M_m; \vec{g_{human}})$
    for possible m **do**
        Build M_m with *Encoder* and m (See Eq. 3.7.)
        $f_{N,\mathbf{ExpectedSelf},m} \leftarrow$ Train f_N with $(L^+_{f_N}, M_m, \vec{s}, \vec{a}, \vec{u})$
        Calculate L_{M_m} with dataset_evaluation
    end for
    Calculate $L_{M_m}/L_{M_{m^*}}$ for each $m(\neq m^*)$

    // For comparing accuracy
    $f_{N,\mathbf{optimal},m^*} \leftarrow$ Train f_N with $(L_{f_N}, M_{m^*}, \vec{s}, \vec{a}, \vec{u}')$
    $f_{N,\mathbf{ablation},m^*} \leftarrow$ Train f_N with $(L_{f_N}, M_{m^*}, \vec{s}, \vec{a}, \vec{u})$
    Calculate how much the outputs of $f_{N,\mathbf{ExpectedSelf},m^*}$ match those of $f_{N,\mathbf{optimal},m^*}$
    Calculate how much the outputs of $f_{N,\mathbf{ablation},m^*}$ match those of $f_{N,\mathbf{optimal},m^*}$
end for

3.7 節のユーザスタディにおける、「推測されるエージェント」モデルが予告を提示する時間と、ロケットの動きをユーザが予測する際の難易度の兼ね合いを考慮して、$N = 100$（5 秒）に設定した。

疑問 (i) を検証するため、3.4.1 項の手順を異なる 100 の乱数シードをもとに実行した。それぞれの乱数シードで、m の異なる 6 種類の (M, f_N) が得られる。また、m の中には正解ラベル g_{human} を M_m が予測する際の精度を最も高くする射影 m^* が存在する。ここでは、$L_{M_{m^*}}$ に対する L_{M_m} $(m \neq m^*)$ の比を、結果を評価する指標とした。比が 1 より大きくなれば、損失関数 L_{M_m} が最適な射影 m^* を選択できることを意味し、M の損失関数としての L_{M_m} を正当化する結果となる。

疑問 (ii) を検証するため、「期待されるエージェント」モデルの精度を計算した。ここ

表 3.3: A2C の学習に用いたハイパーパラメータ

Hyperparameter	Value
Update steps	5
Discount factor γ	0.995
Optimizer	RMSprop
RMSprop epsilon	1e-5
Learning rate	7e-4
Hidden activation	ReLU
Hidden sizes	[512, 512, 512]

で精度とは、「期待されるエージェント」モデルによって獲得された f_N の出力結果がどれだけ *optimal* の結果と一致するかを示す。$L^+_{f_N}$ の評価から L_{M_m} の影響を排除するため、ここでは L_{M_m} によって最適な射影 m^* が選ばれることを前提に学習を行った。結果の比較対象として、*ablation* の精度も検証した。

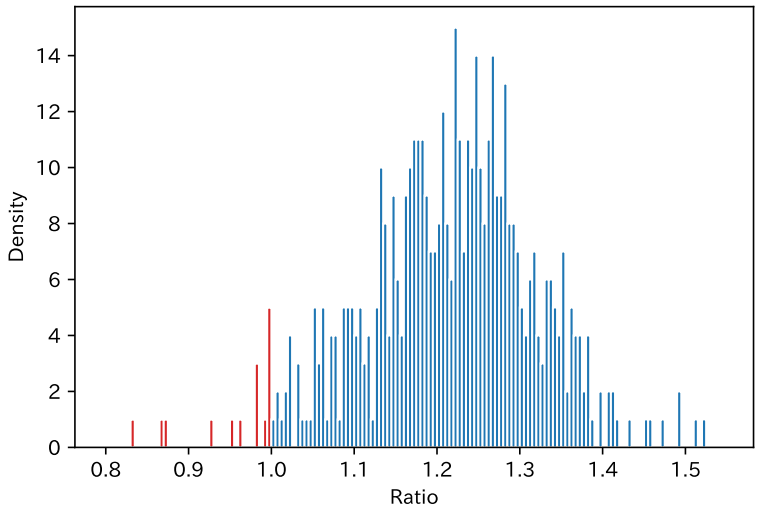
3.6.3 結果

疑問 (i) に対する検証の結果を図 3.4 に示す。unskilled データセットと skilled データセットのそれぞれで、500 のサンプルのうち 485/487 サンプル（97.0 % / 97.4 %）で比の値が 1 を超えており、L_M によって最適な射影 m^* を選択することができたといえる。データセット間で、分布の幅に差が表れたものの、精度の違いはほとんど生じなかった。unskilled データセットでの比の平均は 1.22 (95% CI[3] 1.01, 1.42)、skilled データセットでの比の平均は 1.14（95% CI 0.99, 1.29）であった。最適な射影 m^* を選択できたという結果は、「期待されるエージェント」モデルが損失関数 L_M によって指示者の背後にある目標を正しく推定できるということを示唆している。

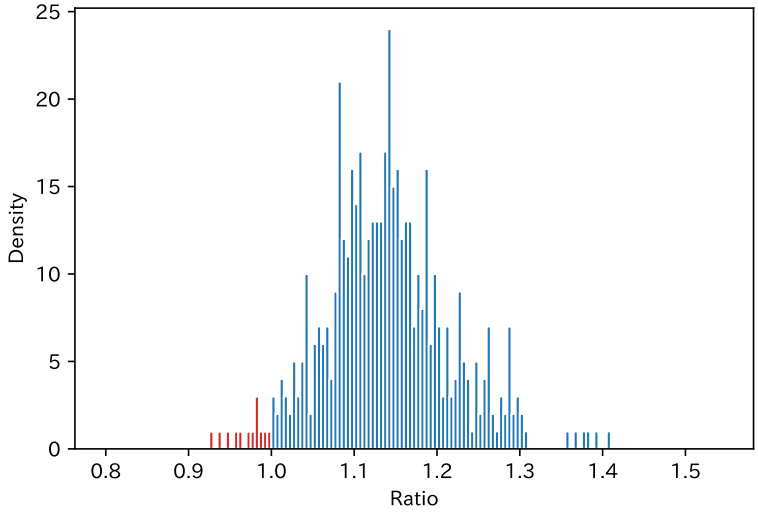
図 3.5 に、「期待されるエージェント」モデルと *ablation* の精度を示す。unskilled データセットと skilled データセットのいずれにおいても、「期待されるエージェント」モデルの学習結果の精度は *ablation* と比較して有意に高かった (Mann-Whitney's U test, $p < .001$)。「期待されるエージェント」モデルの精度は、平均値で.700（unskilled）と.870（skilled）である一方、*ablation* では.303 と.522 にとどまった。unskilled データセットで精度が低下した理由としては、ロケットが月面に衝突したり、画面外に出てしまうという失敗が多く、実際に指示にしたがって動く例が少ないために、指示の語彙とロケットの動きを対応づけることが難しかったことが挙げられる。

数値実験の結果は、3.6.1 節に挙げた 2 つの疑問を肯定する結果であった。「期待されるエージェント」モデルは、損失関数 L_M によって、目標の非対称性が存在する状況で指示者の背後にある目標を正しく推定できた。また、損失関数 $L^+_{f_N}$ によって、指示の背後にある目標の推定結果をもとに、指示語彙を適切に解釈することができた。

[3]Confidence interval



(a) Unskilled dataset



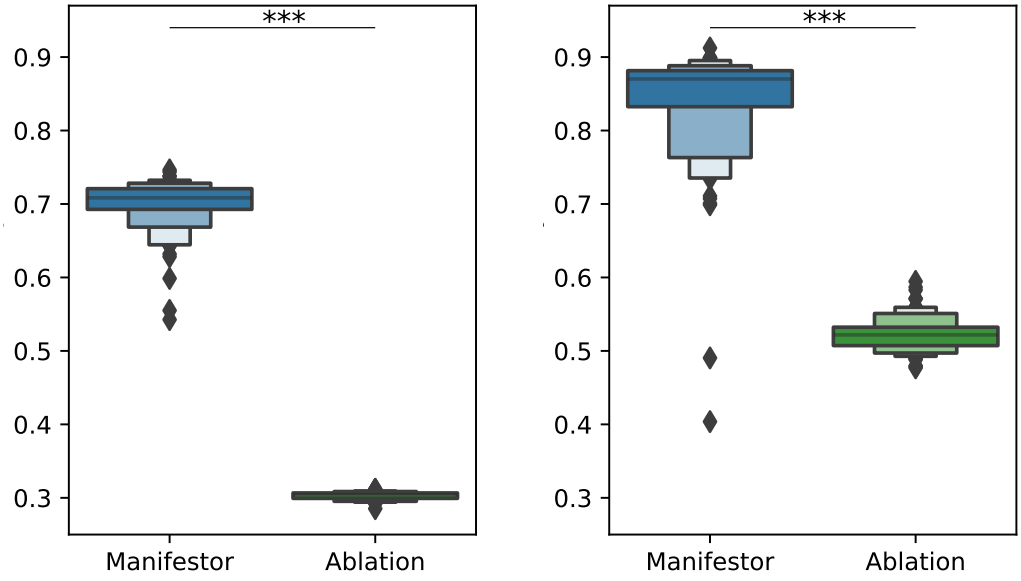
(b) Skilled dataset

図 3.4: 比のヒストグラム
青い領域は比が 1 より大きい場合、赤い領域は比が 1 以下の場合を表す。[21]

3.7 ユーザスタディ

3.7.1 目的

ユーザスタディでは、「期待されるエージェント」モデルをより実践的な場面で評価した。ここでは、「期待されるエージェント」モデルが獲得した f_N をもとに AI エージェントの動きの予告を生成し、実験参加者に提示することで、エージェントの動きの予測性を向上させることができるか検証した。



(a) Unskilled データセット　　(b) Skilled データセット

図 3.5: 「期待されるエージェント」モデルと *ablation* によって獲得された f_N の精度 [21]

3.7.2 方法

参加者は、AI エージェントが操作するロケットの動きを途中まで提示され、その後の 5 秒間（$N = 100$ フレーム）でロケットが月面のどの位置に着陸するかを問われた。図 3.6 に、実験に用いたインタフェースを示す。参加者には、ロケットの動きとともに、生成された予告の確率が棒グラフで提示した。予告の提示に棒グラフを使用したのは、先行研究 [18] において、語彙が当てはまる程度を表現することが動きの予測可能性を向上に寄与することがわかったためである。比較のために、予告として「期待されるエージェント」モデルの他に、*ablation* と *optimal* の結果も用意した。さらに、予告を全く見せないベースライン条件も設定した。

実験は、参加者間計画で実施した。参加者は、日本のクラウドソーシングプラットフォームである Lancers[4]を利用して募集した。結果、100 人の参加者が集まった (女性 26 名、男性 74 名; 21-67 歳, $M = 41.3, SD = 8.6$)。参加者には謝金として 120 円を支払った。参加者に始めに実験の内容を説明し、全員から参加への同意を得た。実験に先立って、タスクの理解度を問う簡単なクイズを 4 問出題し、正答しなかった 14 名は評価から除外した。最終的に、「期待されるエージェント」モデル条件、*optimal* 条件、*ablation* 条件、ベースライン条件に、それぞれ 18、20、22、26 名の参加者が振り分けられた。タスクでは、参加者はロケットの動きと生成された予告を見て、月面に記載された数字の中からロケットが着陸する地点として最も近いと予測するものを回答した（図 3.6）。参加者に提示する

[4]https://www.lancers.jp/

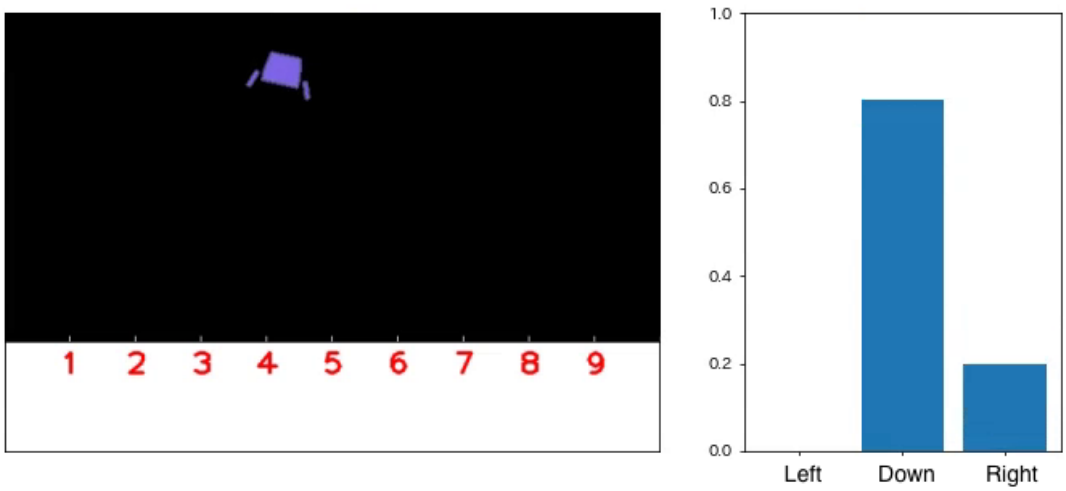


図 3.6: ユーザスタディに用いたシステムのインタフェース [21]

ロケットの動きとして 20 エピソード分を用意し、条件に応じて予告だけ差し替えた動画を、ランダムな順番で提示した。

参加者が予測する着陸地点と実際の着陸地点の距離を指標とし、より誤差が少なくなる予告を適切な予告とした。

3.7.3 仮説

「期待されるエージェント」モデルが目標の非対称性を考慮して f_N を適切に獲得していること、獲得された f_N による予告がロケットの動きの予測性を高めることができるか調べるため、2 つの仮説を設定した。

(H1) 「期待されるエージェント」モデルの予告を見た参加者が予想する着陸地点の誤差は、*optimal* と同程度に小さい。

(H2) *Ablation* の予告を提示された参加者の誤差は、「期待されるエージェント」モデルの場合より大きい。

3.7.4 結果

図 3.7 に、参加者の予測の絶対誤差と、統計分析の結果を示す。図 3.6 の数字の間の距離が、図 3.7 の誤差では 0.2 となる。Kruskal-Wallis 検定の結果、4 条件の間に有意な差があることが分かった ($p < .001$)。マン・ホイットニーの U 検定に Bonferroni の補正を施す方法で、多重比較による事後検定を行った。結果、「期待されるエージェント」モデル-*optimal* を除く全ての組み合わせで差が有意であった。

「期待されるエージェント」モデルによる予告と *optimal* による予告ははどちらも、ベースライン条件と比べて誤差を小さく抑えることができた ($p < .001$)。誤差の平均値は、「期待されるエージェント」モデルで 0.172、*optimal* で 0.176、ベースライン条件で 0.260 であった。「期待されるエージェント」モデルとベースラインの間の効果量 r は.25、*optimal*

