社会的承認によって定義された心がある AI：評価方法と有効性の基礎検討

飯田 愛結 [1]　大澤 正彦 [1]
[1] 日本大学 文理学部
chay21052@g.nihon-u.ac.jp　osawa.masahiko@nihon-u.ac.jp

概要

「"それ"であると認められることをもって、"それ"の定義を満たす」とする社会的承認という方法で定義された「心がある AI」の実現を目指す．そして人と AI 自身が「心がある」ことを認める AI のつくりかたを考える．本論文では，社会的承認による定義やこの場合の評価方法の実現性，作成する人工物の有効性を議論する端緒を形成することを目的に，簡単な発話応答に対する評価を実施した．実験ではダニエルデネットの提唱する「設計スタンス」「意図スタンス」の考え方に基づく，設計発話/応答と意図発話/応答の組み合わせを作成し印象を調べた．

1　はじめに

「心がある AI を実現する」などと謳った論文は，だいたい怪しい．そもそも多くの人々が「ある」と信じている心を，誰も実際に見たことも触れたこともなく，何ができれば心ができたことになるのかを工学的に定義できている状況にない．そのような中で本論文は，心がある AI を実現することを目指す．

「どのような機能が実現できれば、それができたといえるのか」を定める方法を機能要件集合による定義と呼ぶ [1]．これまで心を工学的に定義できていなかったことは，機能要件集合を定められていなかったと言い換えられるだろう．ところが，我々は常に物事の定義を機能要件集合によって行っているわけではない．例えば自分にとっての「友達」の定義を要件を列挙して定めていることは珍しく，「(お互いに）友達だと思ったら，友達」といったおおらかな定義をしていると考えられる．著者らはこのように「"それ"であると認められることをもって，"それ"の定義を満たす」といった定義を社会的承認による定義と名付けた [1, 2]．本論文では，社会的承認によって定義された心がある AI について検討する．

2　社会的承認によって定義された心

2.1　機能要件集合による定義

機能要件集合による定義 [1] を以下に示す．

> **機能要件集合による定義**：人工物がラベル Y の名で呼ばれるための機能要件集合 $\boldsymbol{R}^Y$ を Y の定義と呼ぶ．ある人工物 x が機能要件集合 $\boldsymbol{R}^Y$ に含まれる全ての機能を実現しているとき，x は Y である．

心の機能要件集合 $\boldsymbol{R}^{心}$ を決めれば，要件を 1 つずつ実装していくことで心の実現に近づく．しかし，広く合意される $\boldsymbol{R}^{心}$ を設定することは難しい．

2.2　社会的承認による定義

社会的承認による定義（[1] を改変）を示す．

> **社会的承認による定義**：ある人工物 x が Y であるという信念を，全ての主体 $\forall s$ が持ったとき，x は Y である．

信念とは「その主体が認識している世界の情報や知識」である．ここでは全ての主体から認められることを条件としているが，現実的には難しい．そこでいくつかの緩和方策がありうる．

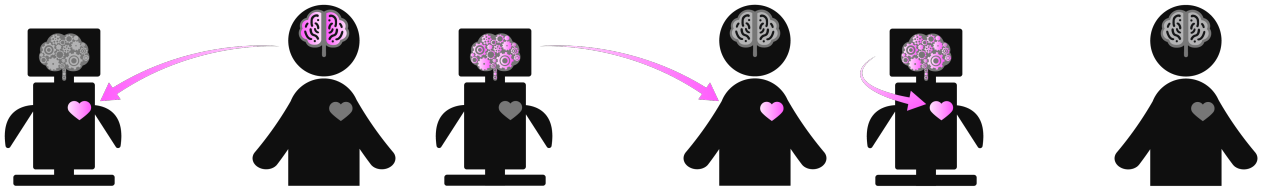
例えば，時期を緩和する方法である．

> **社会的承認による定義'**：全ての主体 $\forall s$ が，x とインタラクションを開始した後のある時刻 $T(T > 0)$ 以降において x が Y であるという信念を持つならば，x は Y である．

コミュニティを制限する方法もありうる．

> **社会的承認による定義''**：ある人工物 x が Y であるという信念を，あるコミュニティ $\boldsymbol{c}$ に所属する全ての主体 $\forall s \in \boldsymbol{c}$ が持ったとき，x は $\boldsymbol{c}$ において Y である．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).



