パターンの発見に基づいた好奇心のモデルに対する主観評価

Estimation of Model's Trait Based on Pattern Discovery via Subjective Human Evaluation

長島一真 [1] *　森田純哉 [2]
Kazuma Nagashima[1], Junya Morita[2]

[1] 静岡大学創造科学技術大学院
[1] Graduate School of Science and Technology
[2] 静岡大学大学院
[2] Graduate School of Integrated Science and Technology

Abstract: 好奇心の研究は，主にエージェントの研究に応用されてきた．しかし，これらの好奇心の振る舞いが，人間が想定する好奇心を正確に表現しているかどうかは示されていない．本研究では，好奇心のモデルの振る舞いを人間の主観評価の結果から検討する．我々は，ACT-R を用いて実装した好奇心のモデルを視覚化し，そして被験者によるモデルの評価実験を実施した．

1　はじめに

HAI (Human-Agent Interaction) に関する研究では，エージェントの性格や印象などの属性帰属に関する研究が行われてきた [24, 25, 26]．しかし，少数の例外 [21] は存在するものの，多くの研究は，見た目や行動などを操作要因とし，エージェントの行動生成の基になる内的要因を直接操作してこなかった．

人の行動からの属性帰属に関しては，機械学習の研究が多々行われている [8, 22]．これらの研究では，機械学習を用いて，人間の振る舞いから人間よりも人間の特性を性格に推定可能であると主張している．しかし，これらの機械学習に関わる研究には，正解ラベルの付与に関する問題がある．性格のラベルは一般的には主観的な観測によって付与されるものであり，個人の内部に存在する真の要因を定義することは難しい．

一方で，属性帰属などに関する社会心理学などの知見では，人間の行動からの特性の推定は不正確であると言われている [19]．人間は行動から自発的に特性を推定するが，多くの帰属エラーが報告されている．特に，人はランダムな行動からでも意図を見出すことがある [11]．また，ランダムな行動と意図を持った行動の区別が困難であることを示す研究もある [26]．

これら属性帰属に関わる問題を踏まえたうえで，本研究では性格を操作した認知モデルに対する属性帰属を扱う．我々はエージェントの内部モデルとして，近年の機械学習エージェントの研究に応用されている「好奇心」に注目した [4]．複数の種類の好奇心を有する認知モデルを用意し，それぞれの行動からの属性帰属を検討する．参加者による帰属の結果と，モデルの予測とのマッチングを検討することで，人間による属性帰属の特性を明らかにすることを目指す．

*連絡先：静岡大学創造科学技術大学院
〒 432-8011 静岡県浜松市中区城北 3 丁目 5 − 1
E-mail: nagashima.kazuma.16@shizuoka.ac.jp

2　関連研究

本研究は，好奇心の内部モデルを持つエージェントの特性を人間の評価から推定可能かどうかを検討する．この目的と関連した研究として，(1) 好奇心の内部モデルに焦点を当てた研究，(2) モデルを持つエージェントの特性の評価に関する研究を紹介する．

2.1　好奇心の内部モデル

好奇心は，教育やエンターテイメントなど，幅広い分野で個人の活動を促進するために重要な内的特性である．近年，好奇心に関する人間の神経科学に基づいた数理モデル [10] が提案され，それを基にした機械学習のエージェントの研究が注目されている [4]．これらのエージェントは，エージェントの学習の「探索（exploration）と活用（exploitation）」の問題に対する一つの解決策を提供し，特定の環境（例：ゲーム）でのパフォーマンスを向上させてきた．

上記の好奇心のエージェントは，内部処理が不透明な深層学習を用いて実装されている．このような研究に対して，著者らはこれまで，人間の脳内で生じるプ

ロセスを対象とした認知モデルの研究領域の中で，好奇心のモデルを研究してきた [23]．認知モデルを用いることで，人間が好奇心を発生させるプロセスをトレースすることが可能になる．

著者らの研究では，好奇心の認知モデルを実装するために，認知アーキテクチャの 1 つである ACT-R ((Adaptive Control of Thought-Rational[2]) を用いた．ACT-R は，個々の認知機能をモジュールという基本的な単位に割り当てる [9]．ACT-R のモジュールは，脳部位と対応し，fMRI を用いた脳計測などによってその挙動に関わる検証が進められている [3]．さらに，モジュールの領域の間のマッピングは，神経科学的な知見に基づいている [20]．