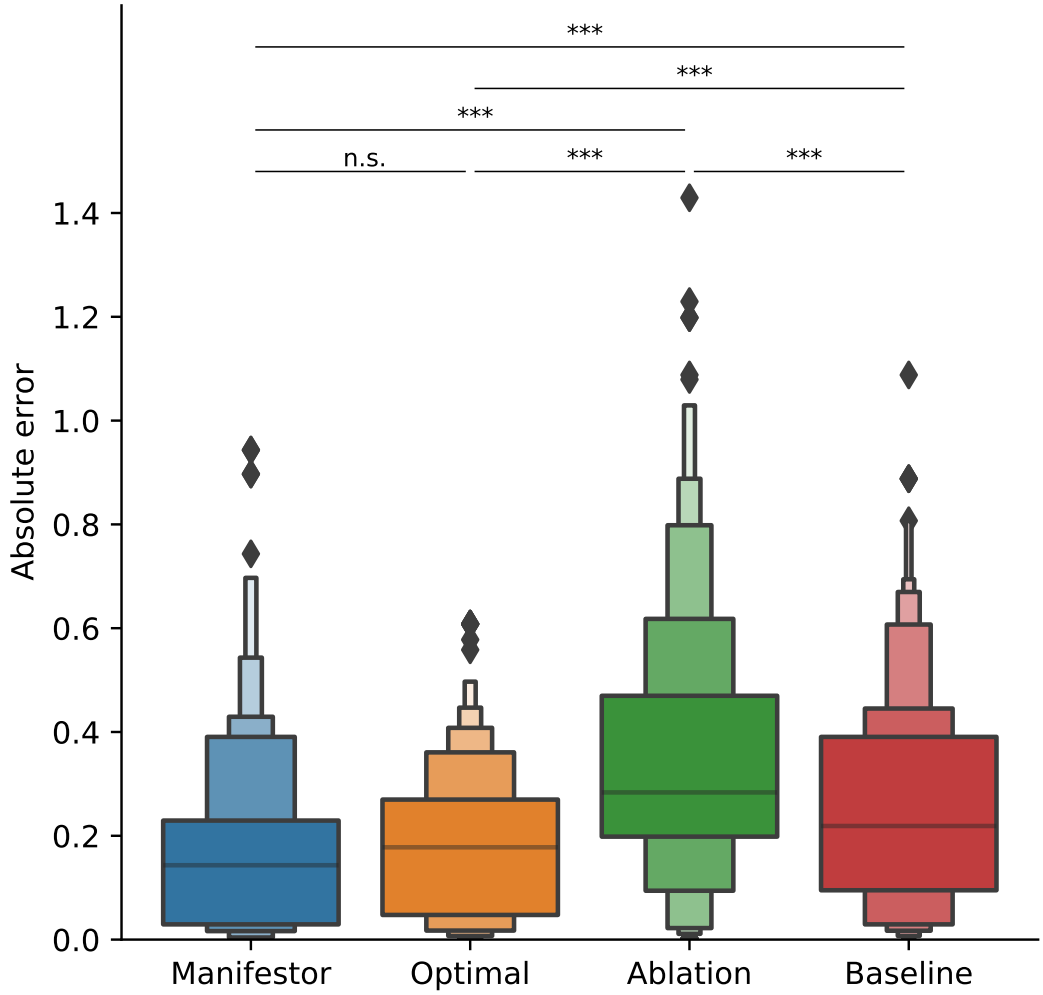


図 3.7: 実験参加者の予測の絶対誤差 [21]

とベースラインの間で .22 であった。「期待されるエージェント」モデルと *optimal* の間には有意な差が確認されなかった（$p = .25$）。両者の間の r は、.06 であった。以上の結果は仮説 H1 を支持している。つまり、「期待される」モデルは、学習時に目標の非対称性があったにも関わらず、目的の非対称性を考慮する必要のない *optimal* と同等の結果を得ることが出来ている。

ablation による予告は参加者の予測誤差を減らすことはできず、むしろ参加者をミスリードした。*ablation* の場合の誤差の平均値は 0.346 であった。ベースライン条件と *ablation* の間の r は.16、「期待されるエージェント」モデルとの間では.37 であった。以上の結果は仮説 H2 を支持している。つまり、目標の非対称性が存在する場面で、従来の指示遂行研究のように $g_{agent} = g_{human}$ を仮定する L_{f_N} では語彙を正しく解釈することができないこと、逆に、「期待されるエージェント」モデルが語彙を正しく解釈できたのは、$L^{+}_{f_N}$ が目標の非対称性を適切に扱っているためであることを示唆している。

3.8 限界と今後の展望

ここまでの実験で、「期待されるエージェント」モデルが人とエージェントの間の目標の非対称性を適切に扱うことで、人の指示の解釈と指示の背後にある人の目標を推定できるようになること、解釈した指示の語彙を流用することで、AI エージェントの動きを人に予告できるようになることを示した。しかしながら、モデルの実装や実験設定は、式 3.2-3.4 に表現されるアイデアの検証に焦点を当てており、「期待されるエージェント」モデルを実際の人・AI エージェント間のインタラクションに応用するためには、さらなる検討が必要である。

例えば、人の指示はルールベースによる定式化を行った（式 3.1）。しかし、人の指示を AI エージェントの行動学習に応用するインタラクティブ強化学習（IRL）では、不規則、一貫性の欠如、準最適などといった人のフィードバックシグナルの特徴が指摘されている [23] IRL 分野では人間のフィードバックの特徴を扱う手法も検討されており、こうした知見を活かすことは「期待されるエージェント」モデルの実際の応用に向けて有望だと期待される。

3.2.2 節でも言及した通り、本研究では指示の解釈とその背後にある目標の推定の整合性を取るというアイデアの有効性を検証するため、指示者からは、エージェントの動きが指示に従っていたか、というようなフィードバックが手に入らない設定を採用した。一方で、フィードバックの活用と「期待されるエージェント」モデルは相補的である。フィードバックは、「期待されるエージェント」モデルの指示者モデル M の学習を加速させる。そして M の学習は、人から必要なフィードバックの量を減らすことができる。実際のインタラクションでは、M の学習が進んでいない初期の段階に人からのフィードバックを活用し、妥当な M がある程度構築出来てきた段階で人に求めるフィードバックを減らすことで、より効率的に「期待されるエージェント」モデルを構築できると考えられる。

文脈の情報も、学習に貢献すると考えられる。例えば、本研究の設定では g_{human} はエピソード毎にランダムに決めた。しかし、実際の人の目標を考えると、g_{human} は周囲の状況やその人のそれまでの行動といった文脈と関わりがあることも多く、M の学習に有益な情報を持っていると考えられる。また、指示の解釈にも、その指示が発せられた文脈の情報は有用であろう。

「期待されるエージェント」モデルの改良として最も有望な方法の一つは、語彙の解釈における意味論や辞書、コーパスの活用である。「期待されるエージェント」モデルでは語彙に関する事前知識がゼロの状態から学習を始めている。辞書的意味の情報を「期待されるエージェント」モデルと組み合わせることで、f_N の学習コストを削減できると期待できる。

「期待されるエージェント」モデルの実装では、人間の選びうる目標の集合 $\mathcal{G}(\ni g_{human})$ が既知であり、またエージェントの目標の集合 g_{agent} と同じであるという仮定がある。これは、「期待されるエージェント」モデルの評価器 E を実装するうえで必要だった仮定である。しかし、AI エージェントの設計を十分に知らないユーザは、エージェントの取りうる目標以外をエージェントに求める場面も想定できる。逆強化学習 [5] を人に対して用

い、人の行動から g_{human} を推定できるようにすることで、この仮定を緩和することができるかもしれない。

「期待されるエージェント」モデルによる動きの予告は、予測器 T_N に依存している。しかし、環境の遷移を予測する問題は現在も活発に研究されるテーマであり、特に不確実性が高い実世界の予測は依然としてチャレンジングである。この分野で主要なタスクの 1 つは、与えられた動画の先を予想する video prediction である [67, 65, 46]。しかし、本研究の設定では AI エージェントの行動が環境に影響を与えるので、行動条件つき (action-conditioned) な手法が有望である [44]。環境の動態モデルを AI エージェントの意思決定に活かそうとするモデルベース強化学習 [41] の知見もまた、T_N の構築に活かせるであろう。T_N の精度は予測の長さ N に依存しており、説明する動きが長期化することで精度の面で予想の限界が生じると考えられる。「期待されるエージェント」モデルの構成が、どれくらいの長さの動きまで有効なのかは、引き続き検討する必要がある。

3.9 まとめ

本章では、人と AI エージェントが目標を共有できていないという目標の非対称性が存在する場面における、人とエージェントの指示を介したコミュニケーションに着目した。そして、人からエージェントに与えられる指示の背後にある人の目標の推測と、人の目標を考慮した指示語彙の解釈の相互依存関係を解決する「期待されるエージェント」モデルを提案した。「期待されるエージェント」モデルによって、目標の非対称性が存在する場面でも人から与えられる指示語彙を適切に理解できるようになることが分かった。また、「期待されるエージェント」モデルが学習した指示語彙を、エージェントの行動の説明に流用することで、エージェントが見せようとしている動きを人が精度よく理解できるようになった。

4 章

「推測されるエージェント」モデル

4.1 モデルの概要と観測の非対称性

「推測されるエージェント」モデルは、エージェントの動きを見た人がエージェントに対して推測する心的状態を、エージェント自体が推測する、という 2 次の推測をモデル化したものである [20, 19]。「推測されるエージェント」モデルによって、人がエージェントの目標をどのように考えているか推測したり、エージェントの動きが人に誤解を与えている状況を検知することができる。また、「推測されるエージェント」モデルの推測を応用することで、動きによってエージェントの目標を予告する表意動作を生成できるようになる。

「推測されるエージェント」モデルの最も重要な特徴は、人とエージェントの間の**観測の非対称性**が考慮されている点である。ここで、観測の非対称性とは、人とエージェントが異なる視点を持つために生じる観測の違いである。

また本章では、人とエージェントの間の観測の非対称性が生じる場面における「推測されるエージェント」モデルの推定や、モデルによって生成される表意動作を検証し、結果を報告する。実験の結果、モデルが観測の非対称性を正しく考慮できており、それによって非対称性を考慮しない場合よりも効果的な表意動作を生成できることが示された。

4.2 客体的自己認識と予告

本章の研究は、客体的自己認識という心理学における概念に着想を得ている。客体的自己認識（objective self-awareness）とは、人が自分自身を注意の対象として認識する能力である。客体的自己認識は、私的（private）自己認識と公的（public）自己認識の 2 つの側面に分けられる [15, 16]。私的自己認識とは、自己の感覚や感情、思考を内省する能力のことであり、公的自己認識は、自己の外見や振る舞いといった外面に対する注意を指す。

AI エージェントを主体と考えたとき、エージェントにとっての私的・公的自己認識は表 4.1 で定式化することができる。公的自己認識 $f_a(f_h(x_a))$ は、エージェントの挙動を見た人が帰属するエージェントの心的状態を、エージェントの側から推測したものである。また、私的自己認識 $f_a(x_a)$ は、エージェントの意思決定に、目標や意図といった心的状態を当てはめた結果と考えられる。エージェントの私的自己認識は、エージェントの意思

表 4.1: 私的・公的自己認識の数式的表現

私的自己認識	$f_a(x_a)$
公的自己認識	$f_a(f_h(x_a))$

決定モデルが XAI における本来的アプローチで構築されている場合、モデルが明示的に持つ変数と対応する。エージェントの意思決定がブラックボックス化している場合は、事後的アプローチによって推測することになる。

エージェントに対する人の誤解や無理解を、$f_a(x_a)$ と $f_a(f_h(x_a))$ の間の差異、この差異を最小化する予告 u が、効果的な予告であるとする。

$$\mathrm{argmin}_u |f_a(x_a) - f_a(f_h(x_a))| \tag{4.1}$$

4.3 Bayesian Theory of Mind (BToM)

Bayesian Theory of Mind (BToM) は、「推測されるエージェント」モデルの基礎となるモデルである [6]。BToM は、人がエージェントの動きをもとに、エージェントの信念や欲求といった心的状態を推測する 1 次の推測の過程をベイズ推論としてモデル化している。ここからは、行動を行うエージェントを行為者、エージェントの行動を観察する人を観察者と呼ぶ。BToM は、観察者が行為者の心的状態を推定する過程のモデル化と言い換えられる。BToM は式 4.2 によって表現される。

$$\begin{aligned} & P(b^1_{t+1}, d^1) \\ \propto & \sum_{\substack{o^1_{t+1}, s_{t+1}, \\ s_t, a_t, b^1_t}} P(b^1_{t+1}|b^1_t, o^1_{t+1}) \cdot P(o^1_{t+1}|s_{t+1}) \cdot P(o_{t+1}|s_{t+1}) \cdot P(s_{t+1}|s_t, a_t) \\ & \qquad \cdot P(a_t|b^1_t, d^1) \cdot P(b^1_t, d^1, |o_{:t}), \end{aligned} \tag{4.2}$$

ここで、$b_t = P(s|o_{:t})$ は過去の観測 $o_{:t} = (o_0, o_1, ..., o_t)$ が与えられた際に、時刻 t での環境の状態が s である確率の分布である。また、d は行為者の目標である。図 2.2 の例では、d はリンゴとナシのいずれかである。本章において心的状態 x は、欲求と信念の組 (b_t, d) で表現される。変数の添え字の 1 は、それが行為者の心的状態に関する観察者の 1 次の推測の結果であることを示す。

BToM は、観察者の観測 o_t をもとにした環境の状態 s_t の推測、そこから推測される行為者の観測 o^1_t、さらに o^1_t をもとに構築される行為者の信念 b^1_t という、観測の非対称性がモデルに組み込まれている。

4.4 「推測されるエージェント」モデル

「推測されるエージェント」モデルは、観察者が行為者の心的状態を推測する 1 次の推測の過程をモデル化した BToM を拡張したもので、行為者の心的状態が観察者にどのよ

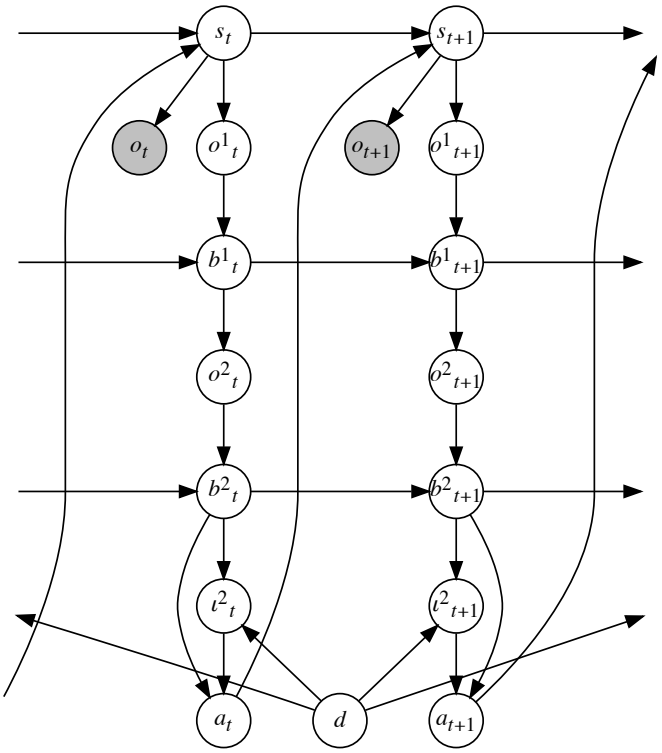


図 4.1: 「推測されるエージェント」モデルのグラフィカルモデル

うに推測されるかを行為者自身が推測する、2 次の推測を行う。BToM と異なり、推測の主体は行為者である AI エージェントである。モデルは以下の式で表現される。また、この式をグラフィカルモデルとして表現したものを図 4.1 に示す。

$$\begin{aligned} & P(b^2_{t+1}, d^2) \\ \propto & \sum_{\substack{o^2_{t+1}, o^1_{t+1}, b^1_{t+1}, \\ s_{t+1}, s_t, a_t, \\ b^2_t, b^1_t}} P(b^2_{t+1}|b^2_t, o^2_{t+1}) \cdot P(o^2_{t+1}|b^1_{t+1}) \cdot P(b^1_{t+1}|b^1_t, o^1_{t+1}) \\ & \cdot P(o^1_{t+1}|s_{t+1}) \cdot P(o_{t+1}|s_{t+1}) \cdot P(s_{t+1}|s_t, a_t) \cdot P(a_t|b^2_t, d^2) \cdot P(b^2_t, d^2, |o_{:t}). \end{aligned} \tag{4.3}$$

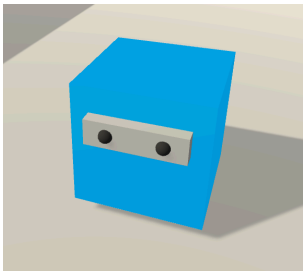
添え字の 2 は、変数が行為者の心的状態に関する 2 次の推測の結果であることを示す。「推測されるエージェント」モデルは、行為者の実際の観測、観察者に対する 1 次の推測、観察者が行為者に帰属する心的状態の 2 次の推測を切り分けて考えることで、観察者と行為者の間にある観測の非対称性を考慮した推測を可能にしている。

