ERP による人・エージェント間内集団関係形成の測定

Detectiong an In-group Relation between Humans and Agents By ERP

本武　陽一 [1*]　福田　玄明 [1]　植田　一博 [1]
Yoh-ichi Mototake[1]　Haruaki Fukuda[1]　Kazuhiro Ueda[1]

[1] 東京大学大学院総合文化研究科
[1] Graduate School of Arts and Sciences, The University of Tokyo

Abstract: The purpose of this study is to detect, by brain activity, the emergence of in-group relation between a person and an artificial agent. For this purpose, we created a three-way discussion setting between a participant and two artificial agents: One agent took the same opinion as a participant while the other did the opposite. This resulted in a feeling of group identity between the participant and the first agent, i.e. the same side. This feeling of group identity was shown to be detected using error-related negativity (ERN), a component of an event-related potential that accompanies errors in speeded performance. We found that amplitude of ERN was bigger for the same side agent's failure than for the other side's. We also found ERN difference correlated to an empathy ability detected by a questionnaire. We could also classify the time-series data of ERN into in-group or out-group state by Support Vector Machine. From these results, we can say that an in-group relation between a person and an artificial agent can be detected by ERN.

1　はじめに

近年の人工知能関連技術の急速な発達によって，Apple 社の「Siri」や，NTT ドコモ社の「しゃべってコンシェル」など，人間とコミュニケーションを行いながら，タスクを処理するような人工エージェント（以下，エージェント）が実用化されつつある．さらに，深層学習などの機械学習技術の発展によって，人間と同等なコミュニケーション能力を持ったエージェントが近い将来登場することも期待されている [Kurzweil 90]．

人間がそうであるように，人工エージェントが 100% の精度でコミュニケーション相手の音声や意図の認識を行うことは困難である．これは，独立したシステムの内部に人間と同等の知能を実現しようとする，従来型の人工知能研究の限界を示唆する．この問題に対処するには，人間とエージェントがコミュニケーションを通して協力しながらタスクを行う際に創発されるような，開かれた知能を検討する必要がある．これは，HAI(Human-Agent Interaction) の主要な課題の一つである [小野 06]．

本研究では，人間が仲間や友人や家族といった，帰属意識の対象となる集団（内集団）のメンバーに対して協力的行動を行う [Brewer 86] という現象に着目した．エージェントとユーザーの間に内集団関係を形成できれば，ユーザーからの協力行動を引き出せると考えられるためである．実際に Reeves ら [Reeves 96] は，簡単な条件付けによって人間がコンピュータ（エージェント）を人間と同様に社会的コミュニケーション法則に従うものとして捉えること，そして，それによってコンピュータとの間にある種の内集団関係が形成できることを示唆している．

そこで本研究では，人工エージェントの実際の利用状況に近いユーザーとの対話環境において，ユーザーとエージェントの間で内集団関係を構築できるかを検討した．その際，工学的応用と客観性の担保を念頭に，測定が容易な脳活動の指標である事象関連電位（Event Related Brain Potential，以下 ERP）を用いた内集団関係の検出を検討した．

このように本研究の目的は，人間とエージェントが，人間同士の場合と同様のコミュニケーション法則に従って内集団関係を形成し得たかを，脳活動の測定を用いた客観的指標によって検出することである．

*連絡先：東京大学大学院総合文化研究科
東京都目黒区駒場 3-8-1　東京大学 3 号館 314
E-mail:mototake@sacral.c.u-tokyo.ac.jp

1.1 仲間の失敗行動の観察と脳活動

人間は，自分の帰属する集団やその構成員と視点や目的，情動などを共有しがちである [Wann 93, Wann 01]．例えば，スポーツを観戦する際には，同じ失点の発生でも，味方チームが行えば悔しく感じ，敵チームが行えば嬉しく感じる．このような人間同士で生じる現象が，人間とエージェントの間でも生じることを，脳活動の測定を通して示した研究 [Newman-Norlund 09] がある．

この研究では，テレビゲームをもとに作成された，実験参加者の自国チームと他国チームのエージェント（ゲームキャラクタ）がサッカーの PK 戦を行っている動画を，実験参加者に観察させた際の脳活動を，機能的核磁気共鳴画像法（functional magnetic resonance imaging，以下 fMRI）により測定している．この結果，ゴールに失敗したことを観察した際の実験参加者の脳活動のうち，エラー処理や報酬予測，意思決定，共感や情動といった認知機能に関わっているとされる前帯状皮質（Anterior Cingulated Cortex，以下 ACC）の一部で，自国チームが失敗したか，他国チームが失敗したかの違いに応じて変化する脳活動が見られた．さらに，アンケート調査で得られた，実験参加者の「他者の失敗や苦痛を自分のものと感じる」共感能力値と，自国・他国条件の違いによる脳活動の変化量との間に相関が見られたことから，この脳活動が他者の視点や情動の共有と関係すると議論している．以上のことは，エージェント（テレビゲームのキャラクタ）の失敗を観察した際の人間の脳活動を測定することで，そのエージェントと実験参加者の間に，内集団関係が存在しているかどうかを検出できることを示唆している．

しかしながらこの実験は，チームや選手という実際に存在する団体や人物をエージェントに投影することで，それに対する選好や知識に基づく内集団関係の検出を行ったものであり，エージェントとの間にコミュニケーションを通した内集団関係の構築がされたとは言えない．そこで本研究では，現実の人間や実在する団体と関係しないエージェントを用意し，実験の過程で内集団関係を形成することを試みた．具体的には，自分の意見に賛成したエージェントに対して，親和的反応を示すという研究 [竹内 00] を参考にして，実験参加者と，実験参加者の意見に賛成・反対する 2 エージェントとの三者間ディスカッションを行い，その場で内集団関係を形成することを試みた．

1.2 ERN による内集団関係の検出

前述した Newman-Norlund の研究 [Newman-Norlund 09] では，fMRI を用いて，実験参加者とエージェントとの関係性の違いに応じた脳活動の変化を議論している．これに対して，本研究では，脳波のうち運動や外部からの刺激に応じて発生する脳の電気活動である，事象関連電位（Event Related Brain Potential，以下 ERP）を用いた．これは，後述するように，Newman-Norlund の研究 [Newman-Norlund 09] と同様の脳活動を，ERP を用いても間接的に測定できると予想される上に，ERP は fMRI に比べて測定が簡便なためである．

特に本研究では，ERP の中でも自身の行動や意思決定の失敗に関連して発生する ERN（Error Related Negativity）に着目した．ERN は，自身の失敗だけでなく，他者の失敗を観測した際にも生じることが知られており [Scheffers 00]，他者との視点の共有と関係していると考えられる．また，Newman-Norlund の研究 [Newman-Norlund 09] でみた脳活動と同様に，ERN の発生源は ACC だと推定されている [Dehaene 94, Ito 03]．以上より，実験参加者がエージェントの失敗を観察した際の ERN に，実験参加者とエージェントとの関係性の違いに応じた差分が生じることが予想される．

