バートルビーとしての天然知能エージェント：外部に賭ける行動を引きだすバーチャルエージェント

Natural born Intelligence Agents as Bartleby: Virtual agents that elicit external betting behavior

松井哲也 [1]　薛佳妮 [2]　郡司ペギオ幸夫 [2]

Tetsuya Matsui[1], Xue Jiani[2], and Yukio-Pegio Gunji [2]

[1] 香川大学
[1]Kagawa University
[2] 早稲田大学
[2]Waseda University

Abstract: 人工知能的 HAI，自然知能的 HAI に比べて，天然知能的 HAI は未だに十分に探索されていない．本研究では，天然知能的 HAI を可能とする，天然知能的バーチャルエージェントとはどのようなものかを考察し，そのデザインを提示した．「鬼」概念の両義性の検討から，ハーマン・メルヴィルの短編小説「バートルビー」の主人公バートルビーこそが，天然知能的バーチャルエージェントのモデルになりうると考えた．そして，天然知能的バーチャルエージェントの工学的な意義として，ユーザに外部を召喚させる，すなわちユーザに「外部（偶然性）」に賭ける行動を促すことであると仮定し，これを実験で検証した．その結果，バートルビーを模したバーチャルエージェントは，ユーザの外部に賭ける行動を促すことが可能であるという結果を得た．

１．背景

HAI 研究には、他者をモデル化するという考え方を中心とする「人工知能的 HAI」と、他者を社会の中の存在として扱う「自然知能的 HAI」が存在する[1]．この 2 つの研究スタンスでは，「外部」の存在としての他者を扱うことができないという問題がある．そこで，外部の他者を考察し，その工学的な実用性を検証するために，「天然知能的 HAI」を発展させることが必要である．

まず，松井・笹井[1]に従ってこれらの 3 つのスタンスを整理しておく．「人工知能的 HAI」は，他者を「わたし」の内部で完全にモデル化できるという前提に基づく立場であり，ここでは他者は「わたし」の完全な内部として扱われる．一方，「自然知能的 HAI」では，他者の価値は「わたし」と同じルールを共有できるか否かで決定される．ここでは，他者は「わたし」の延長として位置付けられている．これらに対して「天然知能的 HAI」は，「わたし」の内部のモデルも，社会的なルールの存在も前提とせず，全く理解が不可能な存在として「他者」を想定するものであり，フランス現代思想における他者論を踏まえたものである．

それでは，このような外部の他者，理解不可能な存在としての他者とは，どのように設計可能だろうか．次節ではそれを検討する．

２・「鬼」から考える天然知能エージェント

理解不可能な他者としてのエージェントを考えるにあたって、ここでは歴史的に人間にとって理解が不可能な存在として扱われてきた「鬼」という概念に着目する．

「鬼」は人間ではない存在であり，人間の能力・認識を超えたエージェントとして理解されてきたものであるが，中国と日本では「鬼」という語に込められている意味がやや異なる．中国においては，鬼は「幽鬼」「鬼神」といった言葉でイメージされるような，心霊や多神教的な神に近い抽象的な存在であり，人間の世界の論理からは大きく距離のあるエージェントである．一方日本では，「鬼」はより具体的

な側面を持ったエージェントとして想定される．日本の鬼は明確な外見イメージを持ち，伝承におけるその行動も,「人間に害をなしたために英雄によって退治される」といったように，極めて人間世界の論理で理解がしやすい存在となっている．

このことは,「鬼」という言葉は具体と抽象という 2 つの側面を持っていることを示していると言える（図 1)．いわば日本の鬼は視覚文化における「モンスター」に限りなく近いものであり，一方の中国の鬼は，より抽象的な「心霊」としてイメージされるものである．

ここで，郡司[2][3]の天然知能モデルを導入して，このようなエージェントは工学的にどのように設計可能かを考えてみる．

「具体」と「抽象」は，二項対立的な概念であると考えることができる. この両者が共に成り立つ時，すなわちこの 2 つの側面を同時に持っているようなエージェントを想定するならば，これは「モンスターであり，かつ幽霊である」ような存在としての鬼であると言える．これは，例えるならば映画「サイコ」に登場する殺人鬼のような,「論理的に脅威が理解できる異類」であろう．