4.5 実装

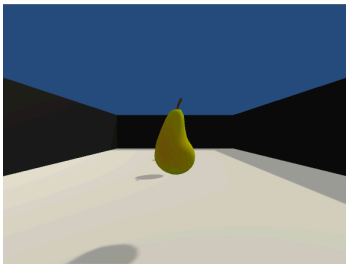
4.5.1 環境と AI エージェント

シミュレーション空間と、行為者である AI エージェントを用意した。空間にはリンゴとナシがあり、エージェントはそのいずれかに向かって進む。観察者である人は固定された視点から行為者の動きを観察し、行為者がリンゴとナシのどちらに向かっているかを推測する。

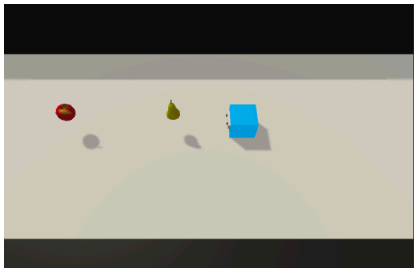
(a) 環境全体



(b) AI エージェント



(c) AI エージェントの視点から見た環境



(d) 人の視点から見た環境

図 4.2: 環境 A

「推測されるエージェント」モデルの実装と検証にあたり、2 種類の環境（環境 A、環境 B）を用意した。どちらも、AI エージェントの行動の学習には、深層強化学習の手法である Asynchronous Advantage Actor-Critic (A3C)[39] を用いた。$\mathcal{A}$ は、前へ加速、時計回りに 22.5 度転換、反時計回りに 22.5 度転換、慣性に従ってそのまま進む、の 4 種類からなる。環境の状態 s は、リンゴとナシの位置、観察者の視界、行為者の状態からなる。行為者の状態は、エージェントの位置、速度、向きと、行為者の視界の範囲からなる。観察者は初期地点から動くことはできず、固定された視界から環境を観測する。つまり、観察者は、リンゴとナシが視界にあるときのみその位置を把握することができる。また観察者、行為者が見えているときは、その位置、向き、速度を把握することができる。本研究では、観察者の見える範囲は行為者にとって既知とする。

環境 A は、横長の空間にリンゴとナシが 2 つずつ存在する（図 4.2a）。行為者（図 4.2b）は一人称視点から得られる環境の情報（図 4.2c）をもとに、リンゴとナシのいずれかに向かうよう行動を学習している。観察者は、環境の中央部を見下ろす形で観測する。環境 A は、「推測されるエージェント」モデルの推測結果を検証するケーススタディに用いた（4.6 節）。

図 4.3 に環境 B を示す。環境 B は、「推測されるエージェント」モデルによって生成される表意動作を検証する際に用いた（4.6 節）。環境 B では、リンゴとナシはランダムな

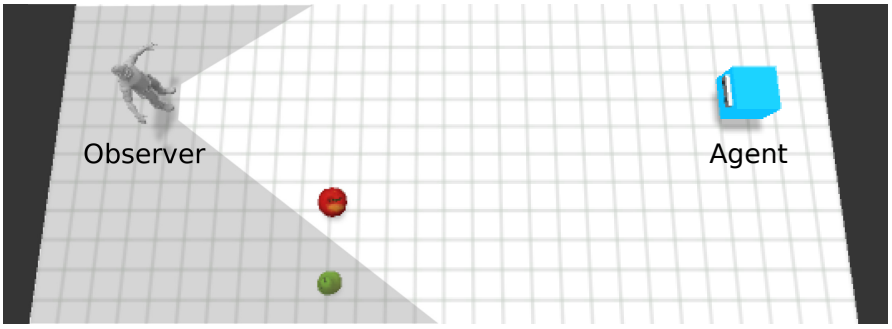


図 4.3: 環境 B
明るい部分が人の観測可能な領域。

位置に 1 つずつ配置され、人はエージェントと同じ空間に立ってエージェントの動きを観測する。人とエージェントの初期位置は、図 4.3 の位置で固定されている。エージェントは、0.5 秒ごとに行動空間 $\mathcal{A}$ から行動を選択する。

4.5.2 「推測されるエージェント」モデル

アルゴリズム 2 に、式 4.3 がどのように計算されるかを示す。s_t は連続値で表現され、また部分観測による不確実性を持つため、観察者や行為者の信念分布を考えるうえで考慮すべき s が無数に存在する。「推測されるエージェント」モデルは、逐次モンテカルロ法を参考にしたランダムサンプリングを行うことで、有限のサンプルによって信念の分布を近似する。時刻 0 において、環境の状態に関するサンプル $S_0 = s_{0,0}, s_{0,1}, ..., s_{0,n-1}$ が抽出される。

「推測されるエージェント」モデルは、2.2 節で取り上げた SDS フィルタをもとに、入れ子構造の推定を行う。モデルは 4 種類のフィルタを持つ:

$$
\begin{aligned}
\Phi^0_t(s_{:t}) &\propto b^0(s_{:t}), \\
\Phi^1_t(s'_{:t}, s_{:t}) &\propto \sum_{b^1_t} b^1_t(s'_{:t}) \cdot \delta(b^1_t, B^{obs}(s_{:t})), \\
\Phi^2_t(s'_{:t}, s_{:t}) &\propto \sum_{b^2_t} b^2_t(s'_{:t}) \cdot \delta(b^2_t, B^{act}(s_{:t})), \\
\Psi^2_t(d^2, s_{:t}) &\propto P(d^2 | s_{:t}).
\end{aligned}
$$

関数 $\delta(\alpha, \beta)$ は、$\alpha = \beta$ の際に 1 を、それ以外で 0 を返す。$B^{obs}(s_{:t})$ と $B^{act}(s_{:t})$ はそれぞれ、環境の状態の系列 $s_{:t}$ が所与の際の観察者と行為者の信念を返す。ここからは環境の状態遷移のマルコフ性を仮定し、$s_{:t}$ を s_t と単純化して表記する。$\Phi^0_t(s_t)$ は、行為者の環境に関する信念を表す。$\Phi^1_t(s'_t, s_t)$ は、環境の状態が s_t であると仮定した際に、観察者が環境の状態を s'_t と信じる確率を、$\Phi^2_t(s'_t, s_t)$ は、s_t を仮定した際に、行為者が環境の状態を s'_t と信じる確率を表現する。最後に、$\Psi^2_t(d^2, s_t)$ は、環境の状態 s_t を仮定した際に、行為者の欲求が d^2 と推測される確率である。これらのフィルタが与えられたとき、人がエージェントに帰属する信念 b^2 において環境の状態が s とエージェントが信じている確

Algorithm 2 「推測されるエージェント」モデルの更新

Require: o_t: the actor's observation at time t

1: **if** $t = 0$ **then**
2: // Initialization
3: Sample n possible environmental states at time $t = 0$ $(s_{0,0}, s_{0,1}, ..., s_{0,n-1})$.
4: $S_0 \leftarrow \{s_{0,0}, s_{0,1}, ..., s_{0,n-1}\}$
5: Initialize filters $\Phi_0^0(s)$, $\Phi_0^1(s', s)$, and $\Phi_0^2(s', s)$ for each $s, s' \in S_0$ as uniform distributions.
6: **end if**
7: Update Φ_t^0, Φ_t^1, and Φ_t^2 based on partial observability.
8: Resample S_t.
9: $S_{t+1} \leftarrow \{Pred(s, a) | s \in S_t, a \in A\}$
10: Generate new filters Φ_{t+1}^0, Φ_{t+1}^1, and Φ_{t+1}^2 based on the previous ones at time t.
11: Calculate $P(a | s \in S_t, d^2, \iota^2)$ for each s, d^2, and ι^2.
12: Generate Ψ_{t+1}^2 based on Ψ_t^2 and $P(action | \iota, s \in S_t)$.
13: Add noise $s \in S_{t+1}$.

率は、以下で計算される。

$$b^2(s) = \sum_{s' \in S_t} \Phi^2(s', s) \sum_{s'' \in S_t} \Phi_t^1(s'', s') \cdot \Phi_t^0(s'').$$

また、人がエージェントに帰属する欲求は、以下の式で計算される。

$$P(d^2) = \sum_{s \in S_t} \Psi_t^2(d^2, s) \sum_{s' \in S_t} \Phi^2(s', s) \sum_{s'' \in S_t} \Phi_t^1(s'', s') \cdot \Phi_t^0(s'').$$

フィルタは、時刻 0 において一様分布に初期化される。$P(s_t | o_t) \propto P(s_t | o_{t-1}) \cdot P(o_t | s_t)$ であるため、行為者が o_t を観測すると、$\Phi_t^0(s_t)$ は観測確率 $P(o_t | s_t)$ を乗ずることで更新される。$\Phi_t^1(s'_t, s_t)$ と $\Phi_t^2(s'_t, s_t)$ も同様に、可能な s_t のそれぞれで観測者と行為者が観測 o^1、o^2 を得る確率 $P(o_t^1 | s'_t)$、$P(o_t^2 | s'_t)$ を乗じることで、更新される。

$$\Phi_t^0(s_t) \leftarrow \Phi_t^0(s_t) \cdot P(o_t | s_t), \tag{4.4}$$

$$\Phi_t^1(s'_t, s_t) \leftarrow \Phi_t^1(s'_t, s_t) \cdot \sum_{o_t^1} P(o_t^1 | s'_t) \cdot \delta(o_t^1, O^{obs}(s_t)), \tag{4.5}$$

$$\Phi_t^2(s'_t, s_t) \leftarrow \Phi_t^2(s'_t, s_t) \cdot \sum_{o_t^2} P(o_t^2 | s'_t) \cdot \delta(o_t^2, O^{act}(s_t)), \tag{4.6}$$

$O^{obs}(s_t)$ と $O^{act}(s_t)$ は、s_t を所与とした際の観測者と行為者の観測を出力する関数である。

更新の過程で、$s \in S_t s.t. \sum_{b^2} b^2(s) = 0$ となる s が出現する。すなわち、サンプルされた s のうち、行為者が環境の状態を s と信じている確率がもはや無い s が生じる。この際、サンプルからは s が取り除かれ、$\sum_{b^2} b^2(s)$ が高いサンプルが分岐する形でリサンプリングが行われる。

環境の状態は、行為者の選択する行動によって分岐する。ここで関数 $Pred : S_t \times A \to S_{t+1}$ を、環境の状態 s_t でエージェントが行動 a_t を選択した際の環境の遷移 s_{t+1} を推測する関数とする。本研究では $Pred$ を教師あり学習によって得ている。$Pred$ によって、環境の状態 $s \in S_t$ がエージェントの行動 a_t によって遷移した先の集合 $S_{t+1}^{a_t}$ を考えることができる。また、S_{t+1} を、$a_t \in \mathcal{A}$ 全てを考慮した際の時刻 $t+1$ における環境の状態 $\bigcup_a S_{t+1}^a$ とする。

時刻 $t+1$ におけるフィルタ Φ_{t+1} は、以下の式で得られる。

$$\Phi_{t+1}^0(Pred(s_t, a)) \leftarrow \Phi_t^0(s_t),$$
$$\Phi_{t+1}^1(Pred(s_t', a'), Pred(s_t, a)) \leftarrow \Phi_t^1(s_t', s_t) \cdot \delta(a', a),$$
$$\Phi_{t+1}^2(Pred(s_t', a'), Pred(s_t, a)) \leftarrow \Phi_t^2(s_t', s_t) \cdot \delta(a', a).$$

Ψ_t は、$P(a|s_t, d^2)$ を乗じることで更新される。

$$\Psi_{t+1}^2(d^2, Pred(s_t, a)) \leftarrow \Psi_t^2(s_t) \cdot P(a|s_t, d^2)$$

$P(a|s^2, d_a^2)$ は、一般的な強化学習によって獲得することができる。

4.6 ケーススタディによる検証

4.6.1 目的

「推測されるエージェント」モデルの推測する “エージェントに帰属される目標” と、人が観察者の視点でエージェントを観察した際に推測するエージェントの目標を比較することで、「推測されるエージェント」モデルの精度を評価する。

4.6.2 方法

実験のため、6 種類のエピソードを用意した。エピソードは、simple、blind、misleading の 3 通りに 2 種類ずつ分類される。simple エピソードでは、行為者は目標となる果物（リンゴ/ナシ）を即座に発見し、そちらの方向へ直行する。blind エピソードでは、行為者の向かう先の果物が観察者の視界外に存在する。観測の非対称性が特に問題になる場面である。misleading エピソードは、行為者が最初に特定の果物に向かっているように見えるが、実際はその果物を無視して別の方向へ進む場面を選んだ。こうした行為者の振る舞いは、行為者の視界に目標とする果物がなく、探索をしている際によく見られた。行為者の目標は、各エピソードでリンゴの場合とナシの場合が両方含まれるように設定した。

実験はウェブブラウザ上で行った (図 4.4)。実験参加者には、観察者の視点から見える行為者の動きを提示した。動きはコマ送りで提示され、各フレームで、エージェントの目標がリンゴとナシのどちらだと思うかを尋た。参加者は、ウェブページ上のスライドバーを操作することで回答を行った。エピソードの順番はユーザ毎にランダム化した。

実験参加者は、大学生と大学院生の合わせて 11 名である。うち 8 名が男性、3 名が女性であった。実験に先立って、参加者には環境の全体図と、その中で観察者が観測できる

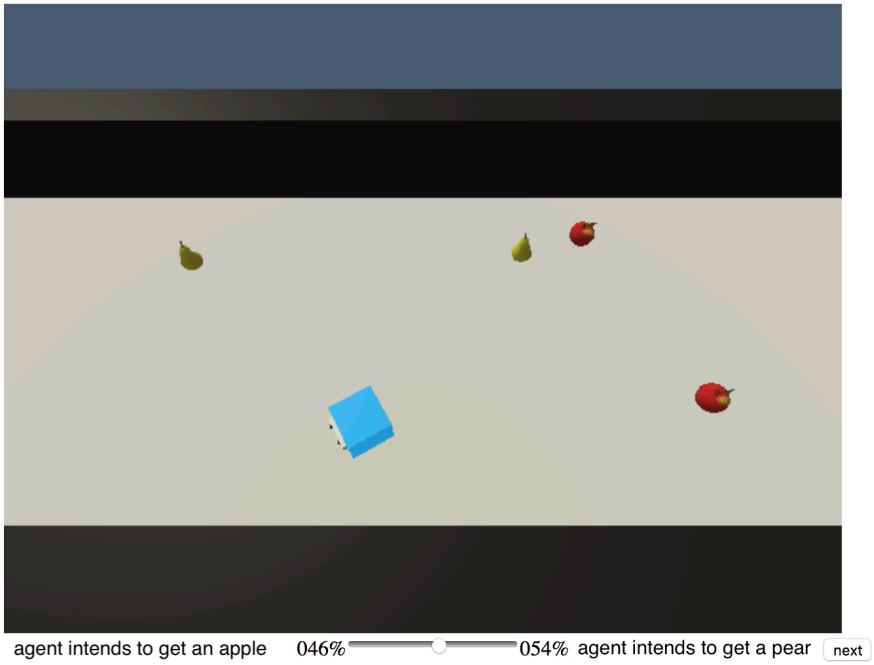


図 4.4: 実験に用いたユーザインタフェース
ユーザはスライドバーを操作することで、自己の推測する確率を回答した。

範囲、エージェントの視界から見た環境を説明した。また、参加者に以下の 4 点を教示として与えた。

1. いずれのエピソードでも、環境にはリンゴとナシが 2 つずつと、行為者が存在する。
2. エージェントが獲得しようとする果物はリンゴかナシの 2 択で、エピソード開始時点でいずれかに決定し、以降変わらない。
3. エピソード開始時点で、エージェントはリンゴとナシがどこにあるか知らない。
4. 果物とエージェントの初期位置はエピソード毎にランダムに決定されていて、エージェントの目標とは無関係である。