著者らは，ACT-R のモジュールとシンボリックプロセスを用いて，好奇心を「パターンの発見」に基づいて表現した．パターンの発見とは，人間が因果関係のパターンを発見し，組み合わせ，利用する能力のことである [5]．パターンの発見は，導入で述べた好奇心の機械学習モデルで用いられる数理モデルの説明から発展するものである．好奇心は外界の認識と経験から得られる予測との差分によって生じる [10]．この予測からの差分が驚き（好奇心）を生じさせ，そのうちの一部は,「楽しさ」などの感情的反応を引き起こす．そしてその「楽しさ」は新しいパターンを発見することと説明される [14, 18]．

2.2 モデルを持つエージェントの特性の評価

人間の特性は，その行動から推定される．同様に，エージェントの特性も，その行動から理解される．HAI の研究は，伝統的に心的傾向の帰属を促進する振る舞いの特徴，あるいはその要因の検討が行われてきた．その中には，人工物の動作を規定する内部モデルを直接操作する研究も少数ではあるが存在する．

たとえば，Rato らはエージェントの特性が異なる文脈にどのように適応するかを評価した [16]．この研究では，3 次元の仮想空間に配置されたエージェントが，環境の文脈に応じて行動するエージェントとランダムに行動するエージェントの振る舞いを参加者に提示した．そして，参加者に動機づけに関連するアンケートを用いて，モデルの特性を評価した．

また，エージェントの内部モデルが有用に機能しない場合もある．前節で述べたように，好奇心の内部モデルを機械学習エージェントに適用することは特定の環境で有効であることが示されている．人間の研究でも，強すぎる動機づけが特定の行動や一連の行動プロセスに依存する行動嗜癖を引き起こすことが報告されている [1]．好奇心を持つ機械学習エージェントの場合も，環境によって「探索（exploration）と活用（exploitation）」のバランスが崩れ，パフォーマンスが低下することがある．この文脈を踏まえて，Walker らは好奇心の内部モデルを持つロボットを実装し，その振る舞いに知性があるかどうかを人間に評価させた [21]．この研究では，ロボットの振る舞いをビデオに録画し，参加者にオンラインでビデオを提示し，アンケートを用いてロボットの特性を評価した．アンケートの作成には，エージェントの振る舞いから特性を判断するための質問紙の Godspeed の「知性の有無」の項目 [6] が利用された．

以上の研究は，人間が機械学習や計算モデルエージェントの行動の振る舞いからエージェントの特性を部分的に推定できることを示唆している．ただし，これらの研究で扱われている内部モデルと人間の認知機能との対応は明確ではない．したがって，本研究では，著者らが開発した好奇心の認知モデルを採用したうえで，モデルの予測と参加者による属性帰属の関係を検討する．

3 実験

3.1 目的

本研究は，エージェントの振る舞いを通して人間の評価からエージェントの特性を推定可能であるかを検討することを目的とする．本実験では，この目的のために，ACT-R の好奇心のモデルの振る舞いを視覚化し，モデルの振る舞いを再現する 3 次元の仮想空間のシミュレータを開発した．そして，先行研究 [21] を参考に実験のためのウェブサイトを作成しオンライン実験を実施した．

3.2 方法

3.2.1 参加者

参加者はクラウドソーシングサービスのランサーズにて募集した．その参加人数は 100 名であった．そのうち，アンケートの回答の収集において問題が生じた参加者を除外し，95 名の回答について分析を行った．

3.2.2 材料

操作要因となるエージェントの特性として，課題にかける心的リソース（思考水準）の異なる好奇心の認知モデルを合計 3 体用意した．これらは 2.1 に示したものである．以下，モデルが実行したシミュレーション課題を示し，その上で本研究で設定した実験条件と対応する 3 体の認知モデルを示す．本項の最後に，認知モデルを参加者に提示する環境を示す．

