Press Release

報道の解禁日（日本時間）：2023 年 7 月 22 日（土）7 時 00 分

2023 年7月21日

記者会、記者クラブ 各位

交渉において心を読む能力を向上させる AI エージェントを開発

～多様な価値観を認めあう社会の実現に向けて～

【研究概要】

東海国立大学機構 岐阜大学の佐藤幹晃（博士前期課程2年）、寺田和憲教授らは南カリフォルニア大学の Jonathan Gratch 教授との国際共同研究によって、交渉の成功に重要な心の状態[1]の一つである相手の選好（価値観の相対化によって得られる順序関係）を読む能力を向上させるための AI エージェント[2]システムを開発し、成人を対象としてその有効性を示しました。

多くの人にとって交渉は難易度の高い社会的相互作用です。そのため、例えば、給与交渉が収入を向上させる可能性を持つにもかかわらず、多くの人が給与交渉をせずに提示された給与をそのまま受け入れていることが知られています。また、交渉への消極性が経済的不平等や賃金停滞を助長する一因となっている可能性も指摘されており、交渉能力の向上は社会的課題であると言えます。一般的に、交渉では Win-Win の関係になれる可能性がありますが、そのためには、複数ある論点に対する選好を相互に正確に見極め、資源の分配を最適化する必要があります。選好は直接観察できない心の状態であるため知ることは容易ではありません。人は日常的に、直接観察可能な相手の行動や表情から相手の選好を推論していますが、交渉においては、相手の選好が自分と一致しているという固定パイバイアスなどの様々な認知バイアス[3]が選好の推論を困難にしており、認知バイアスの克服が課題となっていました。

本研究グループは、人の感情が個人の選好に基づいた状況評価の結果により表出されるという評価理論（appraisal theory）[4]に基づいて設計した感情評価生成モデルを AI エージェントに搭載しました。評価プロセスを視覚的に明示し、ユーザに言語的フィードバックを与えながらシンプルな交渉タスクを行うことで、選好と表出感情の対応関係のモデルを用いて選好を推論する方法を学び、選好を読む能力を向上させるシステムを開発しました。また、成人 187 人を対象とした実験の結果、提案方法によって、選好を読む能力が向上することを確認しました。

本研究グループは、高ウェルビーイング社会の実現のためには、社会構成員が相互に相手の心を読むことによって他者の多様な価値を認め、資源やタスクの分配を数理的に最適化する能力（数理的社会情動能力）を持つことが重要だと考えています。今後は、本研究の成果を発展させ、道徳と算数をハイブリッドした教育プログラムを AI エージェントとのインタラクションに実装することで、コミュニケーションに課題を抱える子どもたちも含めた、社会の未来を担う多くの子どもたちが、相手の心情を理解し、多様な価値観を尊重し、複雑な人関関係にうまく対応する能力を獲得できるシステムを開発する予定です。

本研究成果は、日本時間2023年7月22 日（土）に科学誌「IEEE Transactions on Affective Computing」のオンライン版で公開されます。

【研究の背景と経緯】

交渉は、日常生活のあらゆる場面に存在します。国家間の対立から、家族の夏休みの過ごし方をめぐる議論に至るまで、争い事は、可能な解決策を検討し議論すること、つまり交渉により解決できる可能性が高まります。交渉には複数の論点があり、論点に対する価値観や選好は人によって異なります。したがって、交渉において Win-Win の合意に至るためには、怒りにまかせて対立するのではなく、自他の選好の一致・不一致を正確に見極めて、資源の分配を最適化する必要があります。交渉者は、会話から相手の好みを聞き出すなど明示的な情報交換を通じて相手の選好を知ることもできますが、非言語的な手がかり、特に感情表現を観察することによっても知ることができます。

人の Win-Win の合意を阻害する認知バイアスとして、固定パイバイアス、アンカリング、過信などが知られています。認知バイアスの克服のためには交渉中に交換される情報から正しく相手の選好や限界をデコード[5]することが鍵となります。なお、交渉における重要な数学的なパラメータである選好と限界は主観的に決まるという問題がありますが、心理学の方法によって数値化することで、ある程度客観性を保って扱うことが可能になります。感情は人の状況に対する評価（appraisal）をエンコード[5]しているために、逆評価（reverse appraisal）[6]を用いて相手の選好をデコードすることができます。逆評価は合理的エージェントの生成モデルを尤度（ゆうど）として用いたベイズ推論[7]によって実現できます。