教示 4 は、エージェントの目標に関する初期のバイアスを取り除くために与えた。教示 4 にも関わらず、2 人の参加者はエージェントの初期位置とエージェントの目標を対応づけて回答した。これは例えば、エージェントの初期位置がリンゴの目の前であっただけで、エージェントの目標がリンゴである確率を高く見積もった。このように、エージェントの行動だけでなく、周囲の環境の状態といった文脈情報もエージェントへの目標帰属に影響を与える。しかし本研究では、エージェントの行動が目標帰属に与える影響に注目するため、当該の参加者の結果はその後の分析から除外した。

4.6.3 結果と考察

図 4.5 に、simple エピソードにおける「推測されるエージェント」モデルの推測結果と実験参加者の推測結果を示す。両者には有意な強い相関 ($r = 0.90$) が見られた。この結果

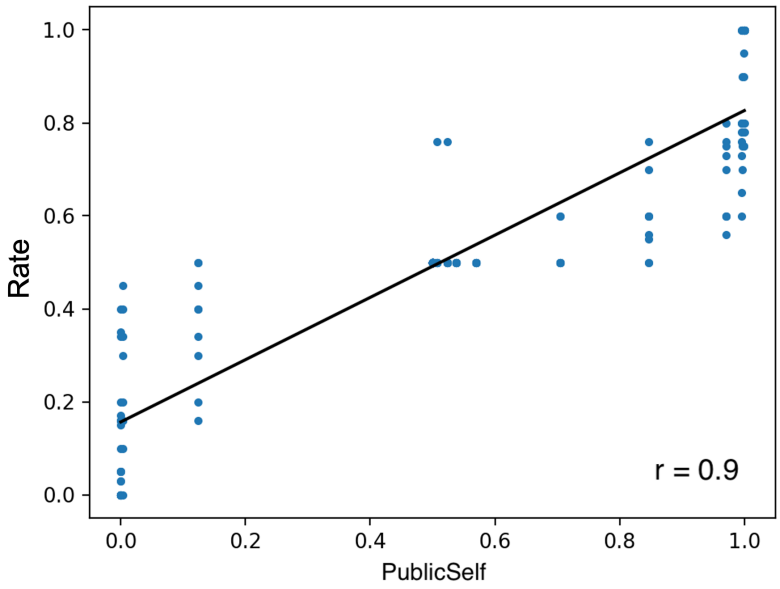


図 4.5: simple エピソードにおいて、「推測されるエージェント」モデルと実験参加者が推測した、エージェントの目標がリンゴである確率

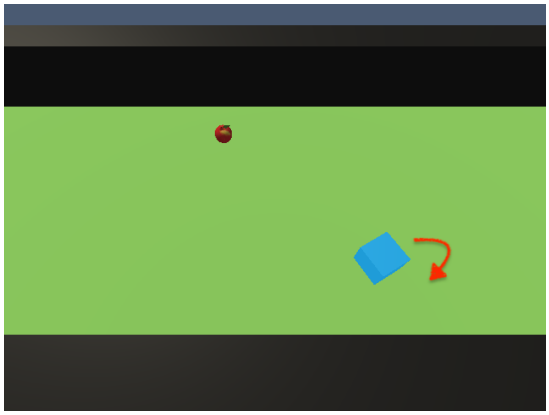
は、「推測されるエージェント」モデルが、人がエージェントに帰属する目標をエージェントの視点から適切に推定できていることを示唆する。

blind エピソードでは、「推測されるエージェント」モデルは、エージェントに帰属される目標が一様だと推測した（図 4.6）。エージェントの視界には目標とする果物が存在し、エージェントは目標に向かって進んでいるにもかかわらず、そのことがモデルの推測結果に影響を与えなかったのは、人とエージェントの間にある観測の非対称性をモデルが適切に考慮できている結果といえる。

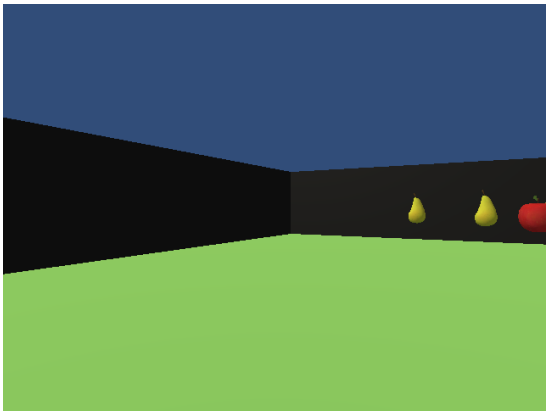
図 4.6 に示すエピソードにおいて、「推測されるエージェント」モデルの推測通り一様な確率を回答し続けた参加者が 6 名いた。一方で、エージェントの目標がナシであると推定する確率を高く見積もった参加者も 3 名おり、彼らはエージェントが画面の右端に向かった動きを見てナシの確率を高めた。インタビューの結果、この推測は以下の思考にもとづいていることが分かった。

- 環境にはリンゴとナシが 2 つずつ存在する。
- 参加者の視点からはリンゴが既に 1 つ見えているが、ナシは見えていない。
- そのため、エージェントが向かった環境の右端には、リンゴがある確率よりもナシがある確率の方が高い。
- よって、エージェントが向かった先にはナシがある確率が高い。

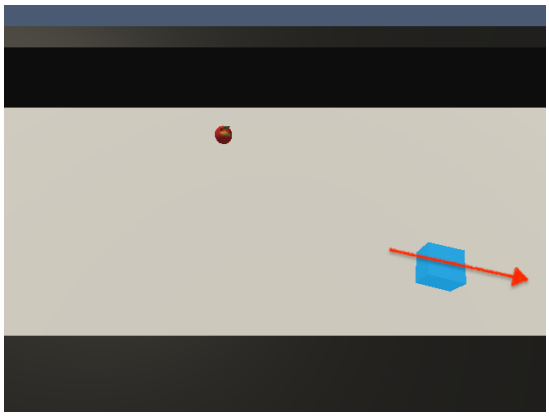
以上の説明は論理的に正しい。しかし、このエピソードでは、参加者の観測できる範囲に



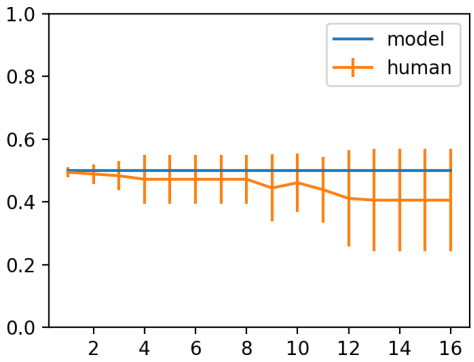
(a) $t = 1$



(b) $t = 1$ (actor's perspective)



(c) $t = 10$



(d) probability

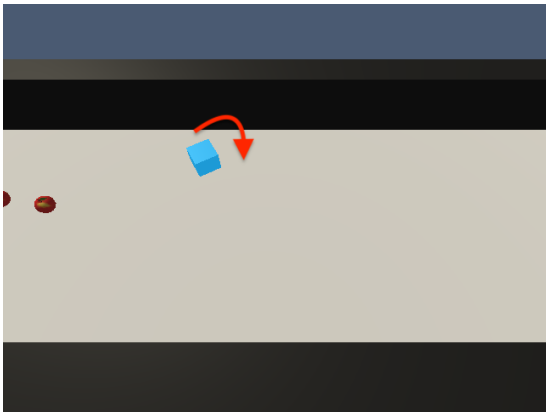
図 4.6: blind エピソードの結果

リンゴが存在することをエージェントが観測していないため、「推測されるエージェント」モデルがエージェントの視点から参加者の思考を認識することは出来なかった。

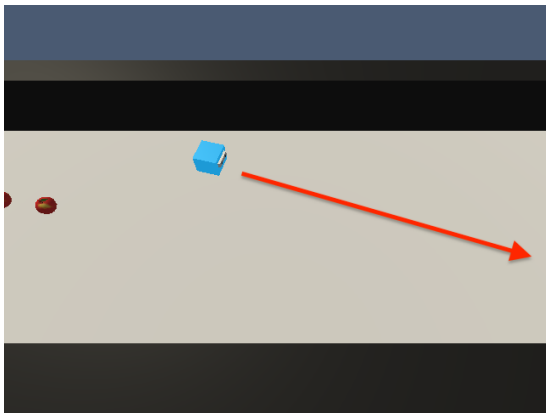
図 4.7、4.8 に、misleading エピソードにおける結果を示す。misleading エピソードにおける実験参加者の推測と「推測されるエージェント」モデルの推測の間には有意な正の相関があったものの、相関の程度は弱かった ($r = 0.46$)。参加者にアンケートを実施したところ、参加者の推測とモデルの推測の間に差が生じた理由として 2 つの可能性が見つかった。

考えられる理由の 1 つは、misleading エピソードにおいて「推測されるエージェント」モデルが参加者の信念の更新を適切に追えていない可能性である。今回の実装では、4.4 の $P(o|s)$ の実装によって、果物が視界に入った時点でその位置に果物があるという信念がほぼ 100%になるよう更新がなされ、それ以外の可能性は切り捨てられた。また、一度切り捨てた可能性を再度考慮する機構が無かった。しかし参加者は、エージェントが果物を通り過ぎたり方向を変えたりした際に、「エージェントが果物を見失った」や、「エージェントには果物が見えていなかった」と解釈し、それまでの推定でほとんど切り捨てていたであろう可能性を新たに考え始めた。このように、信念の推定は視界だけでなく振る舞いからも更新される。そして、それまでの推定に矛盾が生じた場合、人は振る舞いを説明する新たな可能性を考慮に加える。misleading のように、推定の裏切りが発生する場面

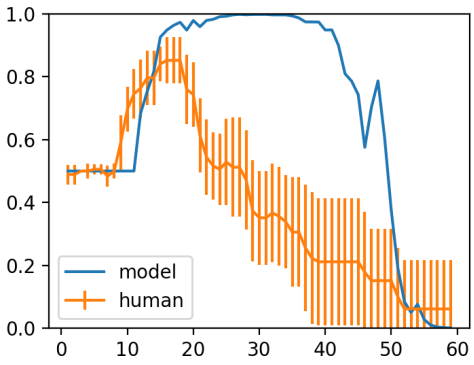
(a) $t = 1$



(b) $t = 32$

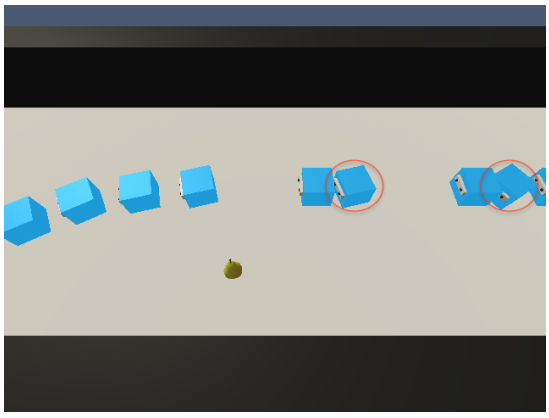


(c) $t = 48$

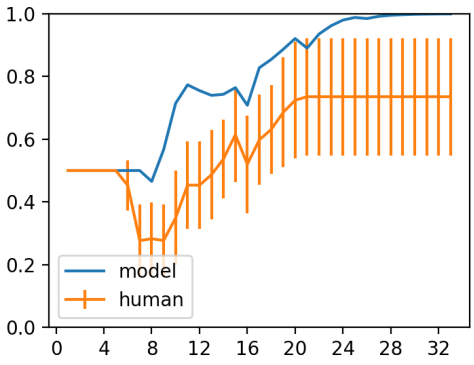


(d) 予測された確率

図 4.7: misleading エピソードの結果 1



(a) エージェントの動き



(b) 予測された確率

図 4.8: misleading エピソードの結果 2

では、確率が低いと一度見積もられた可能性を再び検討したり、新たな可能性を考慮の対象に追加する機構が必要になると考えられる。

もう 1 つの理由は、エージェントのこうした振る舞いを見て人は、エージェントが目標を持って行動しているという意図性に疑問を持った可能性である。misleading エピソードでは、時計回りの転回と反時計回りの転回を繰り返す振動が見られた。これは、エージェントの行動を決定する意思決定器の意思決定が確率的で、特にエージェントが目標を見つけられていないときに行動の選択確率が平準化することに起因している。人は、こうした動きから合理性や一貫性を見出せなくなり、エージェントに意図スタンスを採用して目標を推測するというタスクを諦めた可能性がある。アンケートで参加者は、エージェントの振動に対してネガティブな印象を受けたことや、リンゴとナシという 2 択から回答することに困難を感じたと報告した。1.3 に挙げたように、行動の合理性は意図スタンスの採用に影響を与える。参加者は、エージェントがリンゴかナシのいずれかに向かうことを目標に行動しているという前提に疑問を抱き、どちらでもないということを 50%という回答で表現した可能性がある。本研究では確率という指標だけで評価したが、実際の応用では推測の確信度や、エージェントに採用される意図スタンスの強度を考慮することで、エージェントの「推測されるエージェント」をより精緻にモデル化することができるだろう。

4.6.4 ケーススタディのまとめ

「推測されるエージェント」モデルが、エージェントが目標に向かってまっすぐに進む単純な場面で、人がエージェントに帰属する目標を正しく推測できることが確認された。また、観測の非対称性が存在する場面で、人がエージェントの目標を判断できない状況を適切に捉えられることがわかった。エージェントの行動に一貫性が欠ける場面や行動に矛盾を見出せる場面では、エージェントに帰属される目標を一定の精度で推測することができたものの、一度切り捨てた可能性を再度考慮する機構や、エージェントの行動を説明する新たな可能性をサンプルに追加する必要性が示唆された。

4.7 表意動作の生成

「推測されるエージェント」モデルの推測を応用することで、表意動作を生成する手法を提案する。表意動作は、エージェントの目標を伝達する動きである。提案手法は、観測の非対称性を考慮するという「推測されるエージェント」モデルの特徴によって、人とエージェントの間に観測の非対称性がある場面でも適切な表意動作を生成することができる。

表意動作を生成する手法の概要は以下の通りである：式 4.3 で行為者が特定の行動 a_t を取った際の確率を足し合わせることで、観察者が特定の心的状態 (b^2, d^2) を行為者に帰属する確率 $P(b^2, d^2|o, a)$ を計算できる。この確率を最大化する行動が、エージェントの心的状態を伝達する行動である。

$$a = \mathbf{argmax}_a P(b^2, d^2|o, a). \tag{4.7}$$

特に d^2 に関する周辺確率を最大化するものが、行為者の目標を最も効果的に伝達する行動といえる。

Algorithm 3 表意動作の生成

Require: ι^*: The actor's true intention; a^*: The action chosen under the actor's original policy; s: The state at the time; $\pi(a \mid s)$: The probability distribution of the actions a that the actor takes from the state s.