実際に [Fukushima 09] と [Leng 10] は，他者（人間）の失敗を観察した際の実験参加者の ERN が，その他者との関係性の違いによって変化するかを検討している．具体的には，実験参加者以外の第三者にギャンブル課題を行わせ，この第三者が，当たりを外した場面を観察した実験参加者の FRN（Feedback Related Negativity：ERN と同じ脳内機序であると推定され，予測報酬誤差に応じて発生する）を測定している．[Fukushima 09] では，この第三者の利得と観察者の利得はお互いに相反するよう設定されたのに対して，[Leng 10] では，無関係となるよう設定された．この FRN に，第三者との関係性の違い（実験参加者の知人か他人か）に応じた差分が発生するかどうかを検討した結果，[Fukushima 09] では FRN に第三者との関係性の違いに応じた有意傾向のある差分がみられたが，[Leng 10] では有意な差分はみられなかった．

これら 2 つの研究において，第三者との関係性に応じた FRN 差分がうまく観測されなかった原因として，FRN の測定に用いたギャンブル課題によって醸成される緊迫感や悔しさなどの強度が弱かったため，他者が苦痛を感じていることをはっきりと認識できず，結果として他者の苦痛を自分のものと同一視することで生じる脳活動も強く表れなかったことが考えられる．この問題は，今回の人工エージェントに対する内集団関係の検出においても生じると考えられる．これを解決するため，本研究では，ギャンブル課題を，緊迫感や仲間が負けた際の悔しさがより高いと考えられる対戦型ゲームに変えて実験を行った．実際，前述した fMRI を用いた先行研究 [Newman-Norlund 09] でも，対戦型ゲームであるサッカーの PK 戦を用いて，関係性の違いに応じた脳活動の変化を測定することに成功している．本研究では，インディアンポーカーを改良した，カード

の数字の大小を競うエージェント間対戦型カードゲームを用い，それぞれのエージェントの敗北が決したシーンを観察した際の実験参加者の脳活動を測定した．

2　実験

実験は 3 つのフェーズから構成されていた．

最初のフェーズは，実験参加者とエージェントの間に内集団関係を形成させるための，実験参加者・賛成エージェント・反対エージェントの三者による「ディスカッションフェーズ」であり，賛成エージェントは実験参加者の意見に賛成し，反対エージェントは反対するように設定した．ここで実験参加者が，賛成エージェントとの間に内集団関係を持つと想定した．

第 2 フェーズは，ディスカッションフェーズによって内集団関係が形成されたことを確認するための「脳波測定フェーズ」で，賛成エージェントと反対エージェントが対戦型カードゲームをしている際の実験参加者の脳波の測定を行った．特に，それぞれのエージェントが負けたときに生じる ERP を測定した．

第 3 フェーズは，本実験において内集団関係形成の前提となる，ディスカッション時の賛成エージェントが自分と同じ側の意見を持っていたという意識（以下，同側意見意識）の存在を確認するための「アンケートフェーズ」で，実験参加者にエージェントをどれだけ敵・味方と感じたかを質問した．同時に，[Newman-Norlund 09] と同様に，測定した脳波と実験参加者の共感能力の関係を調べるために，共感能力評定アンケートを行った．また，実験参加者がディスカッションを通して得たエージェントに対する印象と脳波の関係を調べるために，ディスカッション時のエージェントに対する印象評定アンケートも行った．

2.1　実験参加者

実験の目的を知らされていない 18~36 歳までの健康な男女 30 名（男性 : 22 名，女性 : 8 名，平均年齢 : 22.03 歳，標準偏差 : 4.09 歳）が実験に参加した．全員が右利きで同様の実験に参加した経験はなかった．実験参加者は，事前に実験に関する説明を受け，実験参加同意書にサインした上で参加した．なお本実験は，東京大学大学院総合文化研究科の「ヒトを対象とした実験研究に関する倫理委員会」の承認の下で実施された．

2.2　実験手順

§フェーズ 1：ディスカッションフェーズ

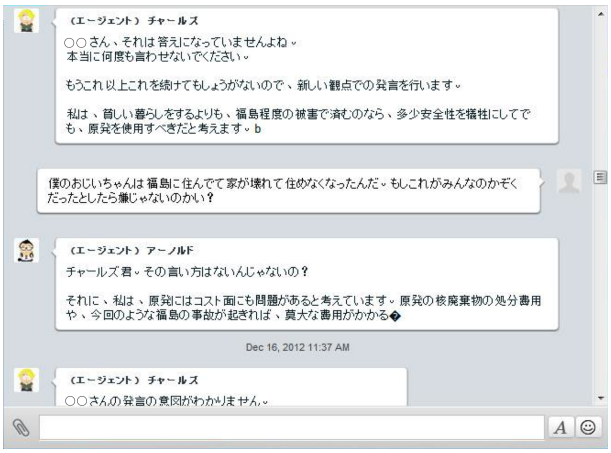

図 1: ディスカッションシステムのスクリーンショット

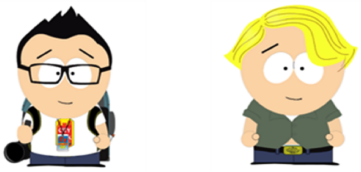

図 2: アバター画像

ディスカッションには，AOL 社チャットシステムを用いた．実験参加者には，図 1 のような画面が提示された．画面には，エージェントの発言と共に，実験参加者自身の発言も表示され，エージェントが発言を行う際には，エージェントを表すアバター画像（図 2）とエージェントの名前が同時に表示された（図 1 中の○○は，実験参加者の名前を表す）．

実験参加者には「ディスカッションで使用するエージェントには人工知能が搭載されている」と教示したが，実際には人工知能エンジンを搭載したエージェントは利用せず，実験者がエージェントを操作した．エージェントを視覚的に表すものとして，図 2 のようなアバターを用いた．このアバターは，米アニメサウスパーク（©2013 South Park Digital Studios LLC. All Rights Reserved）公式サイトで提供されている，アバター作成サービス AVATOR CREATER を用いて実験者が作成した，他に実在しないキャラクタである．それぞれ，「アーノルド（Arnold）」,「チャールズ（Charles）」という名前を与えた．

実験では，事前に実験参加者へアンケートを行い，以下のディスカッションテーマから実験参加者が最も強い意見を持っているテーマを 1 つ選択した．

1. 「日本は死刑制度を廃止すべきである．是か非か」

2. 「日本は核兵器を保持すべきである．是か非か」
3. 「日本はすべての原子力発電を代替発電に切り替えるべきである（ただし，切り替えは 2040 年までに行うこととする）．是か非か」