交渉術を教えられる機会はほとんどありませんが、米国科学アカデミーから世界経済フォーラムに至るまで、交渉術のトレーニングを進めるためのイノベーションを呼びかけています。AI エージェントを用いた交渉トレーニングシステムはこれまでに提案されていますが、フィードバックの与え方などが重視されており、人の感情の認知計算過程は考慮されていませんでした。

【研究の内容】

本研究グループは、評価理論（appraisal theory）に基づいて設計した感情評価生成モデルを AI エージェントに搭載し、評価プロセスを視覚的に明示し、言語的フィードバックを与えることで、人に生成モデルの理解と学習を促し、逆評価の方法をトレーニングするシステムを開発しました。交渉において、相手の心（選好）を読む能力を向上させられるかどうか、また Win-Win 合意に至る能力を向上させられるかどうかを検証しました。提案するシステムは、選好を読む能力を向上させることを目的とした世界初のトレーニングシステムです。評価理論に基づく人の感情の認知過程の計算モデルを搭載した AI エージェントを導入することで開発することができました。

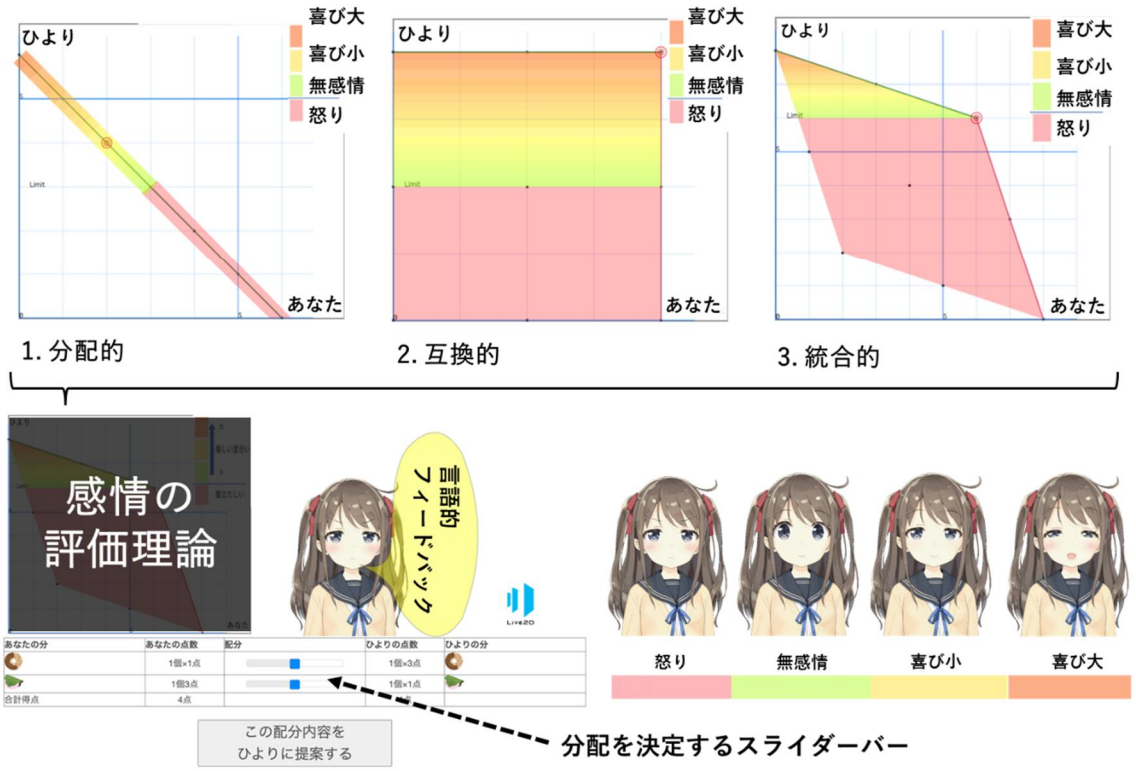


図1　感情の逆評価による選好の推論をトレーニングするインタフェース。キャラクターひよりは株式会社 Live2D の著作物です。このタスクでは 2 種類のスイーツ(この例ではドーナツと柏餅)それぞれ 2 個を実験参加者とひよりで分配します。実験参加者は、ひよりと実験参加者のスイーツの好みが、分配的な交渉の場合一致(利害が対立)、互換的な交渉の場合は正反対(完全に Win-Win になれる)、統合的交渉では部分的に一致しているという設定で交渉を経験します。

