［研究助成（A）］

脳情報デコーディングを用いたロボットに対する心の知覚の定量化

Quantification of mind perception for robots by neural decoding from human brain activity

2191031

研究代表者	㈱国際電気通信基礎技術研究所 脳情報総合通信研究所	主任研究員	堀　川　友　慈
共同研究者	㈱国際電気通信基礎技術研究所 石黒浩特別研究所	グループ リーダー	住　岡　英　信
	㈱国際電気通信基礎技術研究所 石黒浩特別研究所	客員研究員	港　　隆　史

［研 究 の 目 的］

近年，ロボットやバーチャルなエージェントが社会の様々なシーンで利用されるようになるにつれ，人間と高度なインタラクションを交わすことが可能な知的エージェントとして機械を利用する需要が高まってきている。そのような中，インタラクションを交わす相手として，心ある尊重すべき存在であると人に感じさせるロボットの開発は，機械から受けるサービスの印象や効用を向上させるとともに，社会における人間と機械の調和を促進することに貢献すると期待される。そして，ヒトに心ある存在であると感じさせるロボットの開発には，ヒトが評価対象に対してどの程度心があると感じるかを定量的に評価することが必要不可欠となる。

これまでのロボット開発では，ロボットの外見をリアルな人間に近づけることで人間らしさを高める試みや，デフォルメを施してあえてリアルさを落とすことで親しみやすさを高める試みなどが行われてきた。しかしながら，このような外見の人間らしさや親しみやすさは，ロボットに心があると感じることと関連はするが，間接的な評価でしかなく，必ずしもロボットに対してどの程度心があると感じるかに直結しない。そのため，対象に対する心の知覚量をより直接的に評価する指標が必要となる。また，人は必ずしも自らが抱いている印象に対して自覚的ではなく，主観的回答の結果と相反する印象を無自覚的に抱いていることも多くあるため，主観によらない客観的評価指標の開発が望まれる。

そこで本研究では，これらの問題に対し，心理評定実験と脳活動解析を組み合わせ，心理評定実験により推定した“心の知覚量＝心があると感じる度合い”を，機械学習の方法を用いて脳から情報を解読する「脳情報デコーディング技術」を用いて予測することで，評価対象への心の知覚量を客観的・定量的に評価可能とする技術の開発を試みた（図 1）。また，解析を通して，心の知覚に関わる情報が，ヒトの脳のど

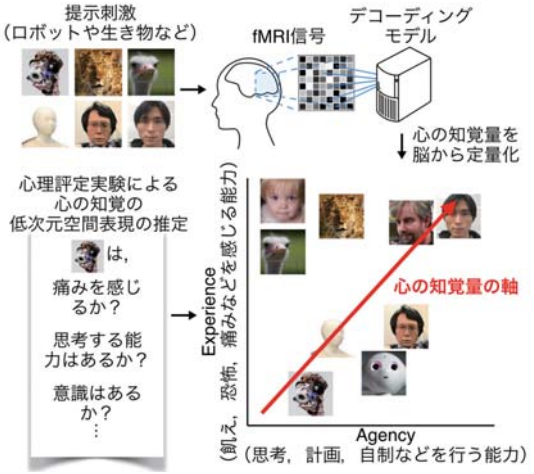

図 1　心の知覚量のデコーディングの概要図

の部位で表現されているかに関する探索を行った。

［研究の内容，成果］

まず，ヒトがロボットや動物などのさまざまな対象（キャラクタ）に，どの程度心があると感じるかに関する指標を構築するため，Gray et al.（2007）の研究をもとに，クラウドソーシングサービスを活用した心理評定実験を行い，さまざまなキャラクタの心的能力に関する評価スコアの収集を行った。この実験では，人間（e. g., ガンジー，ヒトラー），動物（e. g., イヌ，ゴキブリ），植物（e. g., ヒマワリ，サボテン），ロボット（e. g., ジェミノイド，C3PO），その他（e. g., 神，ピラミッド）の 5 カテゴリに属する合計 264 キャラクタに対して，18 項目の心的能力（e. g., 飢えを感じるか，計画を立てることができるか）および 6 項目の個人的嗜好（e. g., 好きか，幸せにしたいか）に関する 9 段階評価（非常に当てはまる－非常に当てはまらない）のスコアリングを，各質問，各キャラクタについて 20 人ずつの作業者に行わせた。