ディスカッションは，次のように実施された．ディスカッションは，3 人（実験参加者，実験参加者の意見に賛成するエージェント，実験参加者の意見に反対するエージェント）で行った．アーノルドとチャールズの見た目の違いを統制するために，アーノルドを賛成エージェントにするか，チャールズを賛成エージェントにするかは，実験参加者間でバランスをとった．最終的に分析対象となった実験参加者のうち，アーノルドが賛成エージェントだった実験参加者は 12 名，チャールズが賛成エージェントだった実験参加者は 12 名となった．またディスカッションでは，反対エージェントの意見に実験参加者を直接対峙させるために，発言の順番を「実験参加者→賛成エージェント→反対エージェント」とあらかじめ決めて行い，最初の発言で賛否の立場を表明するということ以外，発言内容は自由に考えて良いと教示した．各エージェントも最初の発言で賛否の態度表明を行い，実験参加者はこのときに初めて，どのエージェントが賛成エージェントか反対エージェントかを知った．ディスカッションは,「実験参加者→賛成エージェント→反対エージェント」の発言を 1 セットとして，10 セット行った．エージェントの発言は，実験者が事前に用意したテンプレートをもとに，それらを組み合わせてその場で作成した．このテンプレートは，事前に別の実験協力者と行ったディスカッション内容をもとに作成され，これに実験参加者と賛成・反対エージェント間での関係性の形成を助長するような，感情的で否定的な表現を，反対エージェント側の発言に付け加えたものである（例えば図 1 を参照のこと）．一方，賛成エージェントと反対エージェントが直接対立する構造をさけるため，賛成エージェント側には特別に感情的な表現を付け加えることはしなかった．

§フェーズ 2：脳波測定フェーズ

脳波測定には，EGI 社 Geodesic EEG System 300（以下，GES 300）を使用した．ハードウェアレベルでのサンプリングレートは 1kHz で，入力チャンネル数（脳波測定電極数）は 65 であった．電極位置は，図 3 に示すように拡張 10-20 法に従って配置した．電極インピーダンスは 10k Ω以下であった．また，画像刺激提示には，iiyama 社製 HM903DB 19 型 CRT モニタと DELL 製 precisionT5500（CPU: Xeon 2.27GHz/dual core, Memory: 8GB, OS: 64bit Windows7 Professional, Graphic Board: NVIDIA GeForce GT 520）を用いた．モニタと実験参加者との距離は 50cm に設定された．

ディスカッションフェーズ終了後，5 分程度の休憩を挟み，その間に，次の実験で用いるカードゲームを，実験者が実験参加者と実際にゲームを行いながら説明した．このカードゲームのルールを表 4 に示す．

実験刺激（ゲーム画面）は，カードゲームの流れに従って図 5 のような手順で表示された．ERP 算出のために加算平均の対象とする区間を，図 5 の勝敗が決する画面 5 が表示された瞬間から，注視点のみの画面となる画面 0 までの間とした．この区間の開始点を記録するために，画面 5 が現れた瞬間に，脳波データと並行してトリガを記録した．トリガ信号は，実験画像提示プログラムによってコントロールされた，タートル工業製 TUSB-16DIF ボードによって，GES 300 へアナログ情報として伝達された．実験参加者には，画面 4 で，勝敗を決めることを表す音声が流れ終わってから，画面 0 の注視点のみの画面が表示されるまでは，瞬きをせず，ディスプレイ中央の注視点を見続けるように教示した．

このカード分配から勝敗が決するまでの手続きを計 25 回表示することを 1 セットとして，合計 4 セット（計 100 回）実施した．セットとセットの間には休憩をとった．また，エージェントの表示位置による効果を統制するため，セットごとにエージェントの表示位置を左右逆にした．賛成エージェントと反対エージェントは，どちらも 45 回ずつ勝負に勝ち（負け），残りの 10 回は引き分けるように設定した．

また，実験参加者がエージェントに注意を向け続けるよう，各エージェントがゲームに関する学習を行っているように見せた．具体的には，ゲーム後半になるほど，5 以下の数字の場合にカードを交換する確率を高めるように設定した（5 以下で交換する確率を，ゲーム開始時の 50%から 1 セット終了するごとに 10%ずつ高めるように設定した）．

§フェーズ 3：アンケートフェーズ

最後に以下のアンケート調査を実施した．

1.「同側意見意識の発生確認アンケート」

ディスカッション時に，エージェントをどれだけ味方と感じたかを 5 件法で問うアンケートを行った．エージェントを敵と感じる場合を 1 点，味方と感じる場合を 5 点とした．

2.「共感能力評定アンケート」

実験参加者の共感能力を評定するためのアンケートを [Davis 83] に従って作成した．このアンケートは，28 の質問項目（5 件法）に答えることで，以下にまとめるような実験参加者の各種共感能力を評定することができる．（質問項目は付録 1 参照）

PT（視点取得）
自発的に他人の心理的立場をとろうとする傾向．

EC（共感的配慮）
不幸な他人に対して，同情やあわれみを感じる傾向．

PD（個人的苦痛）
他人の苦痛に反応して，こちらが不快や苦痛を経験する傾向．

FS（想像性）
想像上で自分を架空の状況の中に移し込む傾向．

3.「エージェント評価アンケート」
エージェントとのディスカッション内容に対する，実験参加者の主観評定を得るために，[水上 08] に基づいて，対義語からなる形容詞対 30 項目（表 1 参照）に対する印象評価を 5 段階で回答してもらった．[水上 08] では，この形容詞対のアンケート結果を因子分析し，5 つの因子が抽出されている．5 つの因子は，「場の活発度」「議論の多角と統合」「参加者の関係性」「議論の展開と洗練」「参加者の誠実さ」と解釈される．本研究では，アンケート結果をもとに，これら 5 因子の因子得点を算出し，ディスカッション内容の評定を行った．また，ディスカッション時の実験参加者のエージェントに対する共感度合いを直接的に見るため，「共感した」という形容詞を項目に追加した．
最後に，エージェントが人工知能によって操作されていると信じていたかどうかについて報告してもらった．

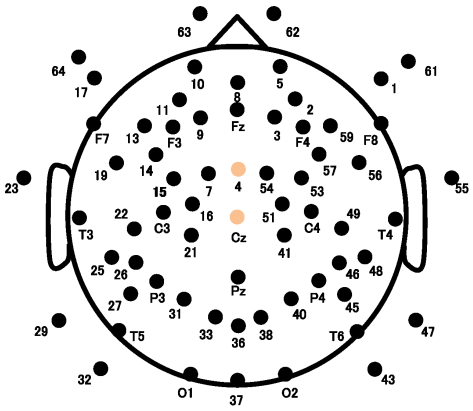

図 3: 電極位置

概要：インディアンポーカーを原型としたカードゲームで，
各プレーヤーは，相手の手札が見えない状況で，お互いの手持ちカード数字の大小を競う．
プレイ人数：2 人
使用カード：1〜10 の数字の書かれたカードが各 4 枚ずつ計 40 枚．
ゲームの目的：相手より大きな数字となるようにカードを交換し勝負に勝つこと．
手順：
①それぞれのプレーヤーにランダムに 1 枚ずつカードが配られる．
②カードを 1 回だけ交換する．(カードを交換しないこともできる．)
③最後にカードを見せ合い，大きい数字のカードを出したプレーヤーが勝ちとなる．
最適戦術：手持ちカードが 5 以下の場合だけ交換する．
※カード配布時，及びカード交換時のカード出現確率は常に一定で変化しないとする．