本研究グループが提案する感情の逆評価による選好の推論をトレーニングするインタフェースを図1に示します。ユーザはこのインタフェースを使って、複数種類のスイーツを分配する交渉タスク(複数論点最後通牒ゲーム[8])を AI エージェント(ひより:キャラクターひよりは株式会社 Live2D の著作物です)と行います。ユーザは分配を決定し、提案します。提案は一度だけ行うことができます。AI エージェントは分配が気に入らなかったら拒否します。相手に拒否されると、ユーザも相手もスイーツは一切得られません。したがって、このタスクは AI エージェントの拒否の限界を見極めつつ自分も満足する配分を探索するタスクと言えます。

交渉はシンプルな 2 論点(スイーツ 2 種類)で、分配的交渉、互換的交渉の 2 つの極端な交渉と、より一般的な統合的交渉の合計 3 つの交渉[9]を経験します。ユーザがスライダーバーを操作すると、2 次元グラフ上で、現在の分配状況を示す点が移動し、自他の効用(総合利益)を逐次に確認することができます。2 次元グラフには AI エージェントの効用と感情表出の対応を可視化した感情評価過程を表すヒートマップが重ねられています。また、AI エージェントはヒートマップの数値に応じた表情を表出します。表情はスイーツが好きな度合いと、自分に割り当てられるスイーツの量のかけ算によって計算される効用を使って表します。交渉を成立させるためには、このかけ算の結果が AI の許容度を超えない範囲になるようにしなければいけません。具体的には表情は効用$u = \boldsymbol{w} \cdot \boldsymbol{x}$使って式(1)に従って表出されます。$\boldsymbol{w}$はスイーツの種類に対する重み、すなわちそのスイーツがどれぐらい好きかを表すベクトル、$\boldsymbol{x}$はその時点で自分に割り当てられているスイーツの個数を表すベクトルです。$limit$はそれ以下の数値であれば AI

が提案を拒否する限界です。

$$expression = \begin{cases} anger & \text{if } u < limit, \\ neutral & \text{if } u = limit, \\ joy & \text{if } u > limit. \end{cases} \quad (1)$$

さらに、AI エージェントはテキストによって分配状況の意味を説明するフィードバックを与えます。これらを経験することで、ユーザは分配、効用、表出感情の対応関係である評価過程を学ぶことができます。この評価過程の学習によって、観測した表情だけから逆に選好を推論することができるようになると考えられます。

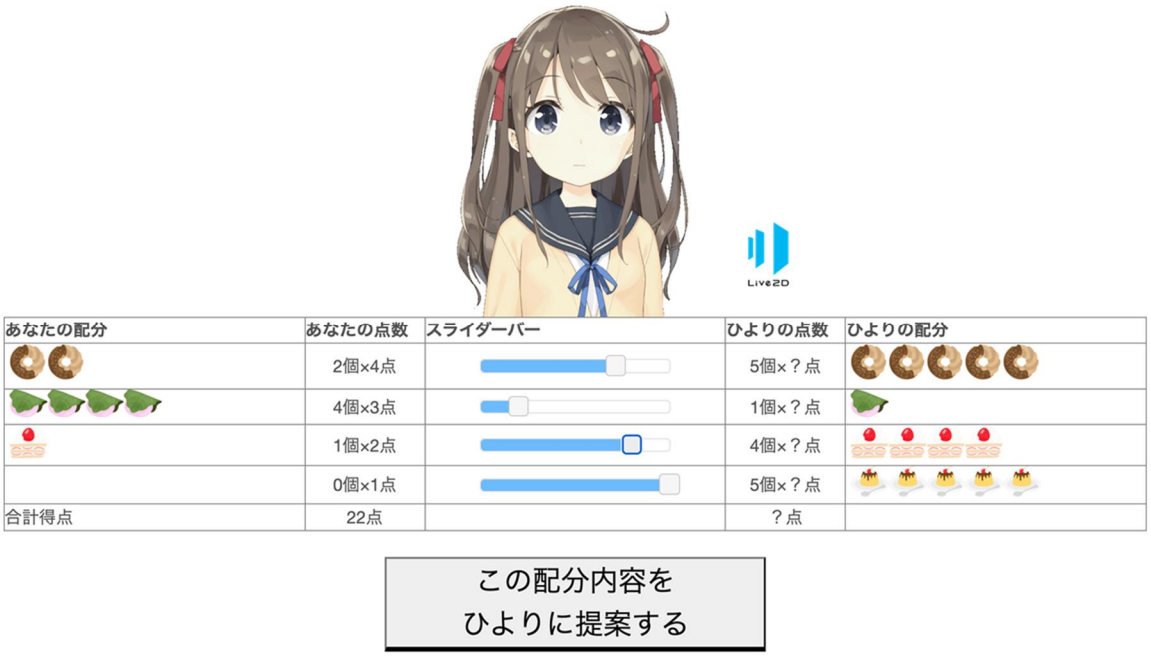