心理評定実験によって収集した心的能力に関する評定スコアに対して主成分分析を行い，キャラクタ間の心的能力の違いを説明する潜在変数（主成分）の推定を行った。その結果，上位の二つの主成分によって全体の変動の 96% 以上が説明され，第 1 主成分が 88.5%，第 2 主成分が 7.9% の変動を説明していた。各主成分に対する心的能力の因子負荷量を見ると，第 1 主成分に対してはほぼ全ての心的能力の変数が正の因子負荷量を示し，第 2 主成分に対しては飢えや痛み，恐怖など，原始的な欲望を感じる能力に対する因子負荷量が正の値を示していた。このことから，第 1 主成分が心的能力全般の有無に対する次元，第 2 主成分が Gray et al.（2007）で報告されていた Experience 次元と対応するような空間が得られたと考えられる。

推定された二つの次元で張られる空間上で，

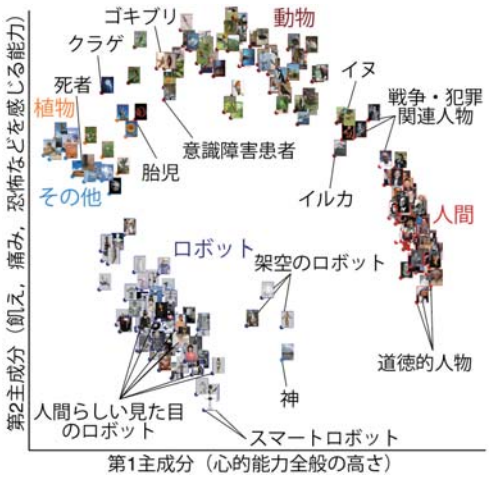

図 2　各キャラクタの主成分スコアに基づく 2 次元空間上におけるキャラクタの分布

各キャラクタがどのように分布しているのかを確認するため，各キャラクタの主成分スコアを 2 次元のマップ上で可視化した（図 2）。マップを見てみると，概ね，人間，動物，植物，ロボット，その他がそれぞれ大きなクラスターとなって分布していた。心的能力全般の高さを表すと考えられる第 1 主成分の次元では，人間が特に高い値をとっており，つづいてイヌやイルカなど高い知能を有すると考えられる動物から，ゴキブリやクラゲなど高度な知的活動を行うことができない生物が続いていた。動物のクラスタに続いて，ロボットが中程度の値を取り，植物や太陽，遺跡など，その他のカテゴリに属するキャラクタが低いスコアを示していた。一方，第 2 主成分の次元では，ロボットの中でも特に近年開発が進められているスマートロボットや，その他のカテゴリの神が低い値を取り，人間カテゴリの中でもマザー・テレサやネルソン・マンデラなど，道徳心の高い人物が比較的低い値を取っていた。ルンバなどのような単純なロボットや，ヒトラーなど戦争・犯罪と関連の強い人物が中程度の値を取り，動物が全体的に高い値を取っていた。

分布全体の形状は 13 キャラクタに対する評価を行なった Gray et al.（2007）の研究で得られたマップをやや回転させたような類似した結果が得られたが，264 キャラクタという多数の

評価対象を用いたことによって，人間やロボット，動物のカテゴリ内においても，個々のキャラクタに対する細かな心の知覚量の違いを数値化することが可能であることがわかった。特に興味深いことに，ロボットのカテゴリ内において，見た目が人間らしいロボット（e. g., ジェミノイド，ERICA）に対しても，他のいかにも機械でできているような外見のロボット（e. g., ASIMO, Sota）と似たスコアとなるロボットが多く，一方で，C3PO, R2D2, ベイマックスなど，フィクションの作品の中に現れる馴染み深い架空のロボットが，見た目は他のロボットとさほど違いはないものの，ロボットの群から大きく離れて分布することがわかった。この結果から，ロボットに対する心の知覚量が必ずしも外見の人間らしさに影響されないことが示唆された。