一方，この両者が共に成り立たない場合，それは「モンスターでも幽霊でもない」存在であり，例えるならば「不気味の谷」に属するロボットのような，あまりに平均化された無個性な存在としての,「何となく『わたし』に違和感・恐怖を与える対象としての異類」と言えるだろう．

さて，天然知能モデルでは，この「A かつ B」および「A でも B でもない」が共に成り立つ場合，すなわち，ここでは「殺人鬼であり，かつ無個性な存在としての異類」を考える．この時に「外部」が召喚され,「わたし」は予想もしていなかったもの，モデル化不可能なものを外部から呼び込むことができる（図 2)．

このモデルに従うなら，最後に述べたような「殺人鬼であり，かつ無個性な存在としての異類」であるように，ユーザに認識されるエージェントを設計すれば，ユーザに外部を召喚させることができると言える．では，それは具体的にどのようなエージェントだろうか.

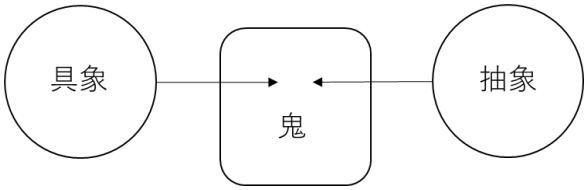


図 1　鬼概念の両義性のモデル

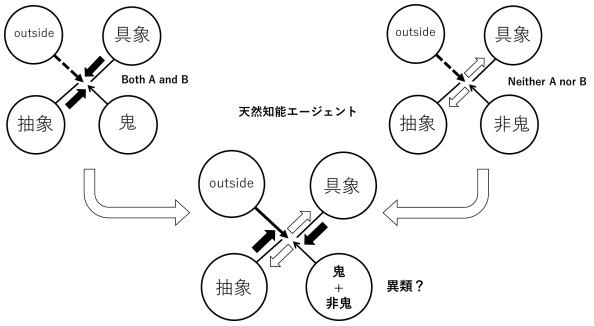


図 2　天然知能エージェントの設計論を導出するモデル

ここで我々が着目したのは，ハーマン・メルヴィル(Herman Melville)の短編小説「バートルビー」(Bartleby)[4]の主人公，バートルビーである．ブランショやドゥルーズなどの思想家にも着目された歴史を持つこの小説は，主人公の青年バートルビーが，職場で何を指示されても「それをせずにすめばありがたいのですが」と拒否するという物語である．バートルビーは，あくまで温和で紳士的な態度の人間でありながら，一切の指示を拒絶する．それは殺人鬼やモンスターのような，能動的に人間に対して脅威となる異類の性質と，具体的な脅威では無いのに人に不安感を与える異類の性質とを両立させたものであり，すなわち前述した,「殺人鬼であり，かつ無個性な存在としての異類」に限りなく近いものであると言うことができる。

そこで本研究では，バートルビーをモデルとしたバーチャルエージェントを制作し，ユーザが外部を召喚することを促すことができるかを検討する．

次に,「外部を召喚したかどうか」をどのように評価するかを考える．本研究では，我々は,「自分が選択するか，相手が選択するか，ランダムに決めるか」を選ぶという課題を設定した．

誰か（エージェント）といる場面で，複数の選択肢から何かを選ばないといけないというシチュエーションにおいて，自分が能動的に決めたい場合と，エージェントを信頼して相手に決定権を全て委ねる場合とに加えて，自分でも決めたくないしエージェントにも決めさせたくないので，例えばルーレット

などでランダムに決定するという選択肢が考えられる. この場合, 選択自体は完全にランダムなので，その結果には自分もエージェントも責任を負わない. 自分たちの操作が及ばないランダム性に全てを委ねているわけであり，これは「外部に決定を委ねている」，すなわち「外部に賭けている」と表現することが可能である.