図 4: 対戦型カードゲームのルール

表 1: ディスカッション印象評価項目

1. 明るい	11. 視野の広い	21. 直線的な
2. にぎやかな	12. 真剣な	22. 協調的な
3. 打ち解けた	13. 注意深い	23. 対等な
4. 積極的な	14. 中立な	24. 連鎖的な
5. 参加している	15. コンパクトな	25. 発展している
6. 動きのある	16. 多面的な	26. 吟味された
7. 自然な	17. 共感した	27. 細かい
8. 開かれた	18. 均一な	28. 整然とした
9. スムーズな	19. 共有している	29. 深まりのある
10. 余裕のある	20. 一貫した	30. 心からの
		31. 固執した

3 分析方法

3.1 同側意見意識（フェーズ 1）に関する分析

アンケートの得点を同側意見意識の得点とした．ディスカッション時に実験参加者の意見に反対したエージェントの得点よりも，賛成したエージェントの得点の方が大きい場合に，同側意見意識が発生したとした．

3.2 ERP（フェーズ 2）の分析

記録された脳波を Net Station Waveform Tools を用いて，1~14Hz の Kaiser type の FIR（Finite Impulse Response）バンドパスフィルタに通し（Roll Off: 0.29Hz, Stop Band Gain : 1.0%），全電極平均を電位基準とした．次に，トリガ信号をもとに分析対象となるセグメントを切り出した．そして，200mV 以上の振幅の絶対値を持つ Bad Channel を周囲のチャンネルの値で補完した上で [Perrin 87]，blink slope threshold を 14[μV/msec.] として，Ocular Artifact Removal[Gratton 83] を行った．チャンネル補完時に，Bad Channel 数が全体の 20%を上回ったセグメントは分析対象から外すこ

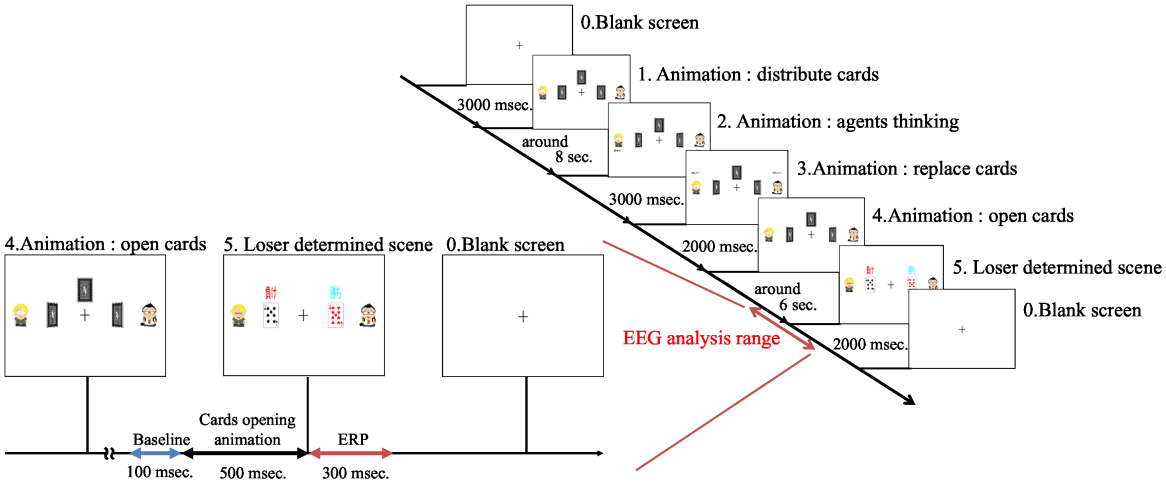

図 5: 実験手続き

ととした．ここまでの処理で，全体の約 12%のセグメントが分析対象から除外された．

勝敗が表示された時刻を 0msec.（図 5 の画像 5 の表示開始時刻に対応）とし，−600msec.∼300msec. を分析対象とした．このうち，−600msec.∼ −500msec. を baseline（図 5 下側参照）として，各試行の 0msec.∼300msec. を加算平均することで事象関連電位を算出した．加算平均は，ディスカッション時に実験参加者の意見に賛成したエージェントが負けた場合と，反対したエージェントが負けた場合とに分けて算出した．

また，使用した脳波計測システム（EGI 社 GES 300）において生じる，アンチエリアジングフィルタの影響によるトリガ時刻のズレ（8msec. の遅れ）を補正した．

3.3 共感能力アンケート（フェーズ 3）の分析

前節で説明した実験参加者の各共感能力値と，賛成・反対エージェント敗北条件間での ERN 振幅値の差分との間でスピアマンの順位相関係数を算出した．

また，ディスカッション時のエージェントの印象評価アンケートの分析では，まず，[水上 08] で得られている因子得点係数を用い，今回のアンケート結果から，各因子得点を計算し，その得点と，ERN 振幅差分との間でピアソンの積率相関係数を算出した（今回追加した「共感した」という質問項目は，単独の要因として扱った）．

4 結果

実験参加者 30 名のうち，脳波測定時にエージェントの名前や顔を忘れていたことが判明した者 1 名，疾患のため脳波測定時にエージェントの画像が視野に入りきらなかったと報告した者 1 名，実験者がディスカッション時に賛成エージェントの設定を取り違える失敗を行った可能性がある者 1 名，エージェントが人工知能によって操作されていたことに疑いをもった者 3 名，の計 6 名を分析対象から除いた．その結果，24 名を分析対象とした（男性：17 名，女性：7 名，平均年齢：21.92 歳，標準偏差：4.11 歳）．

4.1 同側意見意識の発生に関する分析結果

3.1 節に従って同側意見意識を得点化した．結果は，以下の通りであった．

ディスカッション時に実験参加者の意見に賛成したエージェントに対する同側意見意識の平均点は 4.33 であった（標準偏差：0.702，最大値：5，最小値：3）のに対して，意見に反対したエージェントに対する同側意見意識の平均点は 1.38 であった（標準偏差：0.576，最大値：3，最小値 1）．そして，賛成エージェントの得点から反対エージェントの得点を引いた値の平均値は 2.96 であった（標準偏差：0.955，最大値：4，最小値：0）．対応のある t 検定の結果，賛成，反対の各エージェントに対する同側意見意識の得点に有意差がみられた（$t\,(23) = 15.18, p < .01$）．

この結果から，ディスカッション時のエージェントの発言の操作によって，賛成エージェントに対する同

側意見意識が発生していたことが確認された．

4.2 ERP の分析結果

図 6 に，勝敗の決着後 178msec. から 198msec. 後までの 5msec. ごとのトポグラフィックマップを示す．これから，中前頭頂部付近に負の活動電位の集中がみられることがわかる．また，賛成エージェントの敗北時と反対エージェントの敗北時のトポグラフィックマップを比較すると，頭頂から前頭にかけての領域の電位分布に差分が存在することがわかる．