上記の解析で得られた主成分スコアをもとに，各キャラクタに対する心の知覚量を脳活動から予測可能かどうかを検証するため，心理評定実験に用いた 264 キャラクタに対する脳活動を機能的磁気共鳴画像法（functional magnetic resonance imaging, fMRI）を用いて計測する脳計測実験を行なった。この実験では，1 名の被験者（35 歳，男性）を対象に，264 キャラクタに対する 18 項目の心的能力および 6 項目の個人的嗜好に関する評価を行なっている時の脳活動を計測した（264×24＝6336 試行）。一つの試行は，評価対象の画像および名称を提示する期間（1 秒），視覚提示される評価項目に関する質問に対してボタン操作によって評価を行う期間（4 秒）で構成され，各試行後には 1 秒のレスト期間が置かれた。

得られた脳計測信号から，主成分スコアの予測を行うため，各キャラクタに対する主成分スコアを，各キャラクタの評価を行なっている試行中の脳計測信号から予測するモデル（スパース線形回帰モデル）を第 1，2 主成分それぞれについて構築した。解析では，脳のどの部位において高い予測成績が得られるかを調べるため，全脳を 360 の小領域に分割し，各小領域内の脳活動パターンを用いて，スコアの予測を行なった。モデルの学習とテストには，交差検証法（cross-validation）を用いて，8 分割したデータセットをモデルの学習データとテストデータが独立になるよう行なった（8 fold cross-validation）。予測成績の評価には，脳からの予測主成分スコアと，心理評定実験で得られた主成分スコアの相関係数を使用した。

第 1，2 主成分それぞれについて，各脳部位からのスコアの予測成績を調べたところ，いずれの主成分に関しても統計的に有意に高い精度で予測可能である脳部位があることがわかった（図 3）。いずれの主成分に関しても，高次視覚野（higher visual cortex）や背内側前頭前野（dorsomedial prefrontal cortex, DMPFC）付近で高い成績が得られており，これらの脳部位の活動パターンを元に，心の知覚量を表す 2 次元空間上において，各キャラクタの占める位置を

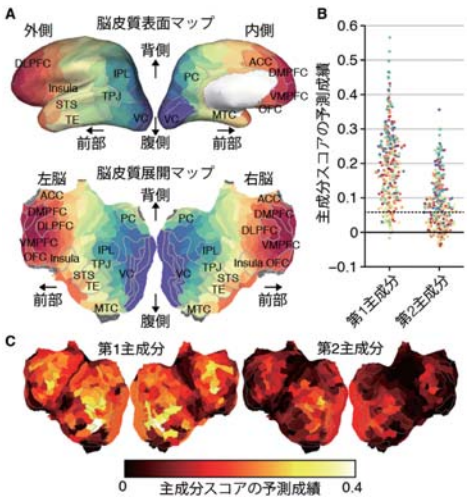

図 3　脳活動からの主成分スコア予測成績。（A）脳の脳皮質マップ。略記：VC, visual cortex ; TPJ, temporo-parietal junction ; IPL, inferior parietal lobule ; PC, precuneus ; STS, superior temporal sulcus ; TE, temporal area ; MTC, medial temporal cortex ; DLPFC/DMPFC/VMPFC, dorsolateral/dorsomedial/ventromedial prefrontal cortex ; ACC, anterior cingulate cortex ; and OFC, orbitofrontal cortex.（B）各脳部位からのスコア予測成績。各点が各脳部位からの予測成績を表す。色は（A）の各脳部位との対応を表す。破線は統計的に有意となる値を示す（相関 r=0.058, t 検定，p=0.01）。（C）第 1，第 2 主成分のスコアに対する各脳部位からの予測成績を表した脳皮質マップ。

予測可能であることが示された。

主成分スコアの予測において高い成績を示した脳部位が，評価対象であるキャラクタのどのような情報を表現しているのかをさらに詳細に調べるため，各脳部位におけるキャラク間の脳活動パターンの違いが，キャラクタの外見や心の知覚量の違いに応じてどのように異なっているのかを調べた。この目的のため，まず各キャラクタを評価している時の脳活動パターンから，評価対象であるキャラクタを判別するデコーディング解析を全ペア間で行ったところ，主成分スコアの予測に有効であった高次視覚野や楔前部，前頭前皮質付近で同様に高い判別成績が得られた（図 4）。