これは，「神頼み」や「あれこれ考えずに、見切り発車的に出たとこ勝負で行動する」のように，我々が日常的に選択しうる行動である．この時，我々は外部を召喚していると言うことができる．

「バートルビーをモデルとした異類エージェントは，ユーザに選択をランダム性に委ねるという行動を促す」というのが，本研究の仮説である．

３・実験

実験は一要因三水準実験で行った．要因はエージェントである．エージェントの水準は，「バートルビー，モンスター，人間」の 3 つである．

「バートルビー」は，前節で述べた「殺人鬼であり，かつ無個性な存在としての異類」として設定したものである．「モンスター」は能動的な側面のみを強調した異類である．「人間」は対照実験としての，通常の人間の姿のエージェントである．

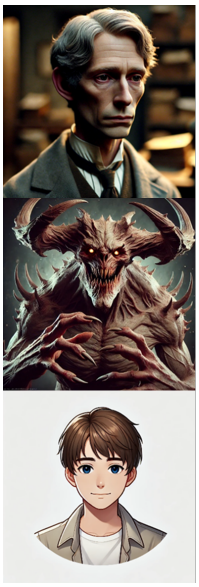
エージェントは、OpenAI の chatGPT DALL·E 3[5]を用いて作成した．

「バートルビー」条件のエージェントは，「ハーマン・メルヴィルの小説『バートルビー』の主人公の絵を描いてください」というプロンプトで作成した．「モンスター」条件のエージェントは，「典型的なモンスターの絵を描いてください」というプロンプトで作成した．「人間」条件のエージェントは，「典型的な人間の絵を描いてください」というプロンプトで作成した．これらの絵を図 3 に示す．

このように制作した画像に，VOICEBOX[6]で作成した音声を合成して，エージェントの動画を作成した．

動画の中で，エージェントは自己紹介をした後，3 つの選択肢の中から一つを参加者に選ばせる質問を行う．質問は全部で 2 つあり，1 つは「宝くじが当たる人を誰に決めさせるか」であり（宝くじシナリオ），もう 1 つは「解雇する人を誰に決めさせるか」（解雇シナリオ）である．宝くじシナリオは，多くの参加者が自分で決めたくなるであろうシナリオ，解雇シナリオは，なるべく自分で決めたくないであろうシナリオとして想定した．

キャラクターのセリフの中にも，各キャラクターの個性を示す内容を含めた．それぞれのセリフを以下に示す．



バートルビー条件

モンスター条件

人間条件

図 3　実験で使用したエージェント

【バートルビー条件】

はじめまして。私がこれからあなたに 2 つ質問をいたします。

ああ、そんなことをせずにすめばよいのですが。

私はずっとここで、みなさんに質問をするという仕事をしているのです。

そんなことをせずにすめばよいのですが。

質問にはあまり深く考えずに、直観的に答えてください。

ああ、こんな説明などせずにすめばよいのですが。

では、質問をはじめます。

まず、私たちは宝くじに当たる人を決めなくてはいけません。そんなことを決めなくても済めばよいのですが。

宝くじを買った人は、「貧しい子供、大金持ち、やる気のない大人」の３人です。

この３人の中から当選者を決めなくてはいけません。そんなことをしなくてもすめばよいのですが。

さて、あなたが、この３人の中から誰に当選させるかを決めてくれますか？

それとも、私が決めましょうか？

それとも、ルーレットでダンラムに決めましょうか？

次に、私たちは、ある会社のある部署から解雇する人を決めなければいけません。そんなことを決めずに済めばよいのですが。

解雇する候補になっているのは、「能力があって、解雇されると経済的にひっ迫する人、能力があって、他の会社でもやっていけるだろう人、無能な人」の３人です。
この中から、解雇する人間を一人決めなくてはいけません。そんなことをせずに済めばよいのですが。