STEP1: 心があるという信念を人間に持たれる AI を作成し，人間がその信念をもつ認知プロセスをモデル化する．

STEP2: STEP1 のモデルを AI に実装し，人間に心があるという信念をもつ機能を実現する．

STEP3: STEP2 の実装を転用して，AI 自身に心があるという信念を持つ機能を実現する．

図 1　筆者の提案する"心がある AI"のつくりかた [3, 4]

2.3　心を感じる主体と認知プロセス

社会的承認による定義を用いれば, 多くの主体に「心がある」という信念を持たれることが重要となる．ここでその信念をもつ主体について考える．人間が信念を持てば良いのなら，AI が人間から見て心があるかのような振る舞いさえできれば，その AI は心があることになる．このとき，AI の実装方法は一切問わない．しかし AI 自身もその主体に含めると，妥当な方法で AI が AI 自身に心があるという信念をもつための実装が必要になる．そこで，図 1 のような STEP1 から STEP3 で社会的承認による定義をされた心がある AI のつくりかたを提案した [3, 4].

心があるという信念を持っている状態とは具体的にどのような状態であろうか．本稿では，Human-Agent Interaction(HAI) 研究の中で頻繁に扱われてきた，ダニエルデネットの提唱する「意図スタンス」[5] で他者の振る舞いを予測・解釈している状態がこれに共通する要素があるのではないかと考えて議論を進める．意図スタンスとはある主体の行動や振る舞いをその主体の意図に基づいて予測・解釈することである．さらに意図は信念や願望といった心的状態から形成されると考えられてきた [6, 7]. 意図スタンスの他に，主体をその設計や振る舞いのルールに基づいて予測・解釈する「設計スタンス」がある．

著者らは意図スタンスで他者の振る舞いを予測・解釈するモデルとして，他者モデル [8] が有望と考えてきた．他者モデルは人の認知プロセスをモデル化した認知モデルの一種であり，他者の心的状態や行動の予測/解釈モデルである．同様に，自己の心的状態や行動の決定/解釈モデルを自己モデルと呼ぶ．つまり図 1 における STEP1 でモデル化して STEP2 で実装するのが他者モデルに，STEP3 で実現するのが自己モデルにそれぞれ対応する．

他者モデルを転用して自己モデルを構成することは，我々人間が普段行う思考に照らし合わせてみても納得がいくものである．例えば，「自分はこう思うから，きっと相手もこうだろう.」「あの人がこうしてうまくいったから，自分も真似してみよう」など，自己と他者が共通する認知プロセスを持っていることを前提にインタラクションしている．

特に人間以外の動物とのインタラクションではこれが顕著である．例えば我々人間は犬や猫に対して「喜んでる」「怒っている」などの心的状態を想定しながらインタラクションしているが，本来は人間と大きく異なる脳のアーキテクチャをもつ動物は，人間と全く異なる心的状態を持っているはずである．ところがいくら長期的なインタラクションを重ねても，我々が自分と全く違う認知プロセスを獲得することはなく，ある程度人間と同じアーキテクチャを持つと仮定することで「人間（私）で言うところの、こういう状態のはず」という想定の仕方をしていると捉えられるだろう．人間と動物の異種間インタラクションと同様に，人間と AI という異種間インタラクションでも「ある程度同じ認知プロセスである」という仮定が重要となる可能性は十分にある．

2.4　人間同士のインタラクションの分類

人間同士のインタラクションは，意図スタンスに基づいて行われるのが通例であるが，身近な例に設計スタンスでのインタラクションもある．例えば，上司からの指示に機械的に従っている場合や，組織内のルールに基づいて行動している場合，その行動が習慣化している場合などである．

本論文では，このように設計スタンス的に解釈できる発話と応答を設計発話・設計応答とそれぞれ呼ぶ．また，主体の信念や願望から形成される意図に基づいた発話と応答を，意図発話・意図応答それぞれと呼ぶ．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1 シチュエーション．設計発話/応答は指示・ルール・習慣に基づく．意図発話/応答は自身の願望・信念に基づく．

ID	人間の発話			AI の応答		
	設計発話	意図発話 (願望)	意図発話 (信念)	設計発話	意図発話 (願望)	意図発話 (信念)
1	出勤しなくて良いと言われたので，	落ち着いた環境で作業したいし，	テレワークができる環境なので，	対面で行う会議が入ったので，	あなたの上司に進捗を確認してもらって欲しいし，	明日は上司も出勤するので，
	家で作業します．			出勤してください．		
2	チームの方針なので，	自分の時間を大切にしたいし，	予定があるので，	緊急時では通常の方針とは異なるため，	後であなたが無理することになってほしくないし，	多少作業を進める時間もあるはずなので，
	週末は作業を進めません．			週末に作業を進めてください．		
3	夕飯の時はビールをいつも飲むので，	夕飯ではビールを飲みたくなるし，	昨日買ってあるので，	明日は健康診断なので，	あなたには健康でいてほしいし，	最近飲みすぎているので，
	今日もお酒を飲みます．			今日はお酒は飲まない方がいいでしょう．		

