Win-Win 関係構築のための感情認知計算
Affective Computing for building Win-Win relationship

佐藤 幹晃 †，寺田 和憲 †，ジョナサン グラッチ ‡
Motoaki Sato, Kazunori Terada, Jonathan Gratch
† 岐阜大学，‡ 南カリフォルニア大学
Gifu University, University of Southern California
motoaki.sato@ai.info.gifu-u.ac.jp

概要

Win-Win な交渉をするためには，事前にコミュニケーションすることで相手について学ぶことが重要である．人-人の交渉において，感情表現は偽りがなく，信頼できる信号だと考えられているが，人-AI エージェントの交渉結果にどのような影響を与えるかは未知であった．そこで本研究では，交渉前のコミュニケーションで感情表現から AI エージェントの選好を学習することが，Win-Win な交渉結果に寄与するか検討した．

キーワード：交渉, 表情， 評価理論， 心的状態の推測

1. はじめに

近年のバーチャルリアリティ（VR）と人工知能（AI）の進歩により，言語・非言語コミュニケーションを通じて人と交渉する具現化 AI エージェントが実現されつつある [1]．交渉では，交渉前や交渉中にチープトークを通じて様々な情報が交換され，信頼などの相手の印象や選好などの心理状態といった交渉相手のモデルを形成している．チープトークとは，ゲーム理論において，ゲームのペイオフに直接影響を与えないプレイヤー間のコミュニケーションのことである．AI エージェントは後ろめたさを感じることなく嘘をつくことができるため，競争的な状況ではエージェントのチープトークを無視した方が良いが，協力的な状況では無視しない方が Win-Win な解決に至ることができる [2]．非言語によるチープトークは交渉に有効である [3]．非言語信号が効果的な理由の 1 つは，口頭での情報よりもより正直で信頼性が高い [4] ため，偽ることが難しいからである．したがって，人間が AI エージェントとの交渉で Win-Win となるためには，交渉前に交わされる非言語情報を用いて，人間が AI エージェントをどのようにモデル化しているのかをより知る必要がある．本研究では，交渉前の感情表現から AI エージェントの選好を学習することで，人間と AI との交渉において Win-Win な解決策を導くことができるかどうかを検討した．

2. 実験方法

実験参加者は Yahoo！クラウドソーシングで募集した，19 歳から 70 歳までの男性 119 人，女性 28 人であった ($M_{age} = 46.31$, $SD_{age} = 11.092$)．ID の不一致とアテンションチェック該当者を取り除き，147 人 (選好学習フェーズ有:75 人，無: 参加者 72 人) をデータとして集計した．

参加者は Yahoo!クラウドソーシングから Qualtrics で作成したアンケートページに進み，年齢と性別を答えたのち，複数論点最後通牒ゲームの説明を受け，理解を確認する問題に答えた．次に参加者は 16 種類の食べ物に 1 から 4(1: 嫌い，2: どちらでもない，3: まあまあ好き，4: とても好き) の水準で評価した．この 4 段階に少なくとも 1 つ以上の食べ物が割り当てられるよう，選好がばらけるような食べ物を事前に選定した．加えて，選好が極端に偏る参加者を考慮し，各水準に対して，少なくとも 1 つ以上の食べ物を割り当てるように強制した．各水準からランダムに 1 つの食べ物を選択し，4 つのアイテムを後のタスクに用いた．プリプレイコミュニケーションを行う前にエージェントの選好をどのように推測したかを評価するために，参加者はエージェントの選好を推定し，1 から 4 まで順位をつけた．