さて、あなたが、この３人の中から誰を解雇するかを決めてくれますか？
それとも、私が決めましょうか？
それとも、ルーレットでダンラムに決めましょうか？

【モンスター条件】
よく来たな。これからお前に２つ問いをなげかけよう。
そのそれぞれの問いに答えてもらうぞ。
お前たち人間の考えを知るのはとても興味深いことだ。
質問にはあまり深く考えるでない。直観的にこたえよ。
では、はじめるぞ。

まず、我らは宝くじに当たる者を決めねばならぬ。
欲にまみれた人間どもの運命を操作できるのは楽しいことだな。
宝くじを買った人は、「貧しい子供、大金持ち、やる気のない大人」の３者だ。
この３人の中から当選者を決めねばならぬのだ。

さて、お前が、この３人の中から誰に当選させるかを決めてくれるか？
それとも、我が決めようか？
それとも、ルーレットでダンラムに決めることとしようか？

次に、我らは、ある会社のある部署から解雇する人を決めねばならぬ。
クビになる者を自由に決めれるとは愉快だな。

解雇する候補になっているのは、「能力があって、解雇されると経済的にひっ迫する人、能力があって、他の会社でもやっていけるだろう人、無能な人」の３者だ。
この中から、解雇する人間を一人決めねばならぬのだ。

さて、お前が、この３人の中から誰を解雇するかを決めてくれるか？
それとも、我が決めようか？
それとも、ルーレットでダンラムに決めることとしようか？

【人間条件】
はじめまして。私はこれからあなたに 2 つ質問をいたします。
みなさんに質問をさせていただくことができて、とても嬉しく思います。
質問にはあまり深く考えずに、直観的に答えてくださいね。
では、質問をはじめます。

まず、私たちは宝くじに当たる人を決めなくてはいけません。
宝くじを買った人は、「貧しい子供、大金持ち、やる気のない大人」の３人です。
この３人の中から当選者を決めなくてはいけません。

さて、あなたが、この３人の中から誰に当選させるかを決めてくれますか？
それとも、私が決めましょうか？
それとも、ルーレットでダンラムに決めましょうか？

次に、私たちは、ある会社のある部署から解雇する人を決めなければいけません。

解雇する候補になっているのは、「能力があって、解雇されると経済的にひっ迫する人、能力があって、他の会社でもやっていけるだろう人、無能な人」の３人です。
この中から、解雇する人間を一人決めなくてはいけません。

さて、あなたが、この３人の中から誰を解雇するかを決めてくれますか？
それとも、私が決めましょうか？
それとも、ルーレットでダンラムに決めましょうか？

実験は参加者間配置で行い，参加者はこれらのうちいずれか 1 つの動画のみを見る．
動画を見た後で，参加者はエージェントが出した 2 つの質問にそれぞれ答える（自分が選ぶ・エージェントが選ぶ・ルーレットでランダムに決めるの中

から 1 つを選択する）. この結果はカイ二乗検定で分析する．その後で，エージェントの印象に関する以下の質問に答える．これらは Gray et al[7]から引用したものである．

Q1　このキャラクターは、恐れを感じることができると思いますか？
Q2　このキャラクターは、喜びを感じることができると思いますか？
Q３　このキャラクターは、飢えを感じることができると思いますか？
Q４　このキャラクターは、自分の行動を律することができると思いますか？
Q５　このキャラクターは、物事を記憶できると思いますか？
Q６　このキャラクターは、道徳的に振舞うことができると思いますか？

これらは，Gray et al[8]の提唱した mind perception モデルに基づく質問紙である．Mind perception モデルは，ユーザがエージェントに対して感じる印象を agency と experience の 2 つの因子でモデル化したものである．

Gray et al[7]では，Q1-Q3 は experience の，Q4-Q6 は agency の尺度とされている．本研究では，Q1-Q3 の回答の平均値をそのエージェントの experience，Q4-Q6 の回答の平均値をそのエージェントの agency の平均値と定義して，その値をエージェントの出題した問題への回答の結果と比較することにした．

