機械の向こうの私
～ヒューマン-ロボットコミュニケーションにおける fMRI 研究～
Self-reflective mind in a machine
-Neural correlates of mental attribution in human-robot interaction -

○斎藤千夏[1]，高橋英之[1]，寺田和憲[2]，小嶋秀樹[3]，土師知己[1]
吉川雅博[4]，松本吉央[4]，大森隆司[1]，岡田浩之[1]
Chinatsu Saito[1], Hideyuki Takahashi[1], Kazunori Terada[2], Hideki Kozima[3], Tomoki Haji[1]
Masahiro Yoshikawa[4], Yoshio Matsumoto[4], Takashi Omori[1], Hiroyuki Okada[1]

[1] 玉川大学　[2] 岐阜大学　[3] 宮城大学　[4] 産業技術総合研究所
[1]Tamagawa University　[2] Gifu university　[3]Miyagi University　[4] AIST

Abstract: We sometimes attribute a mental state to a non-animate robot. Recent progress of neuroimaging techniques verifies the mental attribution to robots in terms of brain functions (Krach et al. 2008). However required properties of robots that induce the mental attribution are still unclear. In this study, we measured participant's brain activities by functional magnetic resonance imaging (fMRI) method when they play a matching pennies game with the other person, three robots, and a computer program respective. Further participants rated the impression of these opponents with questionnaires (e.g. human likeness, cuteness) and we explored the brain activities correlated with these impression ratings in addition to participant' strategic behavior in the game. Our preliminary results imply that there are several different brain functions for mental attribution for robots in human-robot interaction.

はじめに

近年，心を帰属されやすそうな外見や振る舞いのロボットが数多く開発されている．しかしロボットを実用化する上で，周囲の人間がロボットに心を帰属することは本当に重要なのであろうか？それともロボットは目的遂行の為の高い機能性を有してさえいればいいのであろうか？実用性の高いヒューマンロボットインタラクションを今後デザインする上で，ロボットへの心の帰属が人間の認知や振る舞いに与える影響をより深く知ることは重要であろう．

このような問題意識から，ロボットへの心の帰属が人間に与える影響について，行動解析や質問紙を用いて研究がなされてきた [1]．その一方で，心の帰属は直接観測することができない精神的な営みであり，行動や質問紙だけでは間接的にしか測ることができない．

Krach らはロボットの見た目を人間に近づけていくことにより，心の帰属にかかわる脳部位[2]の活動が高まることを，ロボットとインタラクションしている被験者の脳活動を functional magnetic resonance imaging(fMRI)で計測することにより示した [3]．この研究は，我々がロボットに対して心の帰属を行なっていることを脳活動のレベルから示す重要な知見である．しかし Krach らの研究はロボットの人間らしさを一つの軸として捉えているが，我々はロボットに対する心の帰属がより多元的な複数の軸として処理されていると考えている [4] [5].

本研究では，被験者が人間，三種類のロボット，そしてコンピュータ，それぞれと matching pennies game を行なっている際の行動と脳活動を計測した．さらに被験者には事前に各対戦相手と実際に面会をしてもらい，それぞれの相手に対する印象を多項目に渡る質問紙により評定してもらった．そしてエントロピーで定量化されるゲーム中の行動戦略，そしてロボットに対する印象の質問紙の評定値と相関する脳部位を，心の帰属にかかわるとされる領野の中から探索することにより，ロボットに対する心の帰属が脳機能のレベルでどのような軸で処理されているのかについて検討した．

実験設定

被験者が脳計測中に行う実験課題として，matching pennies game を用いた．このゲームは，被験者と対戦相手は二つの選択肢から一つの行動を選

択することを求められ，両者の行動の組み合わせで被験者の勝ち負けが決まる最も単純なゼロサムの対戦ゲームである．被験者が勝つと一定の金額の報酬が，被験者が負けると同額の罰が生じる．我々はこれまでに，このゲームにおける被験者の行動戦略を過去の行動系列から次の行動を選択する条件付き確率のエントロピーにより定量化し（エントロピーが大きいほど複雑な行動系列であることを意味する），ゲームの相手が人間であると思うことでこのエントロピーが有意に増大すること，そしてこのゲームをロボットと行う状況において，被験者のエントロピーはロボットに対する人間らしさの主観的評定とは相関がない一方，ロボットの視線を被験者が意識している場合に大きくなることを示した．これらの知見から，我々はロボットに対する心の帰属には，主観評定にあらわれるような意識的な軸と，視線の知覚といった無意識的な軸があると考えている [4].