次に，ここで得られた結果を元に，3 体の架空のロボット（C3PO, R2D2, ベイマックス）および 12 体の人間らしい見た目のロボット（e.g., ジェミノイド，ERICA）が，他の人間やロボットとどの程度精度よく脳活動から判別可能かを評価した（図 5）。その結果，高次視覚野では，架空のロボットと人間の判別成績や，人間らしい見た目のロボットとロボットの判別成績が高くなっており，外見の違いが判別成績の高さに大きく影響していることが分かった。一方で，全ペアの解析で平均的に高い判別成績を示していた楔前部周辺の頭頂皮質や前頭前皮質付近の脳部位では，架空のロボットと機械らしい見た目のロボットとの判別成績や，人間らしい見た目のロボットと人間の判別成績が高くなっていた。このことから，これらの脳部位においては，外見の違いよりも，評価対象となるキャラクタの心の知覚量の違いが判別成績の高さに影響していることが示唆された。人間らしい見た目のロボットと人間の判別成績や，架空のロボットとロボットの判別成績が高かった脳部位には，楔前部の周辺部位など，心の理論（theory of mind）と関連があると考えられている脳部位が含まれていたため，ロボットに対する心の知覚においても，人の心を理解するのと同様の神経機構が働いていることが示唆された。

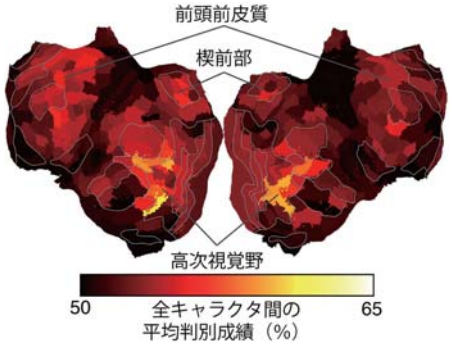

図 4　脳活動からのキャラクタ判別成績。全キャラクペア間の平均成績を示す。

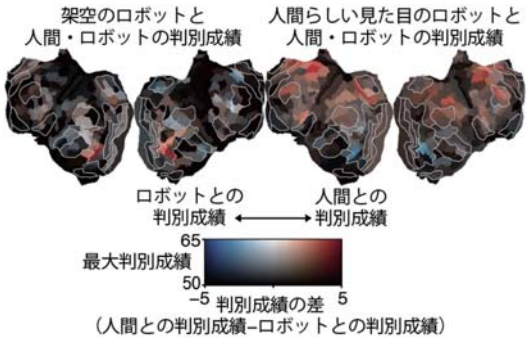

図 5　架空のロボットと人間らしい見た目のロボットに対する他の人間・ロボットの判別成績の差

［今後の研究の方向，課題］

本研究では，心理評定実験によって収集したキャラクタの心的能力の評価スコアから推定した心の知覚量（2 次元空間上の座標位置）を，個人の脳活動から一定の精度で予測可能であることが示された。これにより，実際にロボットとインタラクションを行う人の脳活動から直接推定した心の知覚量を評価基準に用いることで，質問紙などを介さずに，被験者に直感的に心あると感じさせるロボットに求められる振る舞いや能力を評価することができる可能性が示された。さらに，外見の類似度によらずに心の知覚量の違いが，楔前部や前頭前皮質など心の理論と関連のある脳部位において，特に顕著に表現されていることが示唆された。

しかしながら，本研究期間では，1 名の被験者のみの脳計測実験しか行えず，評価対象として用いたキャラクタに対する被験者の事前の知

識や，インタラクションの経験の有無などが，キャラクタに対する脳情報表現にどのような影響を及ぼすかについては検討することができなかった。特に，架空のロボットや人間らしい見た目のロボットに対する心的能力の心理評定結果や脳活動解析の結果が示すように，ロボットなどヒト以外のキャラクタに対する心の知覚量は，必ずしも外見の人間らしさだけでなく，当該キャラクタの能力に関する事前の知識の有無が強く影響することが示唆されている。したがって，ヒトに心ある存在であると感じさせるロボットを開発するために，どのような振る舞い方や能力をロボットに実装することが有効であるかをさらに詳細に分析するためには，今後，同一の被験者において，ロボットとのインタラクション前後での脳活動の変化を検討したり，事前の知識や経験が異なる複数の被験者間での結果の違いを検討することが重要な課題となると考えられる。

［成果の発表，論文等］

Horikawa, T., Cowen, A. S., Keltner, D., Kamitani, Y. The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. iScience 23, 101060 (2020).