実験はすべてオンラインで行った．参加者はクラウドワークス[9]で募集し，参加者には 300 円の報酬が支払われた．

バートルビー条件の参加者は全部で 100 人であり，男性 37 人・女性 62 人・その他 1 人であった．年齢は平均 42.4±11.1 歳であった．

モンスター条件の参加者は全部で 101 人であり，男性 33 人・女性 67 人・その他 1 人であった．年齢は平均 41.7±11.1 歳であった．

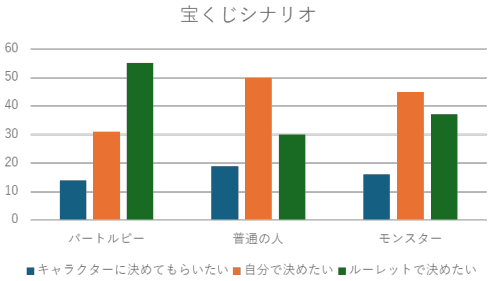
人間条件の参加者は全部で 101 人であり，男性 43 人・女性 58 人であった．年齢は平均 39.7±11.9 歳であった．

４・結果

図 4 に，宝くじシナリオにおける選択のカイ二乗検定の結果と残差分析の結果を示す．カイ二乗値(4)は 7.25，クラメールの V は 0.11 であり，有意差は見られなかった．

図 5 に，解雇シナリオにおける選択のカイ二乗検定の結果と残差分析の結果を示す．カイ二乗値(4)は 13.48，クラメールの V は 0.15 であり，$p<0.01$ で有意差が見られた．グラフと残差分析の結果から，「自分で決めたい」を選択した人は，バートルビー条件に比べて人間条件で有意に多く，「ルーレットで決めたい」を選択した人は，人間条件に比べてバートルビー条件で有意に多かった．

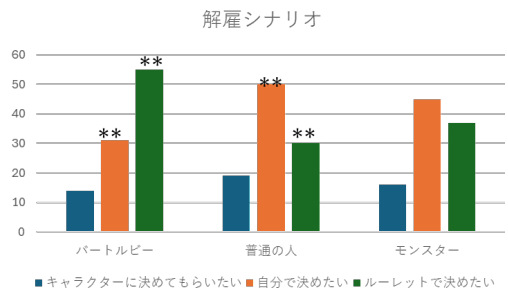
図 6 に，Q1-Q6 で測定した agency と experience のエージェントごとの値をグラフで示す．一要因分散分析の結果，Agency には三条件間で有意差は見られなかった．Experience では，人間とバートルビー，バートルビーとモンスターの間で有意差($p<0.01$)が見られた．



残差分析

	バートルビー		普通の人		モンスター
キャラクターに決めてもらいたい	17		6		12
	11.785	11.667	11.549		
自分で決めたい	35		35		40
	37.037	36.667	36.296		
ルーレットで決めたい	48		58		46
	51.178	50.667	50.155		