σ: Array of floats.
Calculate $P(\iota^2 = \iota^* | action)$ for each action with the 「推測されるエージェント」モデル model.
for all *action* **do**
  if $\pi(action \mid s) > threshold_1$ **then**
    $\sigma[action] \leftarrow P(\iota^2 = \iota^* | action)$
  else
    $\sigma[action] \leftarrow 0$ // do not choose the action
  end if
end for
$a \leftarrow \mathrm{argmax}_{action}\sigma[action]$
if $\sigma[a]/\sigma[a^*] > threshold_2$ **then**
  return a
else
  return a^*
end if

アルゴリズム 3 に、「推測されるエージェント」モデルで表意動作を生成する処理を示す。実際の処理では、式 4.7 を最大化だけでなく、その行動によって伝達できる情報の効率を考慮する。具体的には、まずエージェントの元々の π が選択し得ない行動が避けられる。さらに、式 4.7 を最大化する行動を選択することで生じる、エージェントに d^2 が帰属される確率の変化を、π によって選ばれる行動を選択した場合の確率の変化と比較する。後者の変化と比較して前者の変化が一定以上無い場合は、π が選択する行動をそのまま採用する。これによって、表意動作を生成する効果とコストのバランスを取っている。例えば図 2.1 において、バランスを考慮せずに式 4.7 の最大化のみを行うと、エージェントのカーブが大きくなって元々の目標を達成するのに時間がかかったり、そもそも目標までたどり着けなくなる。

4.7.1 FalseProjective 表意動作

観測の非対称性が表意動作に与える影響を調査するため、「推測されるエージェント」モデルによる動きと比較するための手法として、観測の非対称性を考慮しない FalseProjective 表意動作を用意した。FalseProjective では、行為者自体の信念が観察者にも共有されている

表 4.2: 各シナリオの設定

	行為者の目標（リンゴ）が観察者から観測可能	目標でない果物（ナシ）が観察者から観測可能
Center	True	True
Side-visible	True	False
Side-invisible	False	True
Blind-inside	False	False
Blind-outside	False	False

という前提で表意動作を生成する。FalseProjective の名前は、行為者の信念が観測者に**誤って投射されている**ことに由来する。「推測されるエージェント」モデルが FalseProjective よりも良い表意動作を生成できていれば、表意動作の生成において観測の非対称性が重要な要因であり、モデルがそれを適切に扱えていることを意味する。

「推測されるエージェント」モデルはエージェントの信念 b^0 を計算しており、False-Projective 表意動作は b^0 をもとに生成できる。FalseProjective 表意動作のために、「推測されるエージェント」モデルにフィルタ $\Psi_t^0(s, d^0)$ を追加した。$\Psi_t^0(s, d^0)$ は、行為者の信念を観察者も共有しているという前提でエージェントの目標 d^0 を推定する。$P(d^0)$ は以下の式で計算される。

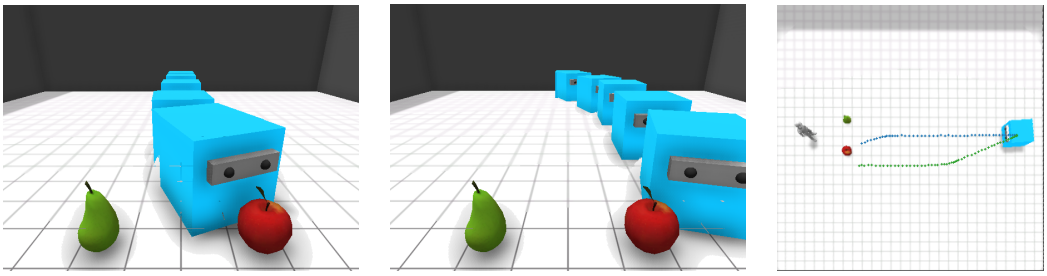
$$P(d^2) = \sum_{s \in S_t} \Psi_t^2(d^2, s) \cdot \Phi_t^0(s).$$

4.8 生成された表意動作

リンゴとナシの位置関係の異なるシナリオにおいて「推測されるエージェント」モデルに生成された表意動作の例を示す。また、比較対象として、深層強化学習によって最短経路を選ぶように学習させた場合の動き（original motion）と、FalseProjective 表意動作も示す。用意したシナリオは、Center、 Side-visible、 Side-invisible、 Blind-inside、 Blind-outside の 5 種類である。表 4.2 に、各シナリオの設定を示す。

例では、行為者の目標となる果物はリンゴに統一している。各シナリオは、行為者の目標であるリンゴと目標でないナシが、それぞれ観測者の視界内にあるかどうかの $2 \times 2 = 4$ 通りの可能性を網羅している。また、全てのシナリオで、リンゴとナシは隣接している。これは、観測者から見てエージェントの目標がすぐに明らかにならない場面や、誤解が生じやすい場面を考えるためである。全ての例で行為者の初期位置からリンゴとナシの両方は観測可能である。つまり、行為者が目標を見つけるための探索が必要ない状況である。

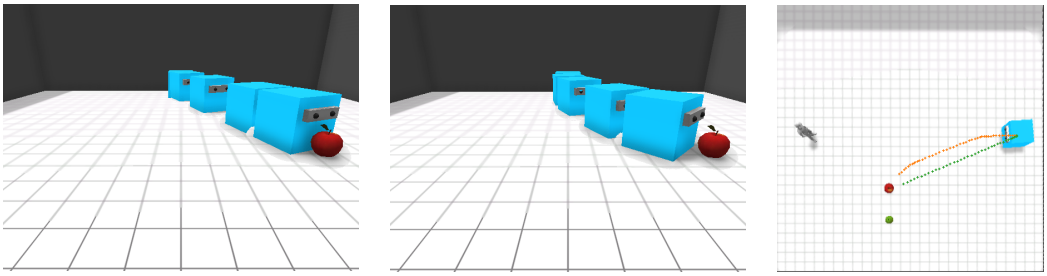
図 4.9 に、Center エピソードの例を示す。Center エピソードでは、リンゴとナシが観察者の目の前に配置されている。最短経路を選ぼうとする original の動きでは、果物が行為者の目前に迫るまで行為者は直線的な動きを見せており、観察者は終盤に行為者がリンゴの方向を向くまで、目標を判定できない (図 4.9a)。一方、FalseProjectve と「推測され



(a) Original　(b) FalseProjective / 「推測されるエージェント」モデル　(c) 鳥観図

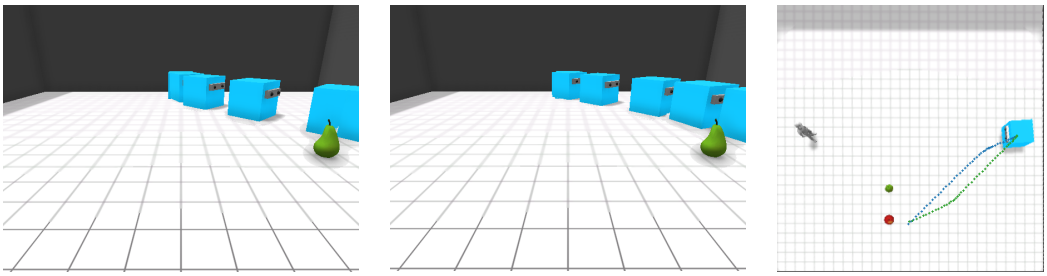
図 4.9: Center シナリオで生成された動き

Center シナリオで生成された動き。青: Original。オレンジ: FalseProjective。緑: 「推測されるエージェント」モデル。



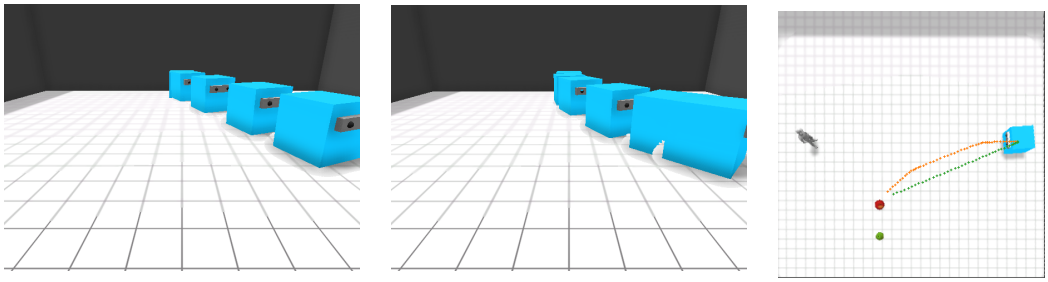
(a) Original / 「推測されるエージェント」モデル　(b) FalseProjective　(c) 鳥観図

図 4.10: Side-visible シナリオで生成された動き



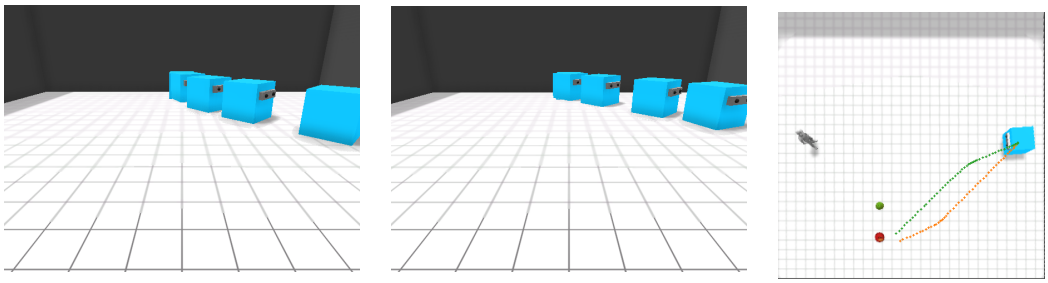
(a) Original　(b) FalseProjective / 「推測されるエージェント」モデル　(c) 鳥観図

図 4.11: Side-invisible シナリオで生成された動き



(a) Original / 「推測されるエージェント」モデル　(b) FalseProjective　(c) 鳥観図

図 4.12: Blind-inside シナリオで生成された動き



(a) Original / 「推測されるエージェント」モデル　(b) FalseProjective　(c) 鳥観図

図 4.13: Blind-outside シナリオで生成された動き

るエージェント」モデルによる表意動作は、行為者が初期からリンゴの側にカーブを描く動きを見せている。これにより行為者は、目標がリンゴであると観察者により早く推測させ、目標がナシだと誤解される可能性を減少することができている。

Side-visible シナリオ (図 4.10) では、観察者の視界にある果物がリンゴのみで、観測の非対称性が問題になる。original では、行為者はリンゴに直線的に向かった。リンゴに隣接するナシが観察者には見えていないので、Center シナリオとは異なり直線的な動きで目標をナシだと誤解されるリスクは少なく、リンゴの方向を向いた早い段階から、観察者は目標がリンゴと判断できると期待される。「推測されるエージェント」モデルの生成する表意動作もまた、original と同じくリンゴに向かって直線的に進む動きとなった。それに対して、FalseProjective では、Center エピソードと同様にリンゴの側に膨らんだカーブが生成された。この動きは、リンゴに近接するナシの存在が観察者にわかっている場合には、目標をナシと誤解される可能性を減じる点で有効であると考えられる。しかし、限定された視界を持つ観察者の視界からは、FalseProjective のカーブは遠回りであり、また、カーブの序盤では目標をリンゴと判断するのが難しいと考えられる。

Side-invisible シナリオ（図 4.11）において、original は最初、リンゴとナシの中間程度の方向に進み、そこから徐々にカーブしてリンゴに向かった。original の前半の動きは、目標がリンゴかナシか区別しにくいといえる。一方、「推測されるエージェント」モデル

の表意動作では、観察者の視界外にあるリンゴのやや外側に向いてから前進し、内側にカーブしながらリンゴへ向かっている。前進するまでの行為者の動きは、目標をナシだと誤解される可能性を低下させており、original と比べて目標を推測しやすくなっていると考えられる。Side-invisible シナリオでは、FalseProjective でも「推測されるエージェント」モデルと同じ動きが生成された。

Blind-inside（図 4.12）と Blind-outside（図 4.13）では、リンゴとナシのどちらの位置も知らない観測者にとって、行為者の動きは目標に関する情報を持たない。そのため、FalseProjective のカーブは目標の伝達には寄与しておらず、移動距離を増やしているのみといえる。対して、「推測されるエージェント」モデルによって生成された動きは、original と同じであった。この結果は、Blind エピソードで行為者の行動を変えることが、目標の伝達という観点からコストにしかならないことを、「推測されるエージェント」モデルが正しく推定で来ている結果だと考えられる。

4.9 表意動作の評価概要

「推測されるエージェント」モデルが、観測の非対称性を考慮することで、効果的に目標を伝達する動きを生成できていることを検証する実験を行った。実験は、シミュレーションスタディとユーザスタディの 2 種類である。シミュレーションスタディでは、4.8 節で示したシナリオ以外でも「推測されるエージェント」モデルが有効な表意動作を生成できているか調べた。ユーザスタディでは、4.8 節のシナリオを実験参加者に提示した際に「推測されるエージェント」モデルの動きが有効であるかを調べた。どちらの実験も、「推測されるエージェント」モデルによって生成された動きを、観察者を考慮せずに最短距離で目標に到達することを目指す original と、観測の非対称性を考慮せずに生成された FalseProjective と比較した。

4.10 シミュレーションスタディ

4.10.1 目的

シミュレーションスタディでは、「推測されるエージェント」モデルが生成する表意動作のスケーラビリティを調査した。特に、観測の非対称性を考慮することで、「推測されるエージェント」モデルが有効な表意動作を生成できていることを検証した。

4.10.2 方法

行為者の動きをもとに目標を分類する機械学習モデルを用意し、これを**機械観察者**とすることで、「推測されるエージェント」モデルの表意動作を検証した。機械観察者は、観察者の視点からの光景を RGB 画像の系列として入力され、その際の行為者の目標を教師あり学習によって学習する。機械観察者は、畳み込みニューラルネットワーク層、長・短期記憶（Long short-term memory: LSTM）層、全結合層からなる深層学習モデルであ

る。実験では、機械観察者が行為者の目標に対して推測している確率を比較し、これがより早く大きく向上するほど有効な表意動作であるとした。

まず、機械観察者の訓練と表意動作の評価のために、2 つのデータセットを作成した。データセットは、各エピソードにおける観察者視点から見たエージェントの 3 種類の動き（original、FalseProjective、「推測されるエージェント」モデル）と行為者の目標のラベルからなる。各エピソードでは、リンゴとナシがランダムの場所に配置される。リンゴとナシは行為者の視界内にあるが、観察者の視界内に存在するとは限らない。リンゴとナシのうち少なくとも 1 つが観察者の視界内に存在しない場合が、観察者と行為者の間に観測の非対称性が存在する場面となる。データセットからは、リンゴとナシのどちらも観察者から観測できないエピソードは除外した。これは、そのようなエピソードでは観察者がエージェントの目標を推測することが不可能であるためである。訓練用のデータセットでは、3 種類の動きの違いが機械観察者の判断に与えるバイアスを排除するため、3 種類の動きが完全に同じになるエピソードのみを抽出した。結果として、訓練用データセットとして 1,847 エピソードが得られた。評価用のデータセットからは、3 種類の動きの相違に注目するため、3 種類が全く同じ動きをするエピソードは除外した。結果的に、695 エピソードが残った。

実験では異なる乱数シードをもとに 5 種類の機械観察者を用意し、機械観察者が推測する目標の確率の平均値を動きの評価値とした。5 種類の機械観察者の間の検者間信頼性は、0.936 だった（$\mathrm{ICC}(3, k)$）。

各エピソードは、動きの種類によって長さが異なる。3 種類の動きを同じ条件で比較するため、各エピソードの 3 種類の動きの中で、目標にたどり着いた最も短い時間に合わせて以降を評価からは切り捨てた。

4.10.3 仮説

観測の非対称性を考慮することで、「推測されるエージェント」モデルが、original と FalseProjective に比べて高い評価値を獲得できると予想した。

H1 「推測されるエージェント」モデルの表意動作の評価値は、他の動きに比べて高い。

具体的には、観測の非対称性が存在する場面では「推測されるエージェント」モデルが他の 2 種類の動きに比べて高い評価値を獲得し、一方で観測の非対称性が存在しない場面では、差が見られないと予想した。

H2 「推測されるエージェント」モデルの評価値は、観測の非対称性が**存在する**状況では他の動きに比べて高く、観測の非対称性が**存在しない**場面では、差が生じない。

4.10.4 結果

R1 図 4.14 に、時刻 t における評価値を、動きの種類ごとに示す。総合的にみて、「推測されるエージェント」モデルが生成した表意動作は original、FalseProjective より高い評価値を獲得した（$3 \leq t$）。時刻 $1, 2$ では、original の評価値が高かった。