ERN は頭頂（図 3 の Cz）から前頭（図 3 の Fz）にかけて最も強く表れるといわれている [Scheffers 00]．また，多くの先行研究 [Gehring 93] で Cz や FCz（図 3 の no.4），Fz 周辺に着目して ERP の分析を行っている．したがって本研究でも，図 6 のトポグラフィックマップで，ERP の空間上のピークに近い Cz，FCz について分析する．

Cz，FCz の事象関連電位の波形を図 7 に示す．賛成エージェント敗北条件（friend）と反対エージェント敗北条件（foe）の間で，200msec. より少し前に位置するピークに違いがみられる．実験参加者ごとの ERP データを用いて，ERP 振幅ピーク（Cz:180msec.，FCz:190msec.）を中心に ±10msec. の平均電位を求め，賛成エージェント敗北時と，反対エージェント敗北時の値に有意な差があるかどうかを，対応のある t 検定で検証した．その結果，有意差がみられた（Cz:$t(23) = -2.243$, $p < .04$; FCz:$t(23) = -2.304$, $p < .04$）（図 7 の右側参照：図中のエラーバーは標準誤差）．さらに，ピーク周辺の各時刻で個別に分析した結果，Cz：167∼193msec. と FCz：175∼195msec. の各時刻に有意差がみられた（$p < .05$）．

4.3 共感能力アンケートの分析結果

反対エージェント敗北時の Cz：170∼190msec. 区間平均電位から，賛成エージェント敗北時の同区間平均電位を差し引いた値と，共感能力値（PD：個人的苦痛）との間に，有意傾向のあるスピアマンの順位相関がみられた（$R = 0.3522, p < .10$. 表 2）．特に，分析区間前半の 170∼180msec. の平均振幅差分との間には，有意なスピアマンの順位相関がみられた（$R = 0.4247, p < .04$. 表 2）．さらに，ピーク周辺での各時刻振幅差分に個別に注目すると，Cz：167∼175msec. の各時刻振幅差分と，共感能力値（PD：個人的苦痛）の間に，有意なスピアマンの順位相関がみられた（$p < .05$）．例えば，175msec. では（表 2：$R = 0.425, p < .04$）であった．

表 2: 共感能力値と各種 ERP 振幅差分値（Cz）との相関分析（上段：spearman の順位相関係数，下段：p-value）

共感能力	FS	PT	EC	PD
FS	1.000			
p-value	0.000			
PT	-0.074	1.000		
p-value	0.730	0.000		
EC	0.091	0.202	1.000	
p-value	0.671	0.343	0.000	
PD	0.067	-0.083	-0.037	1.000
p-value	0.755	0.700	0.863	0.000
平均 ERP 振幅差分	0.101	-0.186	-0.260	0.352 †
170msec.∼190msec.	0.639	0.385	0.220	0.091
平均 ERP 振幅差分	0.115	-0.097	-0.247	0.425 *
170msec.∼180msec.	0.594	0.653	0.244	0.039
ERP 振幅差分	0.111	-0.112	-0.247	0.425 *
175msec.	0.607	0.603	0.244	0.039

$\dagger : p < .1, * : p < .05, ** : p < .01$

4.4 エージェント評価アンケートの分析結果

平均 ERP 振幅差分値（Cz：170∼190msec.）と，ディスカッション印象因子との間のピアソンの積率相関を調べたところ，表 3 のような結果が得られた．

味方エージェントへの印象のうち，活発度との間に有意な負の相関（$R = -0.561, p < .05$）が，展開洗練度との間に有意傾向のある負の相関（$R = -0.356, p < .09$）がみられた．敵エージェントとの間では，「共感した」との間に有意な負の相関（$R = -0.414, p < .05$）が，関係性との間に有意傾向のある負の相関（$R = -0.395, p < .06$）がみられた．

5 総合考察

5.1 ERP による内集団形成の検出の検討

4 章でのデータ解析の結果，200msec. より少し前に，頭頂から前頭にかけて分布し，賛成・反対エージェントの違いに応じて変化する，ERN だと考えられる負の ERP が観測された．この結果から，内集団関係の存在が示されるかどうかについて検討する．

4.3 節の解析結果より，ERP 振幅差分値と実験参加者の「他者の苦痛を自分のものとして感じる」共感能力値（PD）との間に相関関係があるとわかった．したがって，1 章で論じた fMRI を用いた Newman-Norlund の研究 [Newman-Norlund 09] と同様に，観測された ERP は，他者の視点や情動の共有と関係していると推定される．その ERP の振幅に，賛成エージェント敗北時と反対エージェント敗北時で有意な差分が確認されることから，実験参加者と賛成エージェントとの間に内集団関係が存在していたといえよう．

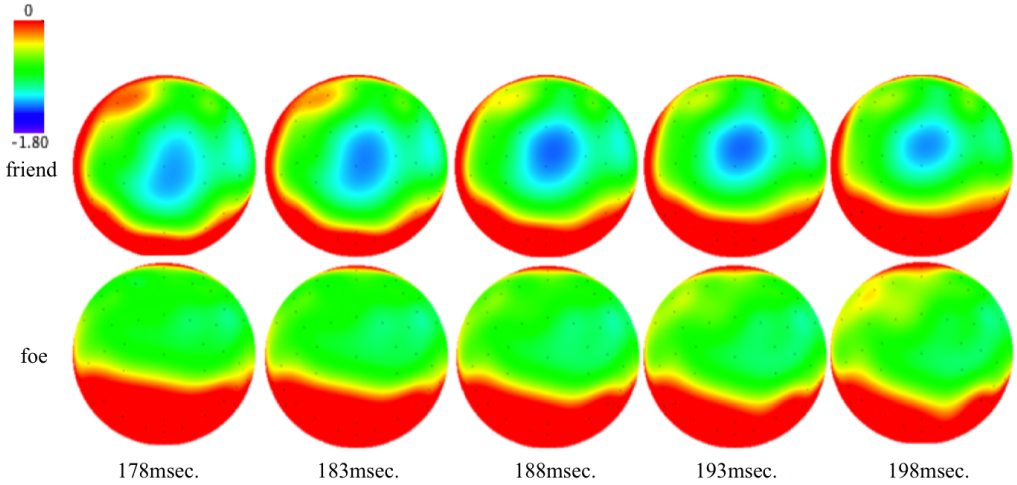

図 6: 脳波トポグラフィックマップ

後述するように，先行研究と比較すると，他者の失敗行動観察時にみられる ERN としては，本研究で観測された ERP のピーク潜時は少し早いものであった．したがって，本研究で観測された ERP が，ERN であったかどうかについての検討を行う．