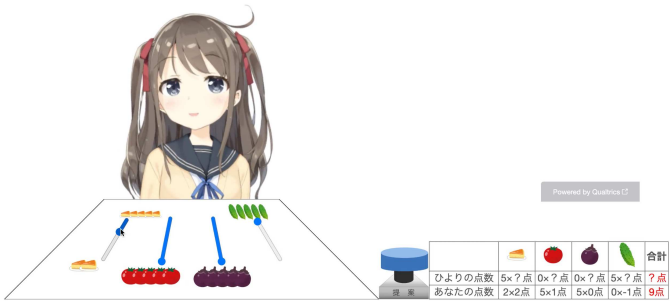
次に，全ての参加者は複数論点最後通牒ゲームを行った．参加者はスライダーバーを用いて，アイテムの配分をエージェントに提示することができた．表 1dのような重みの設定であり，配分されたアイテムの重みの合計がポイントとなった．エージェントはポイントが高いほどより喜んだ表情を表出した．エージェントの喜びの表情は，喜びの段階を適切に表しているか妥当性を検証した 8 段階を用いた (図 1f)．エージェントは許容できる限界のポイント (リミット) を下回ると怒った表情を表出した．参加者がリミット以上で配分を決定すると，参加者とエージェントともに配分に

(a) アイテムごとに表出するエージェントの表情 ($w^{\text{エージェント}} = -1$ (嫌い), $w^{\text{エージェント}} = 0$ (どちらでもない), $w^{\text{エージェント}} = 1$ (まあまあ好き), $w^{\text{エージェント}} = 2$ (とても好き).

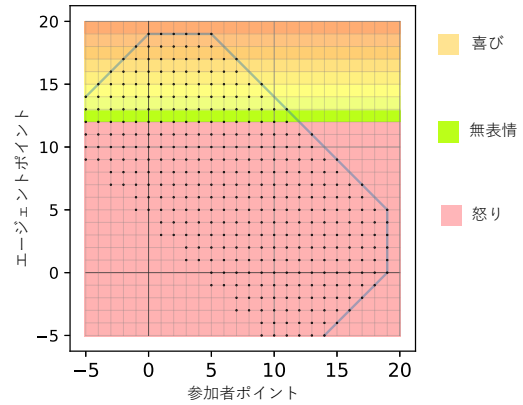
(b) 選好学習フェーズ

アイテム (個数)	A (7)	B (5)	C (5)	D (5)
$w^{\text{参加者}}$	2	1	0	-1
$w^{\text{エージェント}}$	2	0	-1	1