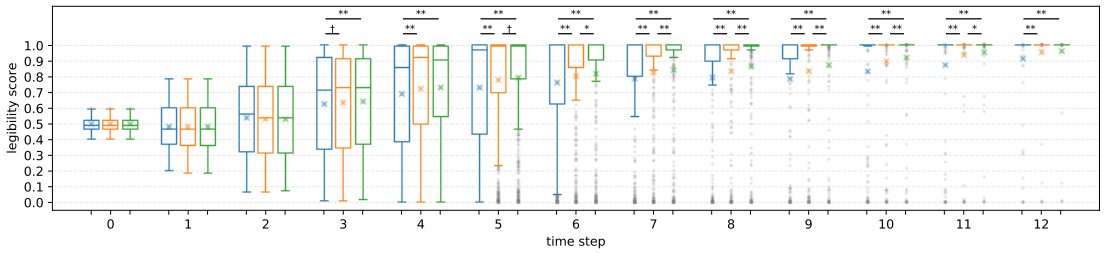


図 4.14: 3 種類の動きが獲得した評価値の遷移
左: original, 中央: FalseProjective, 右: 「推測されるエージェント」モデル。X マークは、平均値を示す。記号は、多重 t 検定の結果 ($** : p < .01, * : p < .05, \dagger : p < .1$)。

結果を統計的に分析するため、各時刻 t における結果を Friedman 検定によって比較したところ、時刻 $3 \leq t \leq 12$ において、3 種類の動きの間に有意な差が見られた ($p = .036$ at $t = 3$ and $p < .01$ at $4 \leq t$)。Friedman 検定で有意差が見られた $3 \leq t$ に関して、Wilcoxon の符号順位検定によって多重比較し、Holm-Sidak 法によって p 値の補正を行う方法で事後検定を行った。結果を図 4.14 に示す。時刻 $4 \leq t \leq 12$ において、FalseProjective と「推測されるエージェント」モデルの両者が original に関して有意に高い評価値を獲得していることが分かった。さらに、時刻 $6 \leq t \leq 12$ において、「推測されるエージェント」モデルの評価値は FalseProjective と比べて優位に高かった。Wilcoxon の符号順位検定における動きの差の効果量 r は、original と FalseProjective の間で時刻 $t = 12$ で最も高く、$r = .23$ だった。また、original と「推測されるエージェント」モデルの間では、$t = 12$ で最大値 $r = .29$、FalseProjective と「推測されるエージェント」モデルの間では $t = 9$ で最大値 $r = .17$ であった。

これらの結果は、仮説 **H1** を支持している。

R2 図 4.15 と 4.16 はそれぞれ、観測の非対称性が存在する場合と存在しない場合における、FalseProjective と「推測されるエージェント」モデルの評価値を示している。観測の非対称性が存在する場合では、「推測されるエージェント」モデルによって生成された動きは FalseProjective より高い評価値を獲得している。Wilcoxon の符号順位検定の結果、時刻 $t \leq t \leq 10$ において、両者の間に有意な差がみられた。一方で、観測の非対称性が存在しない場面では、2 種類の動きの評価値に有意な差は見られなかった。

以上の結果は、仮説 **H2** を指示しているといえる。

4.10.5　シミュレーション実験のまとめ

実験結果 **R1** から、ランダムに生成された 695 シナリオにおいて、「推測されるエージェント」モデルが比較手法である original と FalseProjective に比べて高い評価値を獲得できることが示された。この結果は、仮説 **H1** を支持しており、「推測されるエージェント」モデルが 4.8 節で示したシナリオ以外でも有効な表意動作を生成できているといえる。さらに、仮説 **H2** は結果 **R2** によって支持された。この結果からは、観測の非対称性を考慮することが表意動作の生成において重要な要因であり、「推測されるエージェント」モデ

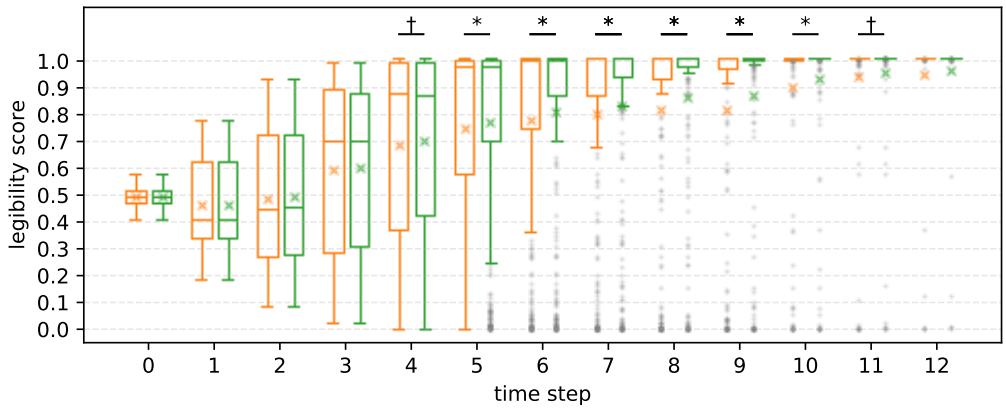


図 4.15: 観測の非対称性が存在する場面における FalseProjective と「推測されるエージェント」モデルの評価値

左: FalseProjective, 右: 「推測されるエージェント」モデル。

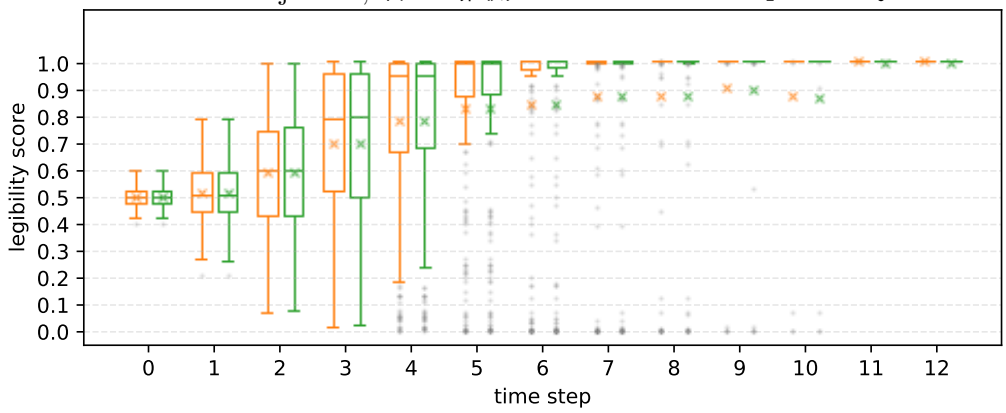


図 4.16: 観測の非対称性が存在する場面における FalseProjective と「推測されるエージェント」モデルの評価値

両者の間に有意な差はみられなかった。

ルがそれを適切に考慮出来ていると結論づけられる。

4.11 ユーザスタディ

4.11.1 目的

ユーザスタディでは、「推測されるエージェント」モデルによって生成される表意動作が人間の観察者にとっても有効であるかを検証した。具体的には、人が推定する行為者の目標の精度を、3 種類の動きに対して比較した。さらに、簡易なアンケートにより、人が 3 種類の動きに対して抱く印象を調査した。

4.11.2 方法

12 人の大学生・大学院生 (男性 6 名、女性 6 名; 20-24 歳, $M = 22.6, SD = 1.83$) を対象に、実験を行った。彼らには謝金として 750 円を支払った。

実験参加者には、観察者の視点からの行為者の動きを動画で提示し、再生する最中に行為者がリンゴとナシのどちらに向かっているかを予測させた。動画のフレームレートは 10Hz とした。参加者は、動画を見ながらキーボードの F キーと J キーを押すことで、予測を回答した。F キー/J キーはそれぞれ、目標をリンゴ/ナシと予測した際に押下させた。キーの割り当てによるバイアスを避けるため、キーと果物の対応関係は参加者毎にランダムに決めた。参加者は F キーと J キーのどちらも押さないことで、行為者の目標が判断できないことを表現することもできる。

参加者には、行為者がリンゴかナシのどちらかを目標に定め、目標に向かって動くこと、行為者の目標は試行毎にランダムで決定していると教示した。また、環境に観察者と行為者が固定地点に、リンゴとナシがランダムな場所に配置されている状況から各エピソードが始まると伝えた。

実験では、original、FalseProjective、「推測されるエージェント」モデルの 3 条件を、参加者内計画で比較した。条件の順序は、参加者間でカウンターバランスをとった。それぞれの条件では、9 種類の動画を提示した。9 種類のうち 5 種類は、4.8 節に挙げたシナリオである。ただし、リンゴとナシの位置関係は試行毎にランダムに入れ替えている。残りの 4 種類は、参加者が条件毎に同じシナリオを提示されていることを気づかせないために追加した。

参加者の推測結果を評価する指標として、*rapidity* と *correctness* の 2 つを設定した。*rapidity* は、行為者の動きからはその目標が明白で、曖昧性が低く、参加者がより早い段階で正しい目標を推定できることを測る。また、*correctness* は、その動きを見て参加者が誤った推定をしないかを測る指標である。

参加者による目標の推測結果に加えて、各条件の動きに対する印象を問う簡単なアンケートを実施した。質問項目は以下のとおりである：

Q1. エージェントがどちらの果物に向かっているか予測するのは簡単だった (Legibility 1)

Q2. エージェントの動きからは、意図が明確だった (Legibility 2)

Q3. エージェントの動きには一貫性があった (Consistency)

Q1、Q2 は、表意動作の先行研究 [12] から引用したもので、動きが行為者の目標を有効に伝達できているかを問うものである。さらに、Side-visible シナリオにおける FalseProjective の遠回りに見えるカーブが、誤解を与える矛盾のある動きと見られると考え、これを検証する Q3 を追加した。

4.11.3 仮説

表 4.3 と 4.4 に、*rapidity*・*correctness* 指標に関する仮説を、Center、Side-visible、Side-invisible のシナリオごとに示す。Center シナリオでは、original の動きが持つ曖昧性が FalseProjective と「推測されるエージェント」モデルの動きにはなく、結果、original と

表 4.3: *rapidity* 指標に関する仮説

	Original	FalseProjective	「推測されるエージェント」モデル
Center	✗	✓	✓
Side-visible	✓	✗	✓
Side-invisible	✗	✓	✓

表 4.4: *correctness* 指標に関する仮説

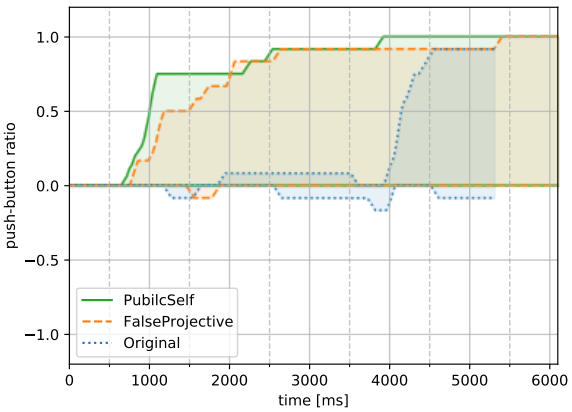
	Original	FalseProjective	「推測されるエージェント」モデル
Center	✓	✓	✓
Side-visible	✓	✗	✓
Side-invisible	✗	✓	✓

比べて他 2 種類の *rapidity* の指標が向上すると期待した。一方、Center シナリオの動きは参加者に誤解を与えるものではなく、*correctness* 指標に関しては差がみられないと考えた。Side-visible シナリオでは、目標の果物の側にカーブを描く FalseProjective の動きは、目標でない方の果物が見えていない参加者からには目標が明確でなく、目標に向かって直進する original、「推測されるエージェント」モデルと比較して、*rapidity* 指標が低下すると考えた。また、FalseProjective の序盤の動きは観察者の側に向かっている動きに見え、エージェントの目標が視界内の果物ではないと誤解を与える可能性があると考え、*correctness* 指標に関しても低下すると仮説を立てた。*correctness* に関しては同様に、Side-invisible シナリオの original が参加者に誤解を与える動きであり、スコアが低下すると考えた。また、Side-invisible シナリオでは FalseProjective と「推測されるエージェント」モデルの動きが、行為者の目標が視界内の果物でないことを強調することで目標を効果的に伝達していると考え、*rapidity* 指標が向上すると仮説を立てた。

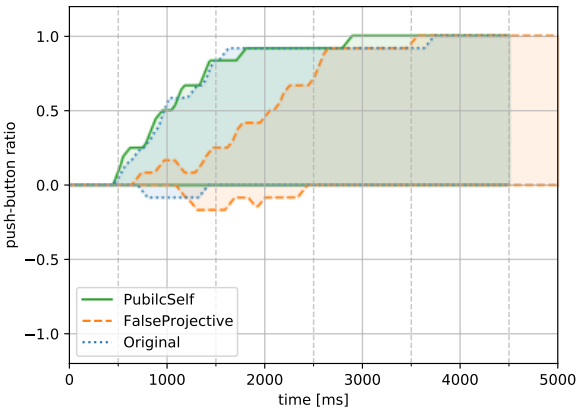
主観評価では、「推測されるエージェント」モデルが観測の非対称性を考慮することで有効な表意動作を生成でき、結果として legibility の指標が最も高くなると期待した。また、original は動きの曖昧から、legibility の指標が最も低くなると予想した。FalseProjective は、観測の非対称性を考慮しなくても「推測されるエージェント」モデルと同じ動きになるシナリオのポジティブな寄与と、Side-invisible や Blind-inside シナリオの遠回りのネガティブな影響によって、original と「推測されるエージェント」モデルの間の評価になると考えた。加えて、consistency の指標では、FalseProjective の遠回りが誤解を与えることで、FalseProjective のみで指標が低下すると考えた。

4.11.4 結果

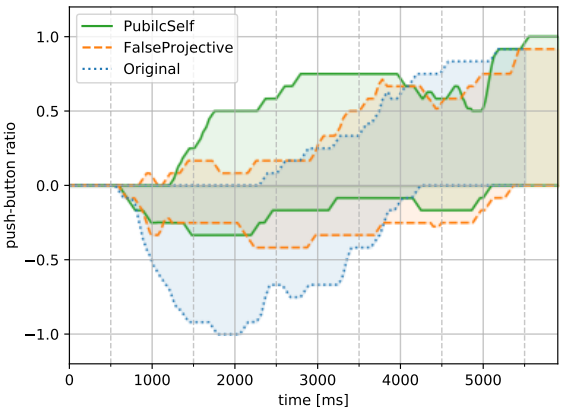
図 4.17 に、Center、Side-visible、Side-invisible シナリオにおいて参加者が行為者の目標を推定した結果を示す。正答した参加者の人数を *rapidity* の指標、誤答した参加者の人数を *correctness* の指標として議論する。表 4.5、4.6 に、仮説の検証結果をまとめる。