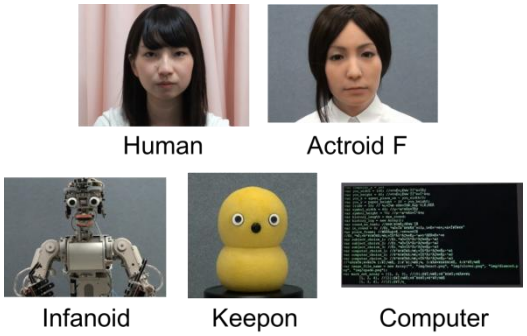

図 1. 五種類の対戦相手

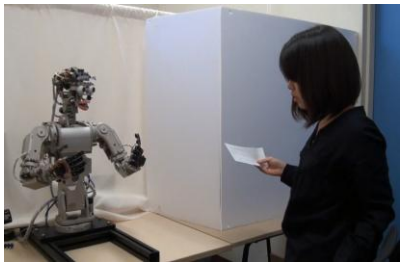

図 2. 事前インタラクションの様子 （上：Actroid F 下：Infanoid）

URL: http://www.youtube.com/watch?v=OrZfkHHqyCc

ゲームの対戦相手として，人間女性，Actroid F（見た目が人間と酷似したアンドロイド），Infanoid（メカニカルなロボット），Keepon（鳥の雛のようなロボット），そして computer の五種類の相手を用意した（図 1）．被験者は MRI 装置の中でゲームを行う前に，実際に各対戦相手と面会し，相手に簡単な自己紹介，対戦相手に対する印象，これから行うゲームに対する意気込み，の三点を述べるように求められた．被験者は各対戦相手と面会した後に，多項目に渡るロボットに対する印象（人間らしさ，賢さ，かわいさ，生物らしさなど）を 7 件法により評定した．その後に，被験者は MRI 装置の中に入り，各対戦相手と matching pennies game を行った．対戦相手の映像はムービーとして被験者に提示された．ゲームにおいて，被験者は一試行あたり 1 秒の制限時間内で左右二つの方向の中から一つの方向を選択することが求められた（この間に行動を選択できない場合はミス試行した）. 実験はブロックデザインで行い，1 ブロックの中で被験者は 20 試行，同一の相手と続けてゲームを行った．被験者は各対戦相手と 6 ブロックずつゲームを行った（ブロックの順番はランダム）. 実際には対戦相手の行動は相手の種類によらず常にランダムであり，被験者の一試行ごとの勝率は常に 50%である．

実験には大学生 20（女性 12 人）人が参加した．そのうち 2 名はミス率が高かったため，2 名は MRI 計測中の体動が大きかったため, 解析から除外した.

ゲーム中の行動解析

行動解析として被験者の選択した行動系列のランダムさをエントロピーにより定量化した．ゲームの過去の被験者と対戦相手の行動系列を c，被験者が次に選択する行動を d とすると，被験者の行動選択確率 $p(d|c)$ からエントロピー H を以下のように計算した.

$$H = \sum_{c}\sum_{d} p(d|c) log_2 p(d|c)$$

詳細については文献[4]を参照のこと.

脳機能画像の解析

脳機能画像は, 3 Tesla の MRI scanner (Trio a Tim System, SIEMENS) を用いて計測を行った．脳画像の計測は TR=3sec で interleave の方式で全脳を対象に 3mm 四方のボクセルレベルで行なった．解析は SPM8 (http://www.fil.ion.ucl.ac.uk/spm/) を用いて行った.

プレプロセッシングとして realign により機能画像の撮像位置のずれを補正し，標準脳空間である MNI 座標系へと normalize を行い、8mm のガアウシアンカーネルを用いて smoothing を行った.