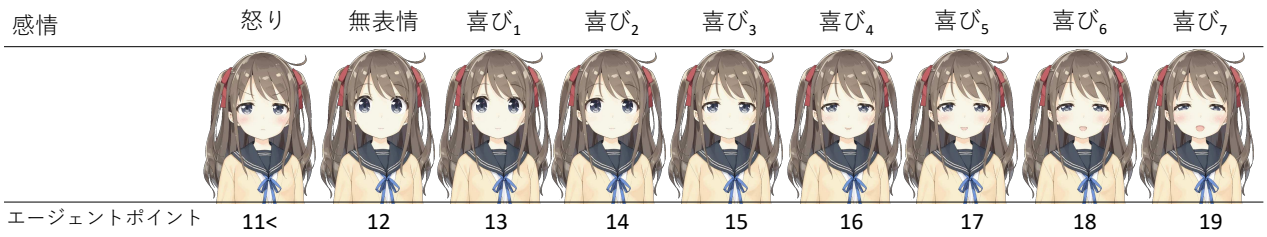
(d) アイテムの重み



(e) 複数論点最後通牒ゲーム



(c) エージェントが配分をどのように評価するかと複数論点最後通牒ゲームにおけるパレート最適を示した評価理論のヒートマップ



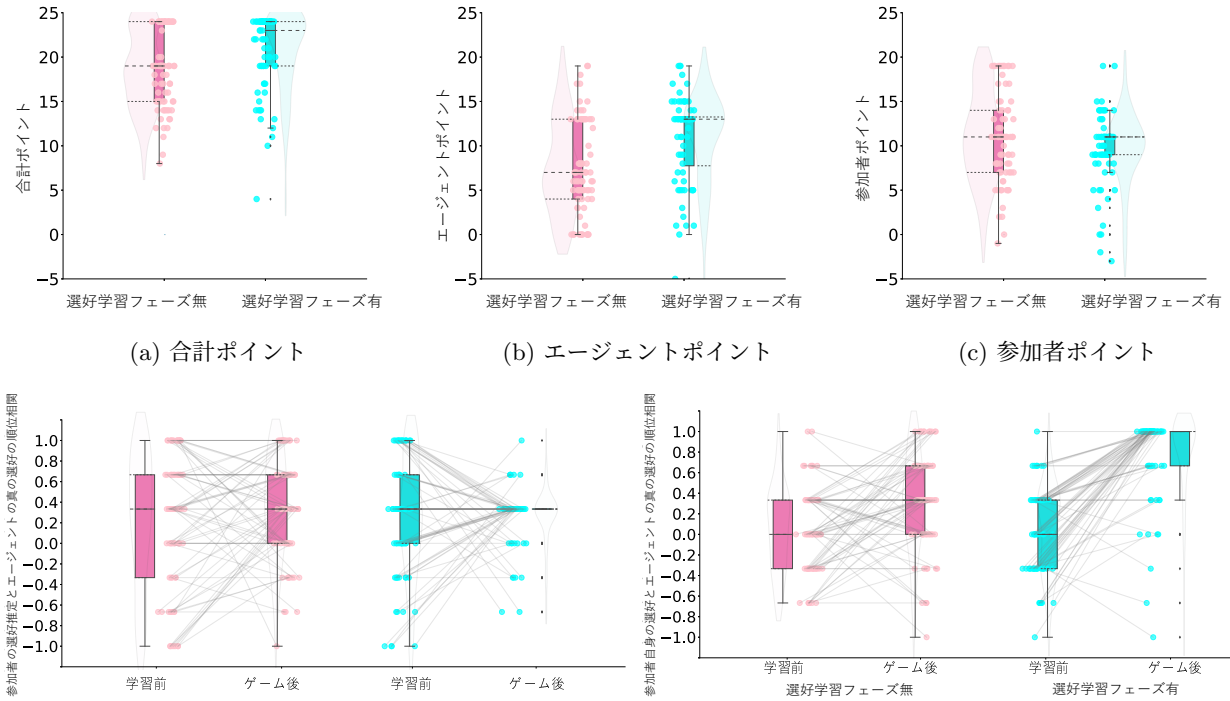
(f) エージェントの利得に応じてエージェントの表情が変化する

図 1: 方法: 選好学習のインタフェース

応じたポイントを獲得した．参加者がリミット未満で配分を決定すると，エージェントに配分を拒否されて両者ともにポイントを獲得できなかった．図 1cのヒートマップはエージェントのポイントと表情の対応関係を示している．参加者の課題は，エージェントに拒否されることなく参加者のポイントを増やすことであった．そのために，参加者はエージェントの好みとエージェントのリミットを推論する必要があった．図 1eのインタフェースを用いることで，エージェントの選好とリミットの同時推論が理論的に可能である．

しかし，選好とリミットを同時に推論することは難しい [5]．そこで本研究では，選好学習フェーズ有条件の参加者のみに，交渉前に選好学習を行う機会を与えた．選好学習フェーズでは，選好を符号化しているエージェントの表情を観察することでエージェントの選好を学習した (図 1b)．エージェントの選好は $w^{\text{エージェント}} = -1, 0, 1, 2$．エージェントが嫌いな問題には-1 が，エージェントがとても好きな問題には 2 が割り当てられている．$w^{\text{エージェント}}$ と表情の関係を図 1aに示す．最後に全ての参加者は再度エージェントの選好推定を行った．

交渉結果に影響を与える可能性のある重要な特徴の 1 つは，エージェントの見た目のリアルさや単純さである．寺田らは，単純な線画からなるソフトウェア・エージェントを用いて，最後通牒ゲームにおいて口の角度が上下するにつれて人々の向社会的行動が変化することを示した [6]．deMelo らは，実際の人間に似たエージェントを用い，囚人のジレンマにおいて，エー