(a) Center



(b) Side-visible



(c) Side-invisible

図 4.17: 参加者の推測結果
グラフの上部は行為者の目標を正しく回答している参加者の数を示し、*rapidity* 指標と対応する。

まず、*rapidity* の結果に着目しよう。Center エピソードの結果は、仮説の通りであった。FalseProjective と「推測されるエージェント」モデルの動きは、どちらも行為者の目標を早期に明確にしており、参加者は早い段階で正答できた。一方 original は、目標の果物に方向転換する瞬間まで目標の判断がつかず、参加者の回答に時間がかかっていることが分かる。Side-visible エピソードの結果も、仮説を支持した。FalseProjective のカーブした動きは行為者の目標を伝達する動きとして適当でなく、original や「推測されるエージェント」モデルの直線的な動きと比べて参加者が回答するまでに時間がかかってしまっている。Side-invisible エピソードでは、仮説のとおり、参加者は original よりも「推測されるエージェント」モデルの動きによってより早く行為者の目標を推定出来た。一方、予想に反して、「推測されるエージェント」モデルと同じ動きである FalseProjective では参加者が正答するのに時間がかかった。その理由として、FalseProjective の動きを見る参加者が判断に慎重になっていた可能性が挙げられる。参加者の 2 人は事後のアンケートで、FalseProjective の動きが遠回りに見えたと指摘した。「推測されるエージェント」モデルの動きも効率の観点から見ると original と比べて遠回りであるが、「推測されるエージェント」モデルに対して遠回りと指摘する参加者はおらず、FalseProjective に特異的であった。参加者が FalseProjectie の動きが遠回りと認識していたとすると、行為者の序盤の動きが目標と即座に結びつかなくなるため、参加者が早期の回答を避けるようになり、*rapidity* 指標が低下したと推測できる。ここまでの結果をまとめると以下のようになる。(i)「推測されるエージェント」モデルは、観測の非対称性を考慮したうえで表意動作を生成することで、人間の参加者に行為者の目標をより早く伝達できた。(ii) FalseProjective の遠回りな動きによって、参加者は序盤の回答に慎重になった可能性がある。

correctness の指標では、Center と Side-visible で誤答した参加者はほとんどおらず、遠回りを見せる Side-visible の FalseProjective においても同様だった。この結果は当初の仮説に反する一方で、前述の結果 (ii) を支持している。つまり、参加者は早期の回答を避けることで誤答を回避する戦略を取った可能性が、ここからも示唆される。Side-invisible シナリオでは、仮説のとおり original の動きを見た参加者が 2,000ms 付近で誤答した。FalesProjective と「推測されるエージェント」モデルの動きでは、数人の参加者が誤答したものの、original と比べて誤答の数を減らすことができた。まとめると、(iii) FalseProjective と「推測されるエージェント」モデルの動きは、original と比較して誤答の数を減少させることができた。(iv) *correctness* の観点からは、FalseProjective と「推測されるエージェント」モデルの間の差はみられなかった。

図 4.18 に印象評価の結果をまとめる。一元配置反復測定分散分析の結果、Q1 と Q2 で有意な差があることが分かった (Q1: $F(2, 22) = 4.52$, $p < .05$, $\eta^2 = .21$、Q2: $F(2, 22) = 4.83$, $p < .05$, $\eta^2 = .26$)。Q1 と Q2 の結果に対して Tukey 法による事後検定を行った結果、FalseProjective と「推測されるエージェント」モデルの動きが original と比較して優位に高く評価されたことが分かった ($p < .05$)。参加者の 2 人は、Center シナリオにおける original の動きが、legibility に関して強くネガティブな印象を引き起こしたと説明した。仮説に反し、FalseProjective と「推測されるエージェント」モデルの間に有意な差

表 4.5: *rapidity* 指標に関する結果

	Original	FalseProjective	「推測されるエージェント」モデル
Center	✗	✓	✓
Side-visible	✓	✗	✓
Side-invisible	✗	✗	✓

表 4.6: *correctness* 指標に関する結果

	Original	FalseProjective	「推測されるエージェント」モデル
Center	✓	✓	✓
Side-visible	✓	✓	✓
Side-invisible	✗	△	△

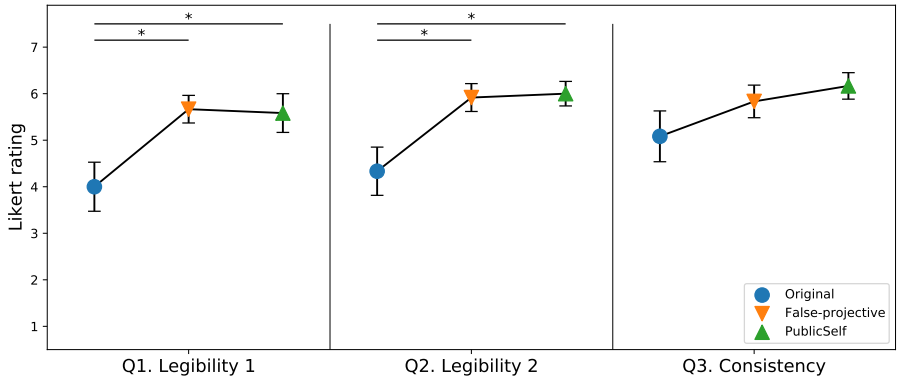


図 4.18: 参加者の主観評価
エラーバーは標準誤差を示す。FalseProjective と「推測されるエージェント」モデルは legibility 指標において original よりも高い評価を得た。FalseProjective と「推測されるエージェント」モデルの間には有意な差は見られなかった。

は見られなかった。2 人の参加者が FalseProjective の遠回りを指摘したものの、その影響は Center エピソードや Side-visible エピソードの成功と比べて限定的であった。また、FalseProjective が consistency を低く評価されることもなかった。まとめると、主観評価の観点では、FalseProjective と「推測されるエージェント」モデルの両方が、original と比べて高い評価を得た。

4.11.5 ユーザスタディのまとめ

「推測されるエージェント」モデルが、観測の非対称性を考慮したうえで表意動作を生成することによって、人間の観察者がより早く行為者の目標を推測できた。観測の非対称性を考慮しない FalseProjective が参加者の誤答を誘発する結果は得られなかったが、

FalseProjective の遠回りな動きによって、人が判断に慎重になり、結果として推測が遅れた可能性が示唆された。

4.12 「推測されるエージェント」モデルの限界と今後の展望

実験の結果、表意動作を生成する上で観測の非対称性を考慮する必要があること、「推測されるエージェント」モデルが観測の非対称性を考慮することで有効な表意動作を生成できていることが分かった。一方で、「推測されるエージェント」モデルをそのまま応用できる範囲に関しては議論が必要である。

例えば実験では、観測の非対称性の議論に注力するため、観察者の視点を固定し、環境にはリンゴとナシが 1 つずつ存在するなどの前提を置いた。また、リンゴとナシが両方エージェントの視界にあるという前提によって、エージェントが環境を探索する状況は扱っていない。しかし、現実の人とエージェントのインタラクションでは、人は動いたり、行動によって環境を変化させることができる。また、環境中の何がどこにあるかといった情報は、多くの場合エージェントと人の両者にとって不確実である。不確実性は、「推測されるエージェント」モデルが考慮すべき可能性を増大させ、計算量の面で問題となる。複雑性がある程度までに達した場合には、現在の「推測されるエージェント」モデルによるボトムアップな推測ではなく、過去の経験や観測に基づく信念にもとづくトップダウンな判断の有効性が高まると予想される。

アルゴリズム 3 の、目標を最も伝達する行動と、目標に向けて最も合理的な動きを選択する閾値は、著者が人手で調整している。しかし、最適な閾値は状況によって異なる。例えば、状況の中で目標を動きで伝達する優先順位・重要性が閾値の決定に影響する。実際の応用においては、強化学習を用いて閾値を自動で最適化するアプローチも考えられる。

「推測されるエージェント」モデルは、エージェントに帰属される心的状態をエージェントの内部で推測するのみで、人から受けるフィードバックは考慮できていない。人を介さない内的な推定と、人との明示的なコミュニケーションを組み合わせることで、より正しく、かつ効率的に帰属される心的状態を推測できるだろう。また、エージェントの意思決定を伝達する方法として、動きと言語のそれぞれ一方でなく、両者を組み合わせたマルチモーダルなコミュニケーションを行うことの有効性を検証していきたい。

ケーススタディ（4.6）で見られたように、人はエージェントが目標をもとに合理的に行動しているという前提自体を疑うことがある。また、強化学習によって行動を学習した際に得られる方策は合理的な最適解になるとは限らず、人から見てエージェントの行動が合理的に見えるかどうかもまた別の問題である。「推測されるエージェント」モデルで前提となっている、人がエージェントに抱くスタンスからモデルに推測させることが必要になる場面もあるだろう。

実験では、各シナリオを独立したものとして提示し、前に提示したシナリオがその後のシナリオに影響を与えないよう教示を行った。しかし、人の他者に関する信念は、インタラクションを通して長期的なスパンで構築されていく。人とエージェントの長期的なインタラクションを考えると、過去のインタラクションが現在の心的状態の帰属に影響をもた

らす仮定をモデルに組み込むことは有望と考えられる。

4.13 まとめ

本章では、エージェントの動きを見た人がエージェントに対して推測する心的状態を、エージェント自体が推測する、という 2 次の推測をモデル化した「推測されるエージェント」モデルを提案した。また、「推測されるエージェント」モデルの特徴として、人とエージェントが異なる視点を持つために生じる観測の非対称性がモデルに組み込まれている点を説明した。さらに、「推測されるエージェント」モデルの推測を応用することで、動きによってエージェントの目標を予告する表意動作の生成手法を提案した。

「推測されるエージェント」モデルを用いた実験では、エージェントに帰属される心的状態の推定や表意動作の生成において、人とエージェントの間の観測の非対称性を考慮することが重要であること、「推測されるエージェント」モデルが観測の非対称性を適切に扱うことができていることを示した。

5 章

研究全体の議論と制約・今後の展望

本論文では、人と AI エージェントの間で、AI エージェントの挙動に関する共通認識を構築する挙動アライメントについて考えた。そして、人からエージェントへの指示と、エージェントから人への伝達を通じて挙動アライメントを行うコミュニケーションモデルを 2 つ提案し、両モデルが人とエージェントの間の非対称性を考慮することで適切に挙動アライメントを行うことを実証してきた。3 章では、人がエージェントに対して抱く期待をエージェントが推定する「期待されるエージェント」モデルを提案し、人からの期待を考慮することで指示を適切に理解できるようになること、理解した指示の語彙をもとにエージェントの動きを伝達することで、人がエージェントの未来の動きを精度よく理解できるようになることを示した。4 章では、AI エージェントの動きを見た人がエージェントに帰属する目標を、エージェントの側から推定する「推測されるエージェント」モデルを提案した。また、「推測されるエージェント」モデルをもとに、AI エージェントが動きによって目標を効果的に伝達できるようになることを示した。これらの研究結果は、〈心〉を前提としてエージェントの存在をメタ的に捉えるモデルがコミュニケーションに果たす役割の一端を示している。

〈心〉を前提にするということは、信念や欲求、意図といった心的状態で他者を説明するというだけでなく、そもそも自己以外の存在がそれぞれに異なる〈心〉を持った他者だと認識することである。従来研究は目標や観測の共有を前提とすることで、心的状態を考慮する必要のない問題設定を構築している。

3、4 章では、AI エージェントと人とのインタラクションにおいて目標や観測の共有は必ずしも成り立たないこと、そうした状況において目標・観測の共有を前提とした手法は機能しないこと、そして、他者の心的状態や他者からみた自己の心的状態を推定することが、コミュニケーションの成立に寄与すること示している。

心的状態の共有を前提とすることは、確かにコミュニケーションを円滑に進めるための基盤となり得る。例えば、協調場面において参加者の間で目標が共有できているならば、目標に関するコミュニケーションを省略して具体的な役割や行動について話し合うという選択ができる。しかし、コミュニケーションを行うか省略するかの選択は容易でない。直接観測できない他者の心的状態は常に不確実性を持っており、また時とともに変化し得るため、心的状態を共有出来た/出来ていないという 2 択で表現できないからである。「期待

されるエージェント」モデルと「推測されるエージェント」モデルは、人と AI エージェントの間で目標と信念が共有できていると考える従来研究の前提を緩和してはいるものの、エピソードという短い区切りの中でのコミュニケーションを扱い、また 3 章の指示者のように、エピソードにおける目標の一貫性も仮定しているため、こうした不確実性を内在したコミュニケーションの問題には踏み込めていない。本研究を発展させる方向として、より長いスパンのコミュニケーションの中で心的状態の共有度合いを継続的に推定し続けること、共有度合いに関する不確実な推定をもとに、コミュニケーションを円滑化させるための戦略を考えることが考えられる。

人の視界の範囲からボトムアップに人の観測を推定している「推測されるエージェント」モデルと異なり、「期待されるエージェント」モデルは観測可能な人の言動とその背後の心的状態の間にある整合性を双方向に解消するという観点から心的状態の推定を行っている。著者は、言動と心的状態の整合性解消というアイデアが、本来観測できない他者の心的状態をコミュニケーションの中で人が徐々に理解していく過程の全般に見出されると考えている。例えば、自分に対して好意を持っているかが気になる相手がいるとする。その相手とのコミュニケーションや、相手の日ごろの言動は、相手が自分に好意を持っているかどうかを判断する材料となる。逆に、相手が自分に対して好意を持っている/いないという信念は、言動の解釈に影響を与える。相手が素っ気ない態度をとってきたとき、相手が自分に好意を持っているという強い信念があると、その態度が天邪鬼や照れであって本心でないという、信念と整合する解釈がより尤もらしく見える。また、こうした解釈は信念をさらに強化する。「期待されるエージェント」モデルで扱った目標の推定と指示の解釈という枠組みを超えて、言動と心的状態の間にある整合性の双方向的解消というアイデアをより一般的な問題設定の中で探求していきたい。

本研究で扱った観測や信念、目標といった心的状態のほかにも、人と AI エージェントの相互理解を考えるうえで重要な要素は多くある。その 1 つに、**信頼**が挙げられる。Lee and See は特定のタスクに従事するコンピュータに対する人の信頼として、「不確実性や脆弱性がある状況下で、エージェントが人の目標達成に貢献しそうだと考える態度」という定義を提案している [32]。AI エージェントの性能を最大限に活かしつつエラーやミスの影響を抑えるために、エージェントの性能を人が過剰に見積る過信や、逆にエージェントの性能を過小評価する不信は回避することが望ましい [45]。エージェントは、人からエージェントの性能がどのように評価されているか推測し、結果をもとに人からの信頼を正しく**較正**できるようにすることが好ましい。新たな研究の方向性として、人とエージェントの間の非対称性の中で人の信頼を推測するよう「推測されるエージェント」モデルを拡張し、また表意動作のようにエージェントの挙動が人の信頼にもたらす効果を予測することで、信頼を較正する振る舞いを生成することが考えられる。

6 章

まとめ

本論文では、AI エージェントが行おうとしている挙動とユーザがエージェントに期待する挙動を一致させ、両者の間に共通認識を構築する挙動アライメントを目的として、AI エージェントと人とのコミュニケーションを考えた。特に、人からエージェントへの挙動の指示と、エージェントから人への挙動の伝達に着目し、これらを達成する挙動アライメント・コミュニケーションモデルを提案した。また、挙動アライメント・コミュニケーションモデルを AI エージェントに実装し、その有効性を実験的に示した。

提案した挙動アライメント・コミュニケーションモデルの 1 つである「期待されるエージェント」モデルは、人とエージェントの間に存在する目標の非対称性を考慮しながら挙動アライメントを行う。AI エージェントが達成しようとする目標と独立に、人がエージェントに達成を期待する目標を考慮することで、人からの指示の語彙を適切に解釈できた。また、解釈した語彙をエージェントの動きの予告に流用することで、人はエージェントがどういった挙動を見せようとしているかを精度よく理解できるようになった。

もう 1 つの挙動アライメント・コミュニケーションモデルである「推測されるエージェント」モデルは、AI エージェントと人がそれぞれの視点から異なる観測をしていることに起因する観測の非対称性を考慮することで、エージェントの目標を人に誤って理解されていることを検知できた。また、「推測されるエージェント」モデルをもとに生成した、エージェントの目標を伝達する動き (表意動作) によって、人がエージェントの目標を早く・正しく理解できることを示した。

両モデルに共通するのは、エージェントと人との間で、目標や信念といった心的状態を帰属し合う〈心〉の読みあいの中で、両者の間に存在する情報の差異（非対称性）を明示的に組み込んでいることであった。モデルの検証結果を通じ、AI エージェントと人の非対称性を考慮することで、挙動アライメントを効果的に達成できることがわかった。

参考文献

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, Vol. 6, pp. 52138–52160, 2018.

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Snderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018.

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, I. Reid, Stephen Gould, and A. V. Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018.

[4] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pp. 1078–1088, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.

[5] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, Vol. 297, p. 103500, 2021.

[6] Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, Vol. 1, p. 0064 EP, 2017.

[7] Benjamin Beyret, Ali Shafti, and A. Aldo Faisal. Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5014–5019, 2019.