フランカータスク[1]のように，行動を起こした瞬間に自身の失敗がわかる課題を用いた研究にみられる ERN のピーク潜時は 80∼150msec. である [Gehring 93]．これに対して，ギャンブル課題や，本研究のような他者の課題実行を観察する課題のように，行動に応じた結果（正解・不正解）が表示された後に，その行動が失敗であったかどうかが判明するような場合には，ERN ピーク潜時が 200∼250msec. 前後になることが報告されている [Fukushima 09, Schie 04]．一方，本研究における ERP のピーク潜時は 180msec. 前後であり，上記の先行研究と一致しない．しかしながら，本研究で脳波測定時に用いた画像刺激は，これらの先行研究とは異なる画像刺激（独自のカードゲーム）を使用したものであり，直接的な比較はできない．さらに，エージェントの勝敗が判明した瞬間に，そのことを文字と色で明確に表示しており，比較的早期に実験参加者が勝敗を認識できるようにもなっていた．そして，以下の 3 点から，この ERP を，エージェントがゲームで負けたことを観察したことによって生じた ERN だと判断した．1 点目は，トポグラフィックマップ（図 6）が先行研究と同様に頭頂から前頭にかけて，負の電位分布を持っていたこと [Schie 04] である．2 点目は， ERP のピーク振幅差分の大きさと実験参加者の共感能力値（PD）との間に有意な相関があったことである．これを理由として挙げたのは，ERN の発生源であると考えられる ACC が，共感と関係しているといわれていること [Decety 04] ，そして今回用いた共感能力値の指標とは異なるものの， ERN の振幅の大きさが共感能力の高さと相関することが先行研究で報告されているためである [Santesso 09]．3 点目は，実験参加者にディスカッションを通して生じた，反対エージェントに対する共感や関係性に関する印象と，ERP 振幅差分が負の相関を持っていることである．これを挙げた理由は，この結果が，ERP の振幅差分と実験参加者の視点との関係，すなわち，反対エージェントを敵とする，賛成エージェントとの視点の共有との関係を示唆していると考えられるからである．

次に，ERP の振幅差分とディスカッション時のエージェントに対する印象との関係について考察する．4.4 節の解析結果から推察すると，実験参加者は，賛成エージェントが弱々しく拙い議論をし，それに対して反対エージェントが，実験参加者の共感できない内容や態度で反論を行っていると感じられた時，賛成エージェントに対して相対的に共感を持つと考えられる．したがって，今回の実験設定で実験参加者が賛成エージェントに共感を抱いた要因としては，同じ意見を共有していたということだけでなく，賛成エージェントがディスカッション時に苦しんでいるのを見て，同情のようなものを抱き，それがその後の脳波測定時（カードゲーム時）にも影響したということも考えられる．

最後に，内集団関係が単にエージェントの見た目の違いによって生じていた可能性がないか検討する．これを調べるために，外見要因（アーノルドかチャールズかの 2 水準，参加者間要因）と同側意見意識要因（賛成エージェントか反対エージェントかの 2 水準，参加者

[1]フランカータスクとは，モニタなどに表示される競合的 (>><>>) または非競合的 (<<<<<) なディストラクタに挟まれた中央の矢印の向きを，ボタン押しなどで答えさせるタスクのことで，できるだけ早く答えさせることで，一定の誤答率を持続できることから，ERN 測定時によく利用される．

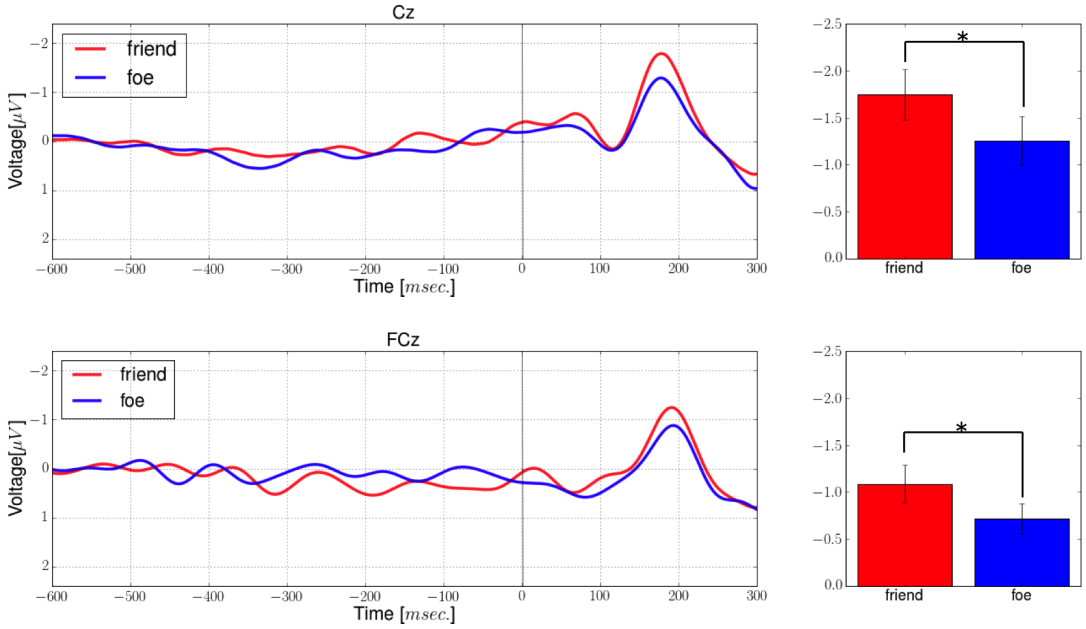

図 7: ERP（左）とピーク周辺平均振幅のグラフ（右）（上段：Cz，下段：FCz）

表 3: ディスカッション時印象評定と平均 ERP 振幅差分（Cz：170msec.∼190msec.）の相関分析（上段：pearson の相関係数，下段：p-value）※項目「共感した」は，ディスカッション時の共感を意味する．

	因子	1	2	3	4	5	6	7	8	9	10	11	12	13
	1. 活発度	1.000												
	p-value	0.000												
	2. 多角統合	0.778 **	1.000											
	p-value	0.000	0.000											
	3. 関係性	0.487 *	0.452 *	1.000										
	p-value	0.016	0.027	0.000										
friend	4. 展開洗練	0.804 **	0.819 **	0.611 **	1.000									
	p-value	0.000	0.000	0.002	0.000									
	5. 誠実さ	0.640 **	0.774 **	0.536 **	0.513 *	1.000								
	p-value	0.001	0.000	0.007	0.010	0.000								
	6. 共感した	0.704 **	0.633 **	0.556 **	0.500 *	0.773 **	1.000							
	p-value	0.000	0.001	0.005	0.013	0.000	0.000							
	7. 活発度	0.302	0.223	0.339	0.367 †	0.036	0.099	1.000						
	p-value	0.151	0.295	0.105	0.077	0.869	0.645	0.000						
	8. 多角統合	0.461 *	0.398 †	0.124	0.552 **	0.073	0.141	0.643 **	1.000					
	p-value	0.023	0.054	0.563	0.005	0.735	0.510	0.001	0.000					
	9. 関係性	0.149	0.172	0.100	0.209	-0.068	0.008	0.620 **	0.604 **	1.000				
	p-value	0.488	0.423	0.642	0.328	0.753	0.972	0.001	0.002	0.000				
foe	10. 展開洗練	0.461 *	0.392 †	0.264	0.582 **	0.185	0.170	0.360 †	0.752 **	0.368 †	1.000			
	p-value	0.023	0.058	0.213	0.003	0.388	0.428	0.084	0.000	0.077	0.000			
	11. 誠実さ	0.228	0.290	0.252	0.298	0.171	0.171	0.507 *	0.639 **	0.474 *	0.419 *	1.000		
	p-value	0.284	0.169	0.236	0.158	0.424	0.424	0.012	0.001	0.019	0.041	0.000		
	12. 共感した	-0.033	-0.173	0.159	-0.018	-0.227	-0.278	0.517 *	0.323	0.723 **	0.265	0.212	1.000	
	p-value	0.877	0.420	0.458	0.933	0.286	0.188	0.010	0.124	0.000	0.210	0.320	0.000	
	13. 平均 ERP 差分	-0.561 **	-0.242	-0.310	-0.356 †	-0.128	-0.247	-0.344	-0.269	-0.395 †	-0.216	-0.045	-0.414 *	1.000
	p-value	0.004	0.254	0.141	0.088	0.551	0.244	0.100	0.203	0.056	0.310	0.834	0.044	0.000