図 1: 迷路環境

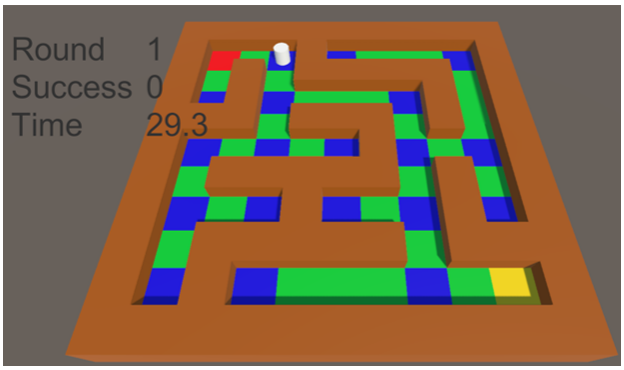

図 2: シミュレータ

シミュレーション課題　本研究が参加者に提示する材料において，モデルは好奇心に基づいて，設定されたマップを探索する，図 1 はモデルが探索する迷路のマップの一例である．マップの広さは，先行研究が対象とした中でも，最も広い 9 × 9 のサイズとした．これらのマップは迷路の曲がり角をノードとするトポロジカルマップであり，ノードとノード間の結合（パス）が知識として表現される．

課題の実行において，モデルは事前に与えられたパスの記憶（2 つの曲がり角とその間の方向の結合）を，モデルの状態とマッチング（パターンマッチング）することで発見する．モデルがスタート位置からゴール位置まで移動するか，または制限時間 (180 秒) に達するまでのプロセスを 1 回のラウンドとし，同一のマップに対する複数のラウンドを，課題全体の制限時間 (3600 秒) に達するか，モデルが課題に「飽きる」まで繰り返す．

モデルの飽きは，2.1 節で述べた著者らの好奇心の認知モデルに基づいて発生する．著者らのモデルにおいて,「飽き」は，知識の圧縮によって，その知識が発見されなくなることで生じると仮定される．このモデルでは，パターンマッチングが発生するとパスの知識と判断のルールが統合（圧縮）され，課題を継続したいという「楽しさ」を感じる．圧縮が発生しないと，ラウンド終了時に課題を継続したくないという「飽き」が生じる．そして最終的に，モデルはラウンドの継続をやめて課題を終了する．

実験条件　上記の設定で刺激されるモデルの好奇心を検討するために，モデルが有するエージェントの特性の条件を思考水準（環境探索の戦略）の観点から操作した．モデルが有する思考水準は，パターンマッチングが発生する量によって区別される．つまり，モデルの特性として思考水準の高いモデルは，戦略を持ち考えて課題を遂行するモデルであり，思考水準の低いモデルは，戦略を持たずに課題を遂行するモデルと仮定される．以下に，モデルの概要を示す．

1. ランダムモデル：ランダムに環境を探索する．
2. 確率的 DFS モデル：確率的な DFS (depth-first search)，あるいはバックトラックによって環境を探索する．
3. 確率的 DFS + IBL モデル：確率的 DFS と IBL (instance-base learning) を組み合わせる．IBL とは，現在の課題の解決に過去の記憶を用いる学習方法である [12, 15].

上記のモデルの思考水準は 1 から 3 の順に高くなると仮定される．1 のモデルはランダムな行動をとるため，次の行き先を考える機会が少なく，思考水準が低いと見なされる．一方，3 のモデルは 2 の戦略に加えて IBL の戦略も含んでおり，より多くの機会で考えることができるため，思考水準が高いと見なされる．

上記の 3 条件について，それぞれ 10 回モデルを実行し，その結果を参加者に対する提示刺激とした．表 1 は，提示刺激となる統計量である．Round Num は課題の継続回数，Goal Rate はモデルがゴールに達した割合，Total Time は，モデルが課題を継続した時間（秒）である．また，Mean はこれらの値の平均値，SE は標準誤差，Min と Max はそれぞれの最小値と最大値である．

エージェント評価用ウェブサイト　先行研究 [21] を参考にして，エージェントの動作を動画に録画し，それに基づいて参加者が主観評価を行うウェブサイトを構築した．

エージェントの動作を示す動画は，Unity で実装されたモデルのログからエージェントの行動を復元するシミュレータによって作成した．図 2 は，そのインターフェイスである．画面には，参加者が現在の状況を把握するために，現在のラウンド数（Round），ゴール達成数（Success），現在時刻（Time）が表示されている．