脳機能画像の統計解析として，対戦相手の種類に応じてゲームの中に活動の高まりを仮定する GLM モデルを用いて，ボクセルごとの差分解析を行った。次に GLM モデルにより計算されたコントラストによって重み付けされたパラメータ加重値を用いて、変量効果モデルによる検定を行った（$p < 0.001$ uncorrected, whole-brain cluster corrected $p < 0.05$)。

実験結果

質問紙と行動の結果

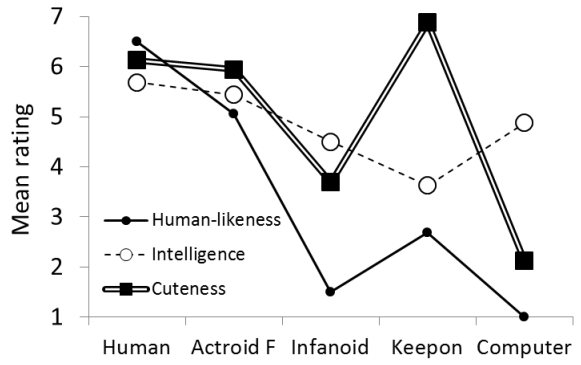

図 3. 対戦相手に応じた人間らしさ，賢さ，かわいさの項目における質問紙の評定平均

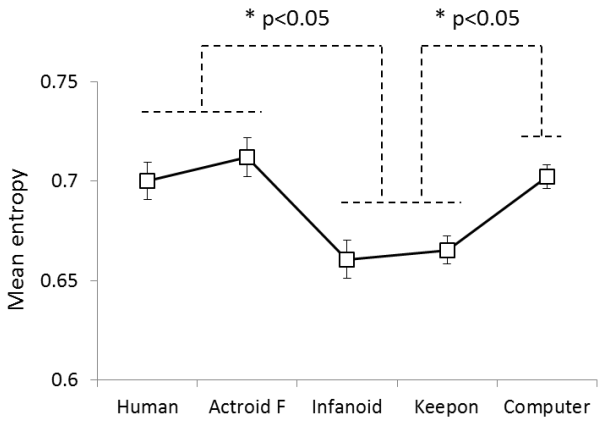

図 4. 対戦相手に応じたエントロピーの平均

印象評定（人間らしさ，賢さ，かわいさ）の対戦相手に応じた平均を図 3 に示す. この印象評定から，人間，そして Actroid F はすべての項目が高いが，他のロボットとコンピュータは一定の傾向はないことが読み取れる．人間らしさの評定は Infanoid, Keepon, computer のどれも低いが，賢さは computer が高い値で，かわいさは Infanoid が高い値で評定された.

また matching pennies game におけるエントロピーの平均を対戦相手ごとに比較した．一要因の対応有りの分散分析の結果，対戦相手の種類に有意な主効果があらわれた（$F[4,15]=6.401$, $p=0.0002$)．下位検定を行ったところ，Human, Actroid F, computer とゲームを行なっている際のエントロピーの平均は，Infanoid, Keepon とゲームを行なっている際のエントロピーよりも有意に高いことが分かった（Ryan's method).

脳活動の解析結果

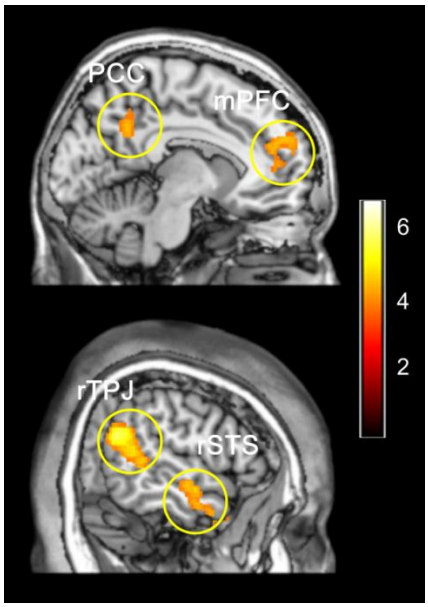

図 5. 人間とゲームを行なっている際の脳活動からコンピュータとゲームを行なっている際の脳活動の差分（$p < 0.001$ uncorrected, whole-brain cluster corrected $p < 0.05$)

図 5 は人間とゲームを行なっている際の脳活動からコンピュータとゲームを行なっている際の脳活動の差分である（$p < 0.001$ uncorrected, whole-brain cluster corrected $p < 0.05$)．その結果，先行研究で心の帰属に関係があると言われている脳部位[2]，内側前頭前野(medial prefrontal cortex; mPFC)，後部帯状皮質(posterior cingulate cortex; PCC)，

右上側頭溝(right superior temporal sulcus; rSTS), 右側側頭頭頂接合部(right temporoparietal junction; rTPJ)に賦活がみられた.