2.5 問い

社会的承認によって定義された心を持った AI を実際に開発していくために，2 つの問いを立てる．

問い 1 社会的承認によって定義された心を持った AI が実現したかを，どのように評価するか．

問い 2 図 1 で示した方法で作成した AI の有効性はどのような点にあるか．

問い 1 に関して，愚直な方法は「この AI に心があると感じますか」という直接質問で評価することである．著者らは「心があるという信念を持っている状態」が「意図スタンスで他者を予測・解釈している状態」と共通すると考えたが，一般の評価者の「心を持った AI」の定義や感覚はそれぞれ異なる．従って直接質問を利用することによる問題点が浮上すると思われるが，どのように顕在化するかは詳しく分析する必要がある．

問い 2 に関して，AI として実装すべき必要な要件が新たに生まれている一方で，この実装がどのように人間や AI 自身に影響を及ぼすかは不明である．本論文では，「人間は，自分と同じ認知プロセスによって AI が発話している感じると，その AI に心があると感じやすい．」という仮説を立てる．もしこの仮説が正しければ，図 1 で示した方法で開発を進めることで，より人間から心があると感じてもらいやすくなるため，有効性があるといえるだろう．

次章では，以上の議論を具体的な実験に落とし込んでいく手がかりを掴むために，単純な問題設定での評価実験を計画する．

3 評価方法の検討

3.1 実験条件

実験に利用したシチュエーションを表 1 に示す．各シチュエーションには 2.4 で定義した設計発話/応答と意図発話/応答が含まれる．実験では，3 つのシチュエーションそれぞれに以下に示す 4 種類の発話応答の組み合わせで 12 種類の発話応答を評価した．

1. 人間の設計発話に対して，AI が設計応答をする
2. 人間の設計発話に対して，AI が意図応答をする
3. 人間の意図発話に対して，AI が設計応答をする
4. 人間の意図発話に対して，AI が意図応答をする

もし，人間と同じ認知プロセスであることが重要であるとすれば，2. よりも 4. の方が心があると感じられる程度が高く評価されると予想される．また，AI が設計応答する場合は，人間が意図発話した 3. よりは，設計発話した 1. の方が心が感じられないというネガティブな効果が少ない可能性がある．

本実験の参加者は 21 歳〜24 歳（平均 21.62 ± 0.96 歳）の 13 名（男性 7 名，女性 6 名）である．質問は「人間の発話と AI の応答をよく読んだ上で、応答した AI をどの程度心のある人間にちかい存在と感じたか，もしくは心のないただの機械のような存在と感じたか，7 段階で評価してください」という 1 問のみである．参加者は 12 件すべての発話応答に対し「1：心のないただの機械のような存在に思う」から「7：心のある人間に近い存在に思う」の中から評価した．加えて，任意の自由記述欄を設けた．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

3.2 実験結果・考察

各シチュエーションの実験結果を図 2，3，4 に示す．3 つのシチュエーションにおいて，設計応答よりも意図応答の方が心があると感じられる程度が高いという共通の傾向が見られた．一方で，どの発話に対する応答かによる差は顕著には表れておらず，本実験の仮説を支持する結果とはいえなかった．特に ID2 のシチュエーションでは意図発話に対する意図応答の方が，設計発話に対するものよりも評価が比較的大幅に低く，仮説とは逆の結果となった．

自由記述欄では，願望・信念から意図が生成されるいわゆる BDI ロジックに対して，（理系的の詰め方ではあるものの）機械的な印象を持ったというコメントがあった．本実験では意図発話/応答を願望と信念から生成されるものとしたが，表現が回りくどく機械的な印象を与えてしまった可能性もある．また，意図発話/応答では願望と信念の両方を提示したため，設計発話/応答よりも文章が長いという問題がある．さらに信念という概念自体は本来主観的な認識ではあるものの，設計発話/応答の指示・ルール・習慣との差別化も難しかった．そこで，意図発話/応答に信念は含めず，願望のみを含めるといった方法で再実験を検討する．

また，心があると感じる要素として共感があるのではないかというコメントもあった．共感はまさに自己と他者が同じ認知プロセスを仮定しているとも考えられるため，本研究の仮説を検証する上で重要なポイントかもしれない．

そのほかにも，本実験は初歩的な検証のため，テキストを実験参加者が読んで評価したが，リアリティに乏しく明確な差が表れなかった可能性がある．従って，動画でそのシチュエーションを視聴してもらったり，実際にロボットなどとのインタラクションを体験してもらって評価を行う必要があるかもしれない．