表 1: モデルの結果

	Round Num				Goal Rate				Total Time			
	Mean	SE	Min	Max	Mean	SE	Min	Max	Mean	SE	Min	Max
Random	33.8	2.62	23	46	0.56	0.05	0.26	0.76	3586.2	13.81	3461.9	3600.0
DFS	9.3	0.98	4	15	0.16	0.03	0.00	0.27	1557.5	501.44	665.9	2499.6
DFS+IBL	6.6	0.65	3	10	0.34	0.11	0.00	1.00	980.7	116.58	329.7	1440.0

また，白い円柱がエージェントを表す．この環境は図 1 のマップと対応し，スタート（赤），ゴール（黄），曲角（青）はトポロジカルマップで繋がっているマスである．エージェントは，これらのマス間を移動する．

この動画に加え，提示刺激の結果を要約した情報（エージェントが迷路を実行した総ラウンド数，ゴール達成率，総課題時間）が提示された．

上記の情報に基づき，参加者はエージェントの属性帰属に関するアンケートに回答した．アンケートは，先行研究 [21] と同様，Godspeed[6] の「知性の有無」の項目を使用した．さらに，好奇心の種類を詳細に検討するために，好奇心の性質を調査するためのアンケートである 5DS[13] を用いた．5DS は 5 つの好奇心のタイプを分類する．その中から本課題に関係のある「Joyous Exploration」と「Deprivation Sensitivity」を用いた．前者は好奇心の喜びや肯定的な経験のために物事を探求に関係があるとされ，後者は問題を解決や知識のギャップを埋めるような知的な物事を探求に関係があるとされる．

5DS のそれぞれの項目は，5 つの質問文で構成される．参加者はその質問に対して同意できるかを 5 点満点で評価した．また，Godsspeed では，知性の有無に関連する 2 つの対立した言葉のセットが 5 つ用意された．参加者は 1 つのセットに対して，エージェントの振る舞いがどちらの言葉に近いのかを 5 段階で評価した（1 には知性がない言葉が配置され，5 には知性がある言葉が配置される）．

なお，本研究では，アンケートに答える参加者が日本人であることを想定した．また，5DS の質問紙が自己回答形式であるため，5DS の質問紙を改変した．まず，参加者がエージェントの振る舞いを評価できるように，質問文の一人称を三人称に変更した（I から He）．その後，その質問紙を，翻訳ツールを用いて日本語に翻訳した．

3.2.3 手続き

ランサーズの募集に応募した参加者には，専用サイトで実験課題に関する説明ページが提示された．以下に，実験の手順を示す．

1. ランサーズの依頼画面にて参加登録
2. 課題説明の提示
3. 3 つの思考水準のモデル分，課題画面にて以下の操作
   (a) 3 分間動画を視聴
   (b) アンケート

参加者が説明を十分に理解したと自己判断した後，課題画面に移動した．表示順は，順序効果を考慮して，参加者ごとにランダムに決定される．各モデルの課題画面では，10 のシミュレーションから無作為に選択された．参加者は動画を 3 分間視聴した後，アンケートに回答でき,「送信」ボタンを押すと次のモデルの課題画面に移動した．

3.2.4 分析

本実験では，好奇心を持つエージェントの行動からモデルの特性を推定するために，人間の主観評価の結果から分析を行う．アンケートで得られた回答について，5DS の 2 項目と Godspeed の 1 項目の平均得点を求めた．先行研究 [13] に従い，この得点を，アンケート項目の最大値（5）に対する割合に変換した．アンケートの分析では，各モデルとこの割合を要因とした 2 要因の参加者内分散分析を行った．

3.3 結果と考察

図 3 は，モデルごとに集計したアンケートの結果である．それぞれのバーの色は，3.2.2 節で示したモデルの思考水準を表し，x 軸の各項目は，5DS（Joyous Exploration と Deprivation Sensitivity）と Godspeed の Perceived Intelligence（知性の有無）の指標を示している．y 軸は，$n = 95$ のアンケート項目に対して得られた評定得点の平均である．

分散分析の結果，モデルの主効果（$F(2, 188) = 24.64$, $p < .01$），指標の主効果（$F(2, 188) = 19.76$, $p < .01$），および交互作用（$F(4, 376) = 6.91$, $p < .01$）が有意となった．本研究において注目するモデルの主効果はいずれの指標においても有意となった (Joyous Exploration: $F(2, 188) = 14.92, p < .01$，Deprivation