$\dagger : p < .1, * : p < .05, ** : p < .01$

内要因）を独立変数，Cz，FCz のそれぞれでのピーク（Cz:180msec.，FCz:190msec.）を中心に ±10msec. の平均振幅を従属変数として，二元配置分散分析を行った．この結果，同側意見意識要因の主効果は Cz，FCz で有意だったのに対して，外見要因の主効果は FCz でのみ有意傾向を示した（Cz: 同側集団意識要因 : $F(1, 23) = 4.797, p < .04$，外見要因 : $F(1, 23) = 0.020, p > .8$; FCz: 同側意見意識要因 : $F(1, 23) = 5.329, p < .04$，外見要因 : $F(1, 23) = 3.214, p > .08$）．このことから，内集団関係の醸成において，同側意見意識要因は外見要因とは独立して効果をもち，一方で，外見要因の効果があった可能性は相対的に低いことが示唆される．

以上の議論をまとめると，ディスカッション中にエージェントが実験参加者の意見に賛成・反対することによって，実験参加者とエージェントとの間に内集団関係を醸成でき，それが ERN を測定することで検出可能だといえよう．

5.2 ERP 時系列データからの内集団関係形成の検出

前述の統計的分析の際に着目したピーク位置などの波形の詳細な構造は，実験設定や実験参加者の個人差，脳波測定機の特性などによって変動しやすい．従って，今回の分析結果をそのままエージェント開発時の性能評価スキームとして利用することは難しい．そこで工学的な応用を念頭に，ERP 時系列データから直接内集団関係形成の有無を検出できるかを検討した．

そのために次のようなデータセットを作成した．説明のため，ディスカッション時に味方としたエージェントを friend，敵としたエージェントを foe とする．ある実験参加者 i のそれぞれのエージェントに対する電極位置 Cz の $0 \sim 300msec.$ の ERP 時系列データを $\vec{X}^i_{friend}$ と $\vec{X}^i_{foe}$ とする．ここから $(\vec{X}^i_{friend} - \vec{X}^i_{foe})$ なる振幅差分時系列データを作成し，このデータに内集団関係があることを示す $True$ ラベルを設定した．一方，$(\vec{X}^i_{foe} - \vec{X}^i_{friend})$ なる振幅差分時系列データも作成し $False$ ラベルを設定した．以上の操作によって，$x^i = \{(\vec{X}^i_{friend} - \vec{X}^i_{foe}), (\vec{X}^i_{foe} - \vec{X}^i_{friend})\}$, $y^i = \{True, False\}$, $i = 0, 1, 2, \cdots 24$ とした，参加した実験者（24 人）の 2 倍のサンプル数をもつデータセットが作成された．

このデータセットを下式のような損失関数を持つ，線形 Support Vector Machine を用いて識別した．

$$E = C \sum_i (y^i - f(\vec{x}^i))^2 + \mathbf{W^T W} \tag{1}$$

$$f(\vec{x}^i) = \sum_k \mathbf{W^T} \phi \tilde{\mathbf{x}}^\mathbf{i} + \tilde{\mathbf{b}}^\mathbf{i} \tag{2}$$

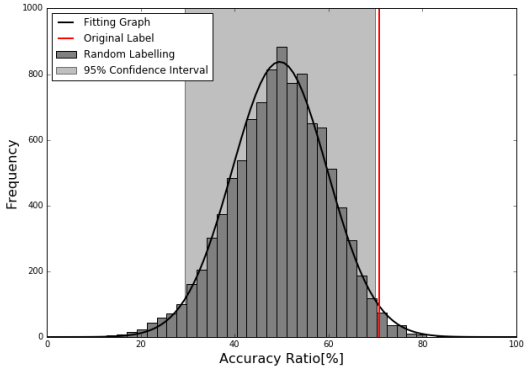

図 8: Permutation テストの結果

C は，モデルの複雑さを表現する正則化項と二乗誤差項のバランスを決定する重み係数である．

SVM の識別精度は leave-one-out 交差検証法によって推定され，その識別精度が優位に優れた値であることを Permutation テストによって検定した（正解ラベルをランダムに振りなおし，その際の交差検証エラーの分布を調べ，得られた識別精度がその信頼区間に入っているかどうかで，それが優位に優れているかを検定する）．C は，交差検証エラーが最も小さくなる 0.8 に設定した．

SVM のトレーニングと leave-one-out 交差検証の結果，70.83% の識別精度が得られることがわかった．そこで，Permutation テストを行ったところ，その精度は優位に優れたものであった（図 8）．

このように，ERP 時系列データから内集団関係形成を直接検出することに成功した．

6 結論

本研究ではこのように，人間と架空エージェントとの間にインタラクションを通して醸成された内集団関係を，ERP という客観的で容易な手法によって検出することに成功した．特に ERP 時系列データからの直接検出に成功したことは，エージェントシステムの開発に際して，本研究の枠組みを容易に適用できることを意味する．本研究の結果に基づいた HAI 研究の進展が期待される．

参考文献

[Brewer 86] Brewer, M. B. and Kramer, R. M.: Choice behavior in social dilemmas: Effects of

social identity, group size, and decision framing., *Journal of personality and social psychology*, Vol. 50, No. 3, p. 543 (1986)

[Davis 83] Davis, M. H.: Measuring individual differences in empathy: Evidence for a multidimensional approach., *Journal of personality and social psychology*, Vol. 44, No. 1, p. 113 (1983)

[デイヴィス 99] デイヴィス M.H:共感の社会心理学 人間関係の基礎（菊池章夫訳）, 川島書店, 東京 (1999)

[Decety 04] Decety, J. and Jackson, P. L.: The functional architecture of human empathy, *Behavioral and cognitive neuroscience reviews*, Vol. 3, No. 2, pp. 71–100 (2004)

[Dehaene 94] Dehaene, S., Posner, M. I., and Tucker, D. M.: Localization of a neural system for error detection and compensation, *Psychological Science*, Vol. 5, No. 5, pp. 303–305 (1994)