一方で，そもそも「心があると感じる」という漠然とした質問自体の問題である可能性も否めない．これは社会的承認による定義に基づく開発の重大な問題となりうる．または，そもそも本実験で計画したような短期的なインタラクションの評価は難しく，長期的なインタラクションの評価を前提にする必要があるかもしれない．いずれにせよ，さまざまな方法による評価を検討・施行していく．

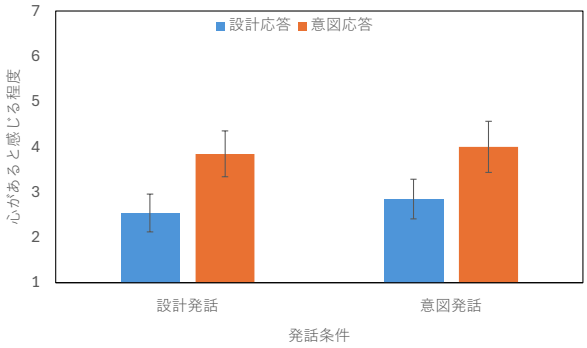


図 2　結果：シチュエーション ID1

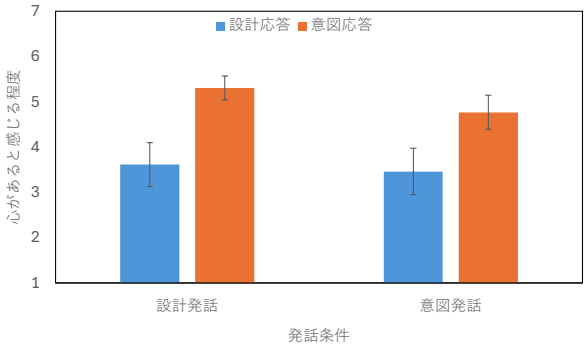


図 3　結果：シチュエーション ID2

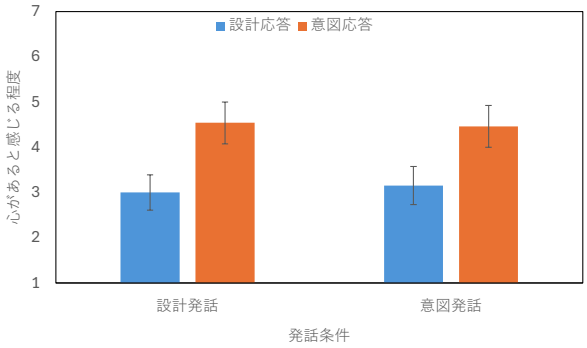


図 4　結果：シチュエーション ID3

4　おわりに

本研究では，社会的承認による定義をした心をもった AI について検討した．また，3 ステップの開発プロセスを考えた上で，社会的承認によって定義された AI の効果や評価方法を模索した．実施した実験は議論の端緒をつくる目的にすぎず，大幅な見直しが必要である．今後も引き続きどのような実験であれば社会的承認によって定義された心を持った AI の評価にふさわしいかを検討していく．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] 大澤正彦. 汎用人工知能実現に向けた人とエージェントの相互適応の研究. PhD thesis, 慶應義塾大学, 2020.

[2] 大澤正彦. ドラえもんを本気でつくる. PHP 研究所, 2020.

[3] 大澤正彦. 大規模言語モデルはドラえもんになれるか. 人工知能, Vol. 39, No. 2, pp. 214–221, 2024.

[4] 飯田愛結, 阿部将樹, 奥岡耕平, 福田聡子, 大森隆司, 中島亮一, 大澤正彦. 意図を読む ai の実現に向けて: 対話型生成 ai と他者モデルの統合を例に. HAI シンポジウム, 2024.

[5] Daniel C Dennett. **The intentional stance**. MIT press, Cambridge, MA, USA, 1989.

[6] Michael Bratman. **Intention, plans, and practical reason**. University of Chicago Press, Chicago, USA, 1987.

[7] Anand S Rao and Michael P Georgeff. Modeling rational agents within a bdi-architecture. In Michael N. Huhns and Munindar P. Singh, editors, **Readings in agents**, pp. 317–328. Morgan Kaufmann, San Francisco, CA, USA, 1997.

[8] Ayu Iida, Kohei Okuoka, Satoko Fukuda, Takashi Omori, Ryoichi Nakashima, and Masahiko Osawa. Integrating large language model and mental model of others: Studies on dialogue communication based on implicature. In **Proceedings of the 12th International Conference on Human-Agent Interaction**, pp. 260–269, 2024.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).