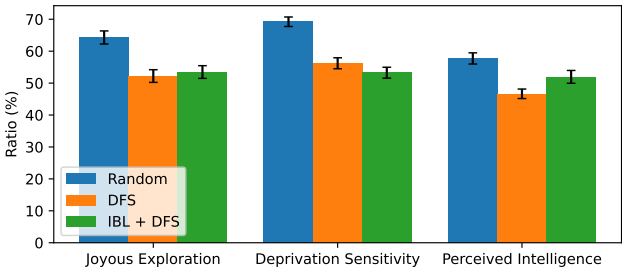

図 3: 結果（エラーバーは標準誤差）

Sensitivity: $F(2, 188) = 38.18$, $p < .01$, Perceived Intelligence:$F(2, 188) = 11.62$, $p < .01$）.

さらに，モデル間の差を明確にするために Holm 法を用いた多重比較を実施した．その結果，思考水準の最も低いランダムモデルがすべての指標で有意に最も高くなった ($p < .05$)．他のモデル間の差異は，Perceived Intelligence において認められ，DFS+IBL モデルが DFS モデルに対して高い結果となった．

以上のように，本研究では，ランダムモデルが最も好奇心を持ち知性的なモデルと評価された．また，他のモデルは 5DS による好奇心の強さが変わらなかったものの，Godspeed においては，DFS+IBL モデルが DFS モデルよりも知的であるという評価が得られた．

上記の結果は，著者らの好奇心のモデルにおける思考水準の設定とは異なっている．よって，本研究の結果は，人間による属性帰属が，モデルの内部状態を推定することがこんなであることを示している．特に，本研究の結果において顕著なのは，参加者が，ランダムなモデルの振る舞いの知的さを過大評価したことにある．参加者がランダムな振る舞いに意図を見出す傾向は指摘されている [11]．また，ランダム性は，問題解決に必要な創造性に近しい振る舞いにつながるという主張する研究も存在する [7, 17]．今後のさらなる検討は必要であるものの，上記のような先行研究の知見より，ランダムモデルの動作に知性を見出す参加者の傾向は解釈できると考えている．

なお，Intelligence の項目において，DFS モデルが他のモデルよりも低い結果となったことについては，このモデルが最も規則性のわかりすい振る舞いをおこなっていたことに由来すると考えられる．

4　まとめ

本研究では，ACT-R の好奇心のモデルを用いて，人間がエージェントの振る舞いからその特性を推定できるかを検討した．この目的のために，それぞれ異なる特性を持つ ACT-R のモデルの動作を復元する仮想の 3 次元空間のシミュレータを実装した．それを用いたオンライン実験では，参加者がエージェントの振る舞いからその特性をどの程度推定できるかを調査した．その結果は，最も思考水準が低いランダムモデルが，最も好奇心を持ち知的な振る舞いをしているという評価であった．これは，エージェントが持つ実際の特性と参加者が評価した特性に差異があることを示すものである．

本研究の意義は，認知モデルを用いて内部モデルを要因として特性の推定を行ったことである．認知モデルは内部プロセスに焦点を当て実装されるため，機械学習エージェントに比べて，人間が認知モデルの特性を推定する際のプロセスをトレース可能であるという利点がある．そのプロセスを解析することで，人間同の特性の推定に応用可能であると考えられる．

今後，このアプローチをさらに進めていく必要がある．今回の研究では，単純な環境においてモデルの振る舞いに焦点を当てて実験を行なった．そのため，参加者はランダム性の高いモデルに好奇心と知性を見出した．このことから，モデルの内面を推定するためには，参加者に提示する情報や環境が不完全であると考えられる．人間がモデルを推定できるようになるためには，モデルの内面に関わる情報の提示や環境の検討を進めていく必要がある．

参考文献

[1] Adam Alter. *Irresistible: The Rise of Addictive Technology and The Business of Keeping Us Hooked*. Penguin, London, 2017.

[2] J. R. Anderson. *How Can the Human Mind Occur in the Physical Universe*. Oxford University Press, New York, 2007.