これらの差分で残った領域を ROI として, GLM における ROI 内のボクセルのβ値の平均を対戦相手ごとに計算し, その対戦相手とゲームを行なっている際のその領域の活動レベルとした.

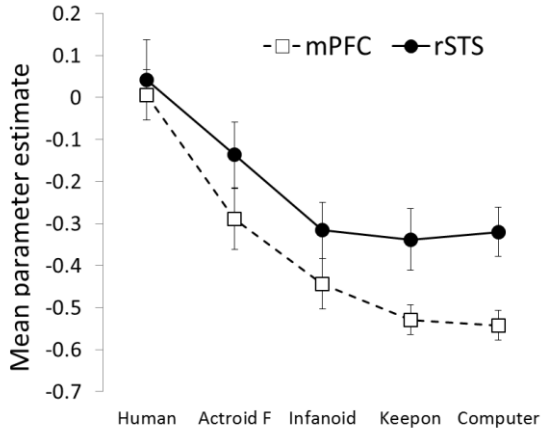

図 6. mPFC と rSTS の対戦相手に応じた活動レベル

図 6 は mPFC と rSTS における対戦相手に応じた活動レベルのグラフである. このグラフをみて分かるように, 人間はこれらの領域の活動レベルが高く, Infanoid, Keepon, Computer はこれらの領域の活動レベルが低い. そして人間そっくりのアンドロイドである Actroid F に対する脳活動は人間相手と機械的なロボットの中間となった.

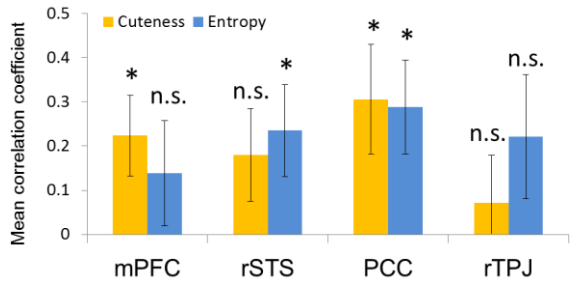

図 7. 個人内での対戦相手間でのかわいさの評定値とエントロピーの値それぞれと, 各脳領域の活動レベルとの間での相関係数の平均値 (* p<0.05 one sample t-test)

次に各領域がロボットに対する心の帰属に関するどのような情報を処理しているのかを調べるために, 個人内での対戦相手間での質問紙各項目の評定値とエントロピーの値それぞれと, 各脳領域の活動レベルの相関係数を計算し, 被験者間で one sample t-test により有意に相関係数が 0 より大きい, もしくは小さいか評価した. 図 7 はかわいさの項目の評定値とエントロピーそれぞれと各脳領域の活動レベルの間での個人内の相関係数の被験者間の平均値である. この結果から, mPFC の活動レベルはかわいさの評定値と有意に正の相関がある一方, エントロピーとは相関が弱い, 一方で rSTS の活動レベルはかわいさの評定値と相関は弱いが, エントロピーと有意に正の相関があった. さらに PCC の活動レベルは, かわいさの評定値ともエントロピーとも共に有意に正の相関があり, rTPJ の活動レベルはかわいさの評定値ともエントロピーとも相関がみられなかった. 従って, 心の帰属にかかわるこれらの 4 領域はそれぞれ異なる属性の情報を処理していることが示唆された.

議論

本研究では, ロボットとゲームを行なっている際に fMRI で計測される脳活動変化を, ロボットに対する詳細な印象評定とエントロピーという対戦相手の認識に敏感な行動指標と合わせて解析することにより, 被験者のロボットに対する心の帰属の軸を, 脳機能のレベルから明らかにすることを試みた. その結果, まだ予備的な結果ではあるが, 心の帰属にかかわる 4 つの脳領域それぞれで異なる属性の情報を処理していることが示唆された.