[8] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

[9] Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, and Ruslan Salakhutdinov. Gated-attention architectures for task-oriented language grounding. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] Daniel C. Dennett. *The Intentional Stance*. MIT Press, 1987.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] Anca Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S. Srinivasa. Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pp. 51–58, New York, NY, USA, 2015. ACM.

[13] Anca Dragan and Siddhartha Srinivasa. Integrating human observer inferences into robot motion planning. *Auton. Robots*, Vol. 37, No. 4, pp. 351–368, December 2014.

[14] Mark Edmonds, Feng Gao, Hangxin Liu, Xu Xie, Siyuan Qi, Brandon Rothrock, , Yixin Zhu, Ying Nian Wu, Hongjing Lu, and Song-Chun Zhu. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics*, Vol. 4, No. 37, eaay4663, 2019.

[15] Adam Falewicz and Waclaw Bak. Private vs. public self-consciousness and self-discrepancies. *Current Issues in Personality Psychology*, Vol. 4, No. 1, pp. 58–64, 2015.

[16] A. Feningstein. Public and private self-consciousness : Assessment and theory. *Journal of Consulting and Clinical Psychology*, Vol. 43, pp. 522–527, 1975.

[17] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, and Michita Imai. Application of instruction-based behavior explanation to a reinforcement learning agent with changing policy. In *International Conference on Neural Information Processing*, pp. 100–108. Springer, 2017.

[18] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, and Michita Imai. Autonomous self-explanation of behavior for interactive reinforcement learning agents.

In *Proceedings of the 5th International Conference on Human Agent Interaction*, HAI '17, pp. 97–101, New York, NY, USA, 2017. Association for Computing Machinery.

[19] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, and Michita Imai. Explaining intelligent agent's future motion on basis of vocabulary learning with human goal inference. *IEEE Access*, Vol. 10, pp. 54336–54347, 2022.

[20] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, Tatsuji Takahashi, and Michita Imai. Bayesian inference of self-intention attributed by observer. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, HAI '18, pp. 3–10, New York, NY, USA, 2018. Association for Computing Machinery.

[21] Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, Tatsuji Takahashi, and Michita Imai. Conveying intention by motions with awareness of information asymmetry. *Frontiers in Robotics and AI*, Vol. 9, , 2022.

[22] Gyrgy Gergely, Zoltn Ndasdy, Gergely Csibra, and Szilvia Br. Taking the intentional stance at 12 months of age. *Cognition*, Vol. 56, No. 2, pp. 165–193, 1995.

[23] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, Vol. 26, pp. 2625–2633. Curran Associates, Inc., 2013.

[24] Bradley Hayes and Brian Scassellati. Challenges in shared-environment human-robot collaboration. In *Collaborative Manipulation Workshop at the ACM/IEEE International Conference on Human-Robot Interaction (HRI 2013)*, Vol. 8, p. 9, 2013.

[25] Bradley Hayes and Julie A. Shah. Improving robot controller transparency through autonomous policy explanation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 303–312, 2017.

[26] Daniel Hein, Alexander Hentschel, Thomas Runkler, and Steffen Udluft. Particle swarm optimization for generating interpretable fuzzy reinforcement learning policies. *Engineering Applications of Artificial Intelligence*, Vol. 65, pp. 87–98, 2017.

[27] Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, Marcus Wainwright, Chris Apps, Demis Hassabis, and Phil Blunsom. Grounded language learning in a simulated 3d world. *CoRR*, Vol. abs/1706.06551, , 2017.

[28] Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar, and Katia Sycara. Transparency and explanation in deep reinforcement learning neural networks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 144–150, New York, NY, USA, 2018. Association for Computing Machinery.

[29] Julian Jara-Ettinger, Hyowon Gweon, Joshua B. Tenenbaum, and Laura E. Schulz. Children ' s understanding of the costs and rewards underlying rational action. *Cognition*, Vol. 140, pp. 14–23, 2015.

[30] W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pp. 9–16. ACM, 2009.

[31] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable agency for intelligent autonomous systems. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 4762–4763. AAAI Press, 2017.

[32] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, Vol. 46, No. 1, pp. 50–80, 2004. PMID: 15151155.

[33] Guiliang Liu, Oliver Schulte, Wang Zhu, and Qingcan Li. Toward interpretable deep reinforcement learning with linear model u-trees. In Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part II*, Vol. 11052 of *Lecture Notes in Computer Science*, pp. 414–429. Springer, 2018.

[34] Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob N. Foerster, Jacob Andreas, Edward Grefenstette, S. Whiteson, and Tim Rocktäschel. A survey of reinforcement learning informed by natural language. In *IJCAI*, 2019.

[35] Yuyan Luo and Renée Baillargeon. Can a self-propelled box have a goal? psychological reasoning in 5-month-old infants. *Psychological Science*, Vol. 16, No. 8, pp. 601–608, 2005.

[36] Hongyuan Mei, Mohit Bansal, and Matthew Walter. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30, No. 1, Mar. 2016.

[37] Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pp. 2772–2778. AAAI Press, 2016.

[38] Dipendra Kumar Misra, John Langford, and Yoav Artzi. Mapping instructions and visual observations to actions with reinforcement learning. In *EMNLP*, 2017.

[39] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy P. Lillicrap, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1928–1937. JMLR.org, 2016.

[40] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, Vol. 518, No. 7540, pp. 529–533, 2015.

[41] Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker. Model-based reinforcement learning: A survey. *CoRR*, Vol. abs/2006.16712, , 2020.

[42] Alex Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo J. Rezende. *Towards Interpretable Reinforcement Learning Using Attention Augmented Agents*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[43] Stefanos Nikolaidis, Anca Dragan, and Siddharta Srinivasa. Viewpoint-based legibility optimization. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, pp. 271–278. IEEE Press, 2016.

[44] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc., 2015.

[45] Kazuo Okamura and Seiji Yamada. Empirical evaluations of framework for adaptive trust calibration in human-ai cooperation. *IEEE Access*, Vol. 8, pp. 220335–220351, 2020.

[46] Marc Oliu, Javier Selva, and Sergio Escalera. Folded recurrent neural networks for future video prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[47] Xiaoshan Pan, Charles Han, Ken Dauber, and Kincho Law. A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations. *AI Soc.*, Vol. 22, pp. 113–132, 10 2007.

[48] David Premack and Ann James Premack. Motor competence as integral to attribution of goal. *Cognition*, Vol. 63, No. 2, pp. 235–242, 1997.

[49] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 4218–4227. PMLR, 10–15 Jul 2018.

[50] Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. Modeling others using oneself in multi-agent reinforcement learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 4257–4266. PMLR, 10–15 Jul 2018.

[51] S. Reddy, A. Dragan, and Sergey Levine. Where do you think you're going?: Inferring beliefs about dynamics from behavior. In *NeurIPS*, 2018.

[52] Stuart Russell. *Human compatible: Artificial intelligence and the problem of control.* Penguin, 2019.

[53] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Mller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *ITU Journal: ICT Discoveries*, Vol. 1, Special Issue, No. 1, pp. 1–10, 2017.

[54] Ştefan Sarkadi, Alison R. Panisson, Rafael H. Bordini, Peter McBurney, Simon Parsons, and Martin Chapman. Modelling deception using theory of mind in multi-agent systems. *AI Commun.*, Vol. 32, No. 4, pp. 287–302, jan 2019.

[55] Brian J Scholl and Patrice D Tremoulet. Perceptual causality and animacy. *Trends in cognitive sciences*, Vol. 4, No. 8, pp. 299–309, 2000.

[56] Pararth Shah, Marek Fiser, Aleksandra Faust, J. Chase Kew, and Dilek Hakkani-Tür. Followmet: Robot navigation by following natural language directions with deep reinforcement learning. *CoRR*, Vol. abs/1805.06150, , 2018.

[57] Tianmin Shu, Caiming Xiong, and Richard Socher. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[58] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton,

Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, Vol. 550, pp. 354–359, October 2017.

[59] Molly Wright Steenson. *AI, Ethics, and Design: Revisiting the Trolley Problem*, pp. 513–533. Springer International Publishing, Cham, 2021.

[60] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

[61] Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. Interpreting black-box classifiers using instance-level visual explanations. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, HILDA'17, New York, NY, USA, 2017. Association for Computing Machinery.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998–6008. Curran Associates, Inc., 2017.

[63] J Waa, J van Diggelen, K Bosch, and M Neerincx. Contrastive explanations for reinforcement learning in terms of expected consequences. In *Proceedings of the Workshop on Explainable AI on the IJCAI conference, Stockholm, Sweden., 37*, 2018.

[64] Gerhard Weiss, editor. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. MIT Press, Cambridge, MA, USA, 1999.

[65] Nevan wichers, Ruben Villegas, Dumitru Erhan, and Honglak Lee. Hierarchical long-term video prediction without supervision. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 6038–6046. PMLR, 10–15 Jul 2018.

[66] Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, Vol. 13, No. 1, pp. 103–128, 1983.

[67] Yufei Ye, Maneesh Singh, Abhinav Gupta, and Shubham Tulsiani. Compositional video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[68] Luke Zettlemoyer, Brian Milch, and Leslie P. Kaelbling. Multi-agent filtering with infinitely nested beliefs. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pp. 1905–1912. Curran Associates, Inc., 2009.

[69] 中川裕志. AI 倫理指針の動向とパーソナル AI エージェント. 情報通信政策研究, Vol. 3, No. 2, pp. 1–24, 2020.

業績

Journal articles

- Shoya Matsumori, Kohei Okuoka, Ryoichi Shibata, Minami Inoue, Yosuke Fukuchi, and Michita Imai. Mask and cloze: Automatic open cloze question generation using a masked language model. *IEEE Access*, Vol. 11, pp. 9835–9850, 2023.

- Riki Satogata, Mitsuhiko Kimoto, Yosuke Fukuchi, Kohei Okuoka, and Michita Imai. Q-mapping: Learning user-preferred operation mappings with operation-action value function. *IEEE Transactions on Human-Machine Systems*, Vol. 52, No. 6, pp. 1090–1102, 2022.

- Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, and Michita Imai. Explaining intelligent agent's future motion on basis of vocabulary learning with human goal inference. *IEEE Access*, Vol. 10, pp. 54336–54347, 2022.

- Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, Tatsuji Takahashi, and Michita Imai. Conveying intention by motions with awareness of information asymmetry. *Frontiers in Robotics and AI*, Vol. 9, , 2022.

Conference papers (Refereed)

- Rintaro Hasegawa, Yosuke Fukuchi, Kohei Okuoka, and Michita Imai. Advantage mapping: Learning operation mapping for user-preferred manipulation by extracting scenes with advantage function. In *Proceedings of the 10th International Conference on Human-Agent Interaction*, HAI '22, p. 95103, New York, NY, USA, 2022. Association for Computing Machinery.

- Ryoichi Shibata, Shoya Matsumori, Yosuke Fukuchi, Tomoyuki Maekawa, Mitsuhiko Kimoto, and Michita Imai. Utilizing core-query for context-sensitive ad generation based on dialogue. In *27th International Conference on Intelligent User Interfaces*, IUI '22, p. 734745, New York, NY, USA, 2022. Association for Computing Machinery.

- Yuta Watanabe, Yosuke Fukuchi, Tomoyuki Maekawa, Shoya Matsumori, and Michita Imai. Inferring human beliefs and desires from their actions and the content

of their utterances. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, HAI '21, p. 391395, New York, NY, USA, 2021. Association for Computing Machinery.

- Teppei Yoshino, Shoya Matsumori, Yosuke Fukuchi, and Michita Imai. Simultaneous contextualization and interpretation with keyword awareness. In *Artificial Intelligence and Soft Computing*, pp. 403–413. Springer International Publishing, 2021.

- Nanase Otake, Shoya Matsumori, Yosuke Fukuchi, Yusuke Takimoto, and Michita Imai. Mixed reference interpretation in multi-turn conversation. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*. SCITEPRESS - Science and Technology Publications, 2021.

- Shoya Matsumori, Kosuke Shingyouchi, Yuki Abe, Yosuke Fukuchi, Komei Sugiura, and Michita Imai. Unified questioner transformer for descriptive question generation in goal-oriented visual dialogue. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1898–1907, October 2021.

- Yosuke Fukuchi, Yusuke Takimoto, and Michita Imai. Adaptive enhancement of swipe manipulations on touch screens with content-awareness. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*. SCITEPRESS - Science and Technology Publications, 2020.

- Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, Tatsuji Takahashi, and Michita Imai. Bayesian inference of self-intention attributed by observer. In *Proceedings of the 6th International Conference on Human-Agent Interaction*. ACM, dec 2018.

- Shoya Matsumori, Yosuke Fukuchi, Masahiko Osawa, and Michita Imai. Do others believe what I believe?: Estimating how much information is being shared by utterance timing. In Michita Imai, Tim Norman, Elizabeth Sklar, and Takanori Komatsu, editors, *Proceedings of the 6th International Conference on Human-Agent Interaction, HAI 2018, Southampton, United Kingdom, December 15-18, 2018*, pp. 301–309. ACM, 2018.

- Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, and Michita Imai. Autonomous self-explanation of behavior for interactive reinforcement learning agents. In *Proceedings of the 5th International Conference on Human Agent Interaction*. ACM, oct 2017.

- Yosuke Fukuchi, Masahiko Osawa, Hiroshi Yamakawa, and Michita Imai. Application of instruction-based behavior explanation to a reinforcement learning agent with changing policy. In *Neural Information Processing*, pp. 100–108. Springer International Publishing, 2017.

Preprints

- Teppei Yoshino, Yosuke Fukuchi, Shoya Matsumori, and Michita Imai. Chat, shift and perform: Bridging the gap between task-oriented and non-task-oriented dialog systems, 2022.
- Yusuke Takimoto, Yosuke Fukuchi, Shoya Matsumori, and Michita Imai. Slam-inspired simultaneous contextualization and interpreting for incremental conversation sentences, 2020.

謝辞

本論文の執筆にあたり、多くの方々のご指導とご協力を賜りました。ここに感謝申し上げます。

はじめに、研究室に配属された学部 4 年生から現在まで指導教官としてご指導いただき、本論文の主査でもある、慶應義塾大学理工学部今井倫太教授に感謝申し上げます。数えきれないほどの議論にお付き合いいただき、もがきながら学んでいった、研究に対するスタンスや価値観は、新たな環境でもなお思考のバックボーンとして活きていることを感じます。

本論文の副査をお引き受けくださり、多くの助言と建設的な意見を頂戴した、慶應義塾大学理工学部斎藤博昭教授、杉本麻樹教授、大澤博隆准教授に感謝申し上げます。

本論文で取り上げた研究は、東京大学山川宏博士、日本大学大澤正彦助教、東京電機大学高橋達二教授と共同で行ったものです。山川博士は、研究構想段階から多大なご指導をいただき、大変な刺激を受けました。大澤助教には、研究室に所属して間もなく、右も左もわからない状態から、面倒を見ていただきました。また、進路や私的な事柄に関しても色々と相談に乗ってくださいました。高橋教授とは「推測されるエージェント」モデルの研究で何度も議論させていただいたことで、論文誌掲載まで辿り着きました。ここに感謝申し上げます。

今井研究室では、多くの同僚に相談に乗っていただき、迷惑もかけました。また、多くの研究に首を突っ込ませてもらいました。お世話になった方は枚挙にいとまがありませんが、木本充彦博士、前川知行氏、岨野太一博士、滝本佑介氏、松森匠哉博士、奥岡耕平氏には特に感謝申し上げたいと思います。また、実験の実施や論文の投稿、会議出席等、研究遂行に伴う無数の事務を処理してくださった、今井研究室秘書の清水奈緒子氏にも感謝申し上げます。

現在の職場の上司である、国立情報学研究所の山田誠二教授は、本論文の執筆に関して深くご理解くださり、寛容に受け入れてくださいました。ありがとうございました。

最後に、ここまでの進学を認め、見守り応援してくださった母と、妻に感謝します。

2023 年 1 月

福地庸介