[3] John R Anderson. Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, Vol. 29, No. 3, pp. 313–341, 2005.

[4] Arthur Aubret, Laëtitia Matignon, and Salima Hassas. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*, 2019.

[5] Simon Baron-Cohen. *The pattern seekers: How autism drives human invention*. Basic Books, New York, 2020.

[6] Christoph Bartneck, TA Cochrane, R Nokes, Geoff Chase, XQ Chen, TT Cochrane, Antonija Mitrovic, AD O'Sullivan, WH Wang, and B Adams. Godspeed questionnaire series: Translations and usage. 2023.

[7] Arthur Cropley. In praise of convergent thinking. *Creativity Research Journal*, Vol. 18, No. 3, pp. 391–404, 2006.

[8] Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine De Cock. Computational personality recognition in social media. *User modeling and user-adapted interaction*, Vol. 26, pp. 109–142, 2016.

[9] Jerry A Fodor. *The Modularity of Mind*. MIT Press, 1983.

[10] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, Vol. 11, No. 2, pp. 127–138, 2010.

[11] Sophie Fyfe, Claire Williams, Oliver J. Mason, and Graham J. Pickup. Apophenia, theory of mind and schizotypy: Perceiving meaning and intentionality in randomness. *Cortex*, Vol. 44, No. 10, pp. 1316–1325, 2008. Special Issue on "Neuropsychology of Paranormal Experiences and Beliefs".

[12] Cleotilde Gonzalez, Javier F Lerch, and Christian Lebiere. Instance-based learning in dynamic decision making. *Cognitive Science*, Vol. 27, No. 4, pp. 591–635, 2003.

[13] Todd B Kashdan, Melissa C Stiksma, David J Disabato, Patrick E McKnight, John Bekier, Joel Kaji, and Rachel Lazarus. The five-dimensional curiosity scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people. *Journal of Research in Personality*, Vol. 73, pp. 130–149, 2018.

[14] Raph Koster. *Theory of Fun for Game Design*. O'Reilly Media, Sebastopol, 2013.

[15] Christian Lebiere, Cleotilde Gonzalez, and Michael Martin. Instance-based decision making model of repeated binary choice. In *Proceedings of the 8th International Conference on Cognitive Modelling*, pp. 67–72, 2007.

[16] Diogo Rato, Marta Couto, and Rui Prada. Fitting the room: Social motivations for context-aware agents. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, pp. 39–46, 2021.

[17] Mark A Runco and Garrett J Jaeger. The standard definition of creativity. *Creativity Research Journal*, Vol. 24, No. 1, pp. 92–96, 2012.

[18] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, Vol. 2, No. 3, pp. 230–247, 2010.

[19] Charles G. Stangor and Jennifer Walinga. *Introduction to Psychology - 1st Canadian Edition*. BCcampus, Victoria, 2014.

[20] Andrea Stocco, Catherine Sibert, Zoe Steine-Hanson, Natalie Koh, John E Laird, Christian J Lebiere, and Paul Rosenbloom. Analysis of the human connectome data supports the notion of a "Common Model of Cognition" for human and human-like intelligence across domains. *NeuroImage*, Vol. 235, p. 118035, 2021.

[21] Nick Walker, Kevin Weatherwax, Julian Allchin, Leila Takayama, and Maya Cakmak. Human perceptions of a curious robot that performs off-task actions. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 529–538, 2020.

[22] Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, Vol. 112, No. 4, pp. 1036–1040, 2015.

[23] 長島一真, 森田純哉, 竹内勇剛. ACT-R による内発的動機づけのモデル化. 人工知能学会論文誌, Vol. 36, No. 5, pp. AG21–E_1-13, 2021.

[24] 石川幸太郎, 飯野直樹, 磯部光裕, 中島亮一, 大澤正彦. パーソナリティ特性に基づく球体の動きに感じるアニマシーの分析. HAI シンポジウム 2022, 2022.

[25] 竹内勇剛, 中田達郎. エージェンシー認知を誘発するコンピュータとのインタラクションと人らしさの帰属. 人工知能学会論文誌, Vol. 28, No. 2, pp. 131–140, 2013.

[26] 細川敦司, 森田純哉. カードゲームにおいて模倣するエージェントへの心的状態の帰属. HAI シンポジウム 2022, 2022.