(a) 合計ポイント (b) エージェントポイント (c) 参加者ポイント

(d) 固定パイ・バイアス (e) 選好推定

図 2: 実験結果

ジェントが競争的な感情表現と比較して協力的な感情表現を示すと，人々はより協力的になることを確認した [7]．そこで本研究では，単純な見た目と実際の見た目の中間であるアニメ風のキャラクターをエージェントとして用いた．

参加者には，参加費として PayPay ポイントを 200 ポイント支払った．さらに，複数論点最後通牒ゲームで獲得した参加者自身のポイントに応じて，より多くのお金を獲得する機会を与えた．今回の実験では 150 人中上位 5%，つまり 7 人の参加者に Amazon ギフト券 3000 円分を与えた．

本研究は，岐阜大学大学院医学系研究科の医学審査委員会 (IRB) の承認を得ている．参加者には，データセットに参加者の情報が使用されることについて，インフォームドコンセントを行った．

複数論点最後通牒ゲームにおいて，エージェントが獲得したポイント，参加者が獲得したポイント，両者の合計ポイントを検証した．また，参加者が推定したエージェントの選好順位を測定し，[8] に倣い，参加者が推定したエージェントの選好と参加者の選好 (固定パイ・バイアス)，参加者が推定したエージェントの選好とエージェントの実際の選好 (選好推定) の間のケンドールの順位相関係数をそれぞれ算出した．実験計画は 2(選好学習フェーズ: 有/無) の参加者間要因配置で行った．エージェントが獲得したポイント (エージェントポイント)，参加者が獲得したポイント (参加者ポイント)，両者が獲得したポイント (合計ポイント) について，2 水準 (選好学習フェーズ: 有/無) の一元配置分散分析を行なった．参加者は，複数論点最後通牒ゲームの前と後の 2 回，エージェントの選好を推定したので，固定パイ・バイアスと選好推定について，最初・最後の 4 アイテムの複数論点最後通牒ゲームのインタラクション要因 (学習前/ゲーム後水準: 参加者内) × 選好学習フェーズの有無の外見要因 (選好学習フェーズ: 有/無) の 2 要因混合計画で行った．

3. 実験結果

合計ポイントについて，一元配置分散分析の結果，選好学習フェーズ有の条件 ($M = 20.6$, $SD = 4.3$) の方が，選好学習フェーズ無の条件 ($M = 18.5$, $SD = 4.3$) よりも有意に高いことがわかった ($F(1, 146) = 9.8, p = .002, \eta_p^2 = .063$ (図 2a))．この結果は，選好学習フェーズが，エージェントと参加者にとって Win-Win な配分をすることに寄与することを示す．エージェントポイントについて，一元配置分散分析の結果，選好学習フェーズ有の条件 ($M = 11.0$, $SD = 5.0$) の方が，選好学習フェーズ無の条件 ($M = 7.5$, $SD = 5.2$) よりも有意に高いことがわかった ($F(1, 145) = 17.3, p = .000, \eta_p^2 = .105$ (図 2b))．この結果は，選好学習フェーズにより，参加者がエージェントに譲歩するようになったことを示している．参加者ポイント

について，一元配置分散分析の結果，選好学習フェーズ有の条件 ($M = 9.6$, $SD = 4.1$) の方が，選好学習フェーズ無の条件 ($M = 10.9$, $SD = 5.1$) に有意な差はなかった ($F(1, 145) = 2.6, p = .107, \eta_p^2 = .018$ (図 2c))．エージェントポイントの結果も踏まえるとこの結果は，エージェントの選好を正しく推定することで，参加者は自身のポイントを下げることなく，エージェントのポイントを上げることを示している．