特に興味深い点として, エントロピーと rSTS の活動は正の相関を示す一方で, mPFC にはこのような正の相関がみられなかった. 今回の研究では, 先行研究と違い対コンピュータの際にもエントロピーが高くなった. この理由として, 実験中に被験者に提示したコンピュータの映像は, 難しいプログラムコードが画面上を流れているというものであり, こちらの動きを計算で予測しているような印象を被験者に与えた可能性がある. 実際に, 相手の賢さに関する印象評定の値はコンピュータが相手の時に高くなった. 我々の先行研究においても, ロボットと matching pennies game を行う際に, ロボットの主観的な人間らしさの印象よりも, 被験者がロボットの視線を意識していた時にエントロピーが増加することが示唆されている [4]. mPFC はどちらかというと再帰的推論で相手の思考を読むといった論理的な他者への心の帰属を反映していることが知られている一方[6][7], rSTS は視線や動きといったより知覚駆動の他者への心の帰属を反映しているという報告がある [8]. そしてエントロピーが後者と相関するということは, matching pennies game の場合はロボットに対する推論的な心の帰属ではなく, 知覚駆

動（ロボットの視線の知覚など）の心の帰属が行動に支配的であるということが脳活動のレベルからも示唆されたのではないかと考える．

さらにPCCやrTPJの活動もmPFCやrSTSとは異なったパターンを示していた．これらの部位の機能を現時点で深く言及することは難しいが，これらの領野もそれぞれ心の帰属を反映する何らかの特異性を持っている可能性は高く，今後の検討が望まれる．

今回はmatching pennies gameという対戦ゲームでの検討であったが，今後はより様々なインタラクション場面においてロボットへの心の帰属に関係する脳機能と被験者の振る舞いや相手への印象の関係を調べていくことにより，ロボットへの心の帰属がどのようにヒューマンロボットインタラクションにおける人間の振る舞いに影響を与えるのか，脳機能のレベルで議論できるようになるかもしれない．このような知見は，より効果的に「人間を動かす」エージェントの開発に大きな知見を与えるものと期待される．

謝辞

本研究は，本研究はJSPS科研費 若手研究B(23700321),新学術領域「人ロボット共生学（No. 4101)」(24118708, 21118002, 24118703)，新学術領域「伝達創成機構（No. 4103)」(21120010)，玉川大学グローバルCOEプログラム「社会に生きる心の創成」の助成を受けた．また人間の対戦相手のサクラ役として協力して下さったアツアンヤ亜伊子さん，岩崎安希子さん，鈴木春香さん，瀧田愛さんに感謝いたします．

参考文献

[1] Osawa, H., Mukai, J, Imai, M., Anthropomorphization Framework for Human-Object Communication, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 11(8), 1007–1014, 2007.

[2] Frith, U., Frith, C. D., The social brain: allowing humans to boldly go where no other species has been. *Philosophical Transactions of the Royal Society B*, 365(1537), 165-176, 2010.

[3] Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., et al., Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLoS ONE*, 3(7), e2597, 2008.

[4] Takahashi, Saito, et al. An investigation of social factors related to online mentalizing in a human-robot competitive game. *Japanese Psychological Research*. (in press)

[5] 高橋，斎藤，古市，岡田，金岡，渡辺，コミュニケーションロボットの擬人化は単一の軸で捉えられるか？ -擬人化における志向要因と情動要因の分離-*HAI シンポジウム 2012*，京都，2012.

[6] Coricelli, G., Nagel. R., Neural correlates of depth of strategic reasoning in medial prefrontal cortex, *Proceedings of the National Academy of Sciences*, 106(23), 9163-9168, 2009.

[7] Yoshida, W., Seymour, B., Friston, K.J., Dolan, R.J. Neural mechanisms of belief inference during cooperative games., *The Journal of Neuroscience*, 30(32), 10744-10751, 2010.

[8] Pelphrey, K.A., Viola, R.J., McCarthy, G., When Strangers Pass: Processing of Mutual and Averted Social Gaze in the Superior Temporal Sulcus, *Psychological Science*, 15, 598-603, 2004.