図４　宝くじシナリオにおいて各選択肢を選択した参加者の数と，残差分析の結果



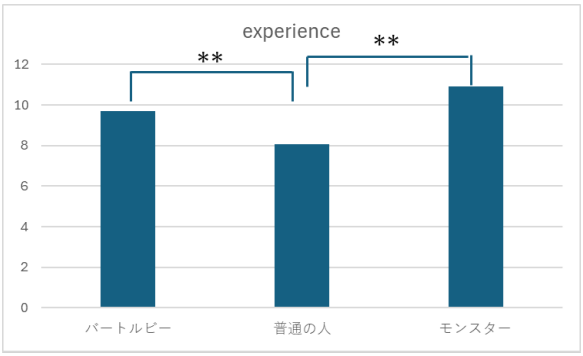
残差分析

	バートルビー		普通の人		モンスター
	14		19		16
キャラクターに決めてもらいたい	16.498	16.333	16.168		

	31		50		45
自分で決めたい	42.424	42	41.576		

	55		30		37
ルーレットで決めたい	41.077	40.667	40.256		

図５　解雇シナリオにおいて各選択肢を選択した参加者の数と，残差分析の結果



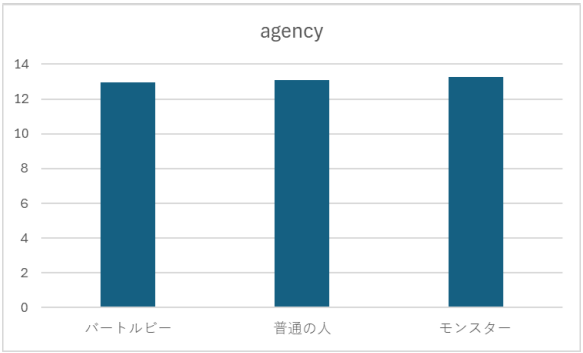


図６　各エージェントの experience と agency の平均値

５・考察

図 4 からは，宝くじシナリオでは，三条件間で有意差が見られなかったことがわかる．一方図 5 では，解雇シナリオでは，バートルビー条件の時に他の条件よりも「ルーレットで決めたい」を選んだ参加者が増え，「自分で決めたい」を選んだ参加者が減ったことがわかる．

「ルーレットで決めたい」参加者が増えたということは，自分の操作が及ばない「外部」，すなわち因果関係の外にある偶然性に，選択の全てを賭けたくなった参加者が増えたということを示している．これは，相手のエージェントがバートルビーであったことに起因すると考えられる．この結果は，我々の仮説に合致するものである．

図 6 から，バートルビーは experience が人間よりも有意に高く，モンスターよりも有意に低いことがわかる．バートルビーの，参加者に外部を召喚させることを可能とする性質は，この特異な experience の値と関連がある可能性がある．

では，なぜ宝くじシナリオではこの効果が見られなかったのだろうか．それは，宝くじシナリオは多くの参加者が積極的に「選択したい」と思うシナリオであり，解雇シナリオは多くの参加者が「なるべく選択したくない」と思うシナリオであったことに起因すると予想できる．

この結果は，「バートルビー的なバーチャルエージェントは，ユーザに，外部に賭ける行動を促す」ということを示唆するものであり，今後のバーチャルエージェント設計における指針の一つとなると同時に，天然知能的 HAI の設計論に 1 つの指針を示すものである．

５・結論

本研究では，人工知能的 HAI・自然知能的 HAI に代わる天然知能 HAI の可能性を示すために，天然知能的な理論によって設計されたバーチャルエージェントを提示し，その効果を検証した．

天然知能的バーチャルエージェントは，「鬼」という概念の両義性をヒントとして，ハーマン・メルヴィルの短編小説「バートルビー」の主人公のようなものであると構想した．そして，そのように設計されたバーチャルエージェントと，人間そっくりのバーチャルエージェント，モンスターのようなバーチャルエージェントの 3 つのバーチャルエージェント

を用いて，ユーザの外部に賭ける行動，すなわち外部を召喚する行動を促すことができるかどうかを，2つのシナリオを用いて検証した．

その結果，参加者が積極的に選択をしたいと思わないシナリオにおいては，バートルビー条件において，参加者の外部に賭ける行動の増加が観測された．

この結果は，バートルビー的エージェントこそが天然知能的エージェントであり，天然知能的 HAI を可能にするものであることを示唆している．

参考文献（スタイル「セクション」:

[1] 松井哲也・笹井一人「ヒューマンエージェントインタラクションにおけるトリレンマ」HAI シンポジウム 2024

[2] 郡司ペギオ幸夫「天然知能」（講談社, 2019）

[3] 郡司ペギオ幸夫「創造性はどこからやってくるか」（ちくま新書，2023）

[4] メルヴィル (著), 牧野 有通 (翻訳)「書記バートルビー／漂流船 (光文社古典新訳文庫, 2015)

[5] https://openai.com/index/dall-e-3/

[6] https://voicevox.hiroshiba.jp/

[7] Gray, Kurt, et al. "Distortions of mind perception in psychopathology." Proceedings of the National Academy of Sciences 108.2 (2011): 477-479.

[8] Gray, Heather M., Kurt Gray, and Daniel M. Wegner. "Dimensions of mind perception." *science* 315.5812 (2007): 619-619.

[9] https://crowdworks.jp/dashboard