固定パイ・バイアスについて，繰り返しありの二要因分散分析を行った結果，学習前/ゲーム後水準に主効果はなく，選好学習フェーズの有/無の主効果もなく ($F(1, 145) = 1.64, p = .202, \eta_p^2 = .011$)，交互作用も確認されなかった ($F(1, 145) = 0.922, p = .339, \eta_p^2 = .006$ (図 2d))．このことから，固定パイ・バイアスは選好学習や学習前後に影響されないことがわかった．選好推定について，繰り返しありの二要因分散分析を行った結果．学習前/ゲーム後水準に主効果が認められた ($F(1, 145) = 105.05, p < .001, \eta_p^2 = .420$)．Bonferroni の方法による多重比較を行った結果，参加者が推定したエージェントの選好は，学習前 ($M = .54$, $SD = .51$) と比較して，ゲーム後 ($M = .05$, $SD = .42$) の方が有意に高かった ($p < .001$)．加えて，学習前/ゲーム後と有/無の交互作用が確認された ($F(1, 145) = 105.05, p < .001, \eta_p^2 = .420$)．Bonferroni の方法による多重比較を行った結果，選好学習フェーズ有の参加者は ($M = .79$, $SD = .42$)，無の参加者 ($M = .31$, $SD = .47$) と比較して，ゲーム後にエージェントの選好をより良く推定した ($F(1, 145) = 41.56, p < .001, \eta_p^2 = .223$(図 2e))．このことは，選好学習フェーズにより，参加者がエージェントの選好をより良く推定できるようになったことを示している．

4. 議論

本研究では，選好学習により，交渉において感情表現から相手の心理状態を推測する能力が向上し，Win-Win な解決策を導くことができるかを検討した．その結果，選好学習フェーズ有の参加者は，無の参加者に比べて，相手の選好をより良く推定し，エージェントと参加者の合計ポイントが増加することが示された．これは，選好推定が正確になることで，より Win-Win な配分をすることにつながったことを示している．

Win-Win な配分の内訳について，エージェントのポイントは選好推定により上がったが，参加者のポイントは変わらなかった．協力主義者は個人主義者よりも Win-Win な配分に容易に到達可能である [9]．そのため，現在の自分のポイントにしか興味がない個人主義から協力主義なエージェントにデザインすることで，参加者自身のポイントも上げることができ，より Win-Win な解決策を導き出すことが容易になる可能性がある．

文献

[1] J. Gratch, “The promise and peril of automated negotiators,” *Negotiation Journal*, vol. 37, no. 1, pp. 13–34, jan 2021.

[2] R. Cooper, D. V. DeJong, R. Forsythe, and T. W. Ross, “Communication in coordination games,” *The Quarterly Journal of Economics*, vol. 107, no. 2, pp. 739–771, may 1992.

[3] H. Takagi and K. Terada, “The effect of anime character’s facial expressions and eye blinking on donation behavior,” *Scientific Reports*, vol. 11, pp. 1–8, Apr. 2021.

[4] L. I. Reed, R. Stratton, and J. D. Rambeas, “Face value and cheap talk: How smiles can increase or decrease the credibility of our words,” *Evolutionary Psychology*, vol. 16, no. 4, p. 147470491881440, oct 2018.

[5] M. Sato and K. Terada, “The effect of visualizing other’s appraisal process on the cognitive ability to reach win-win outcomes in a multi-issue ultimatum game,” in *Conference of the International Society for Research on Emotion (ISRE2022)*, 2022.

[6] K. Terada and C. Takeuchi, “Emotional expression in simple line drawings of a robot’s face leads to higher offers in the ultimatum game,” *Frontiers in Psychology*, vol. 8, pp. 1–9, May 2017.

[7] C. M. de Melo, P. J. Carnevale, S. J. Read, and J. Gratch, “Reading people’s minds from emotion expressions in interdependent decision making,” *Journal of Personality and Social Psychology*, vol. 106, no. 1, pp. 73–88, 2014.

[8] E. Johnson and J. Gratch, “The impact of implicit information exchange in human-agent negotiations,” in *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. ACM, oct 2020.

[9] P. J. Carnevale and A. M. Isen, “The influence of positive affect and visual access on the discovery of integrative solutions in bilateral negotiation,” *Organizational Behavior and Human Decision Processes*, vol. 37, no. 1, pp. 1–13, feb 1986.