[Fukushima 09] Fukushima, H. and Hiraki, K.: Whose loss is it? Human electrophysiological correlates of non-self reward processing, *Social neuroscience*, Vol. 4, No. 3, pp. 261–275 (2009)

[Gehring 93] Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., and Donchin, E.: A neural system for error detection and compensation, *Psychological science*, Vol. 4, No. 6, pp. 385–390 (1993)

[Gratton 83] Gratton, G., Coles, M. G., and Donchin, E.: A new method for off-line removal of ocular artifact, *Electroencephalography and clinical neurophysiology*, Vol. 55, No. 4, pp. 468–484 (1983)

[Ito 03] Ito, S., Stuphorn, V., Brown, J. W., and Schall, J. D.: Performance monitoring by the anterior cingulate cortex during saccade countermanding, *Science*, Vol. 302, No. 5642, pp. 120–122 (2003)

[Kurzweil 90] Kurzweil, R., Richter, R., and Schneider, M. L.: *The age of intelligent machines*, Vol. 579, MIT press Cambridge (1990)

[Leng 10] Leng, Y. and Zhou, X.: Modulation of the brain activity in outcome evaluation by interpersonal relationship: an ERP study, *Neuropsychologia*, Vol. 48, No. 2, pp. 448–455 (2010)

[水上 08] 水上 悦雄, 森本 郁代, 鈴木 佳奈, 大塚 裕子, 竹内 和広, 東新 順一, 奥村 学, 柏岡 秀紀：話し合いにおけるコミュニケーションプロセスの評価法について, 言語処理学会第 14 回年次大会発表論文集, pp. 181–184 (2008)

[Newman-Norlund 09] Newman-Norlund, R. D., Ganesh, S., Schie, van H. T., De Bruijn, E. R., and Bekkering, H.: Self-identification and empathy modulate error-related brain activity during the observation of penalty shots between friend and foe, *Social Cognitive and Affective Neuroscience*, Vol. 4, No. 1, pp. 10–22 (2009)

[小野 06] 小野 哲雄：インタラクションにおけるカップリングと知能 (< 特集> HAI: ヒューマンエージェントインタラクションの最先端), 人工知能学会誌, Vol. 21, No. 6, pp. 662–668 (2006)

[Perrin 87] Perrin, F., Bertrand, O., and Pernier, J.: Scalp current density mapping: value and estimation from potential data, *IEEE Transactions on Biomedical Engineering*, No. 4, pp. 283–288 (1987)

[Reeves 96] Reeves, B. and Nass, C.: *How people treat computers, television, and new media like real people and places*, CSLI Publications and Cambridge university press Cambridge, UK (1996)

[Santesso 09] Santesso, D. L. and Segalowitz, S. J.: The error-related negativity is related to risk taking and empathy in young men, *Psychophysiology*, Vol. 46, No. 1, pp. 143–152 (2009)

[Scheffers 00] Scheffers, M. K. and Coles, M. G.: Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors., *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 26, No. 1, p. 141 (2000)

[Schie 04] Schie, van H. T., Mars, R. B., Coles, M. G., and Bekkering, H.: Modulation of activity in medial frontal and motor cortices during error observation, *Nature neuroscience*, Vol. 7, No. 5, pp. 549–554 (2004)

[竹内 00] 竹内 勇剛, 片桐 恭弘 他：ユーザの社会性に基づくエージェントに対する同調反応の誘発, 情報処理学会論文誌, Vol. 41, No. 5, pp. 1257–1266 (2000)

[Wann 93] Wann, D. L. and Branscombe, N. R.: Sports fans: Measuring degree of identification with their team., *International Journal of Sport Psychology* (1993)

[Wann 01] Wann, D. L., Hunter, J. L., Ryan, J. A., and Wright, L. A.: The relationship between team identification and willingness of sport fans to consider illegally assisting their team, *Social behavior and personality: an international journal*, Vol. 29, No. 6, pp. 531–536 (2001)

A IRI(Interpersonal Reactivity Index) 質問項目 [Davis 83]（日本語訳は [デイヴィス 99] による）

以下の質問項目について,「よく当てはまる」「当てはまる」「どちらでもない」「当てはまらない」「全く当てはまらない」の 5 件法で回答．それぞれ，4 点，3 点，2 点，1 点，0 点の得点に対応.

1. 自分に起こることについて，繰り返し，夢見たり想像したりする．(FS)
2. 自分より不幸な人々について，やさしさや配慮の感情を持つことが多い．(EC)
3. 「他人の」立場から物事を考えるのが，むずかしいと思うことがある．(PT)(R)
4. 他の人びとが問題をかかえていても，気の毒に思うことはあまりない．(EC)(R)
5. 小説の登場人物の感情に，本当にのめり込んでしまう．(FS)
6. 緊急の事故などに出合うと，気になって，落ち着いていられない．(PD)
7. 映画や演劇を見ても，だいたい客観的な態度でいられて，完全にのめり込んでしまうことは少ない．(FS)(R)
8. 物事を決めるには，みんなの反対意見をよく聞いてからにしようとする．(PT)
9. 相手に付け込まれている人を見たときには，その人を守ってあげようという気になる．(EC)
10. 気持ちが落ち着かない場面に出合った際には，独りぼっちだと感じることがある．(PD)
11. 友だちのすることを理解しようとするときには，向こうから見るとどう見えるのかを想像することがある．(PT)
12. 良い本や映画に完全にのめり込んでしまうことは，きわめてまれである．(FS)(R)
13. 誰かが傷ついているのを見ても，平気でいられる．(PD)(R)
14. 他の人びとの不幸が気になることはほとんどない．(EC)(R)
15. 自分が正しいと確信している場合には，他人の意見を聞くのに時間を使ったりはしない．(PT)(R)
16. 演劇や映画を見た後には，自分がその登場人物の一人だったと感じる．(FS)
17. 気持ちが極端に落ちつかなくなる場面に出合うと，恐ろしくなる．(PD)
18. 公正でない扱いをされている人を見ても，かわいそうにと思うことは少ない．(EC)(R)
19. 突発の出来事を処理するのが，いつも上手である．(PD)(R)
20. たまたま出合った出来事によって，気持ちが動かされることが多い．(EC)
21. どんな問題にも 2 つの面があるから，その両方を見るようにつとめている．(PT)
22. 自分はやさしい気持ちの人間だと思っている．(EC)
23. よい映画をみたときには，すぐに主人公の立場に自分を置くことができる．(FS)
24. 突発の出来事に出合うと，自分を抑えることができなくなりがちである．(PD)
25. 相手に腹を立てている時でも，しばらくは「相手の立場に立とう」とすることが多い．(PT)
26. 面白い物語や小説を読んだ際には，話の中の出来事がもしも自分に起きたらと想像する．(FS)
27. 事故にあって助けを求めている人に出会ったら，どうしたらよいか分からなくなる．(PD)
28. 相手を批判する前に，自分が相手の立場だったらどう感じるかと想像してみようとする．(PT)

PT：視点取得，EC：共感的配慮，PD：個人的苦痛，FS：想像性.
(R) の記載のある項目は得点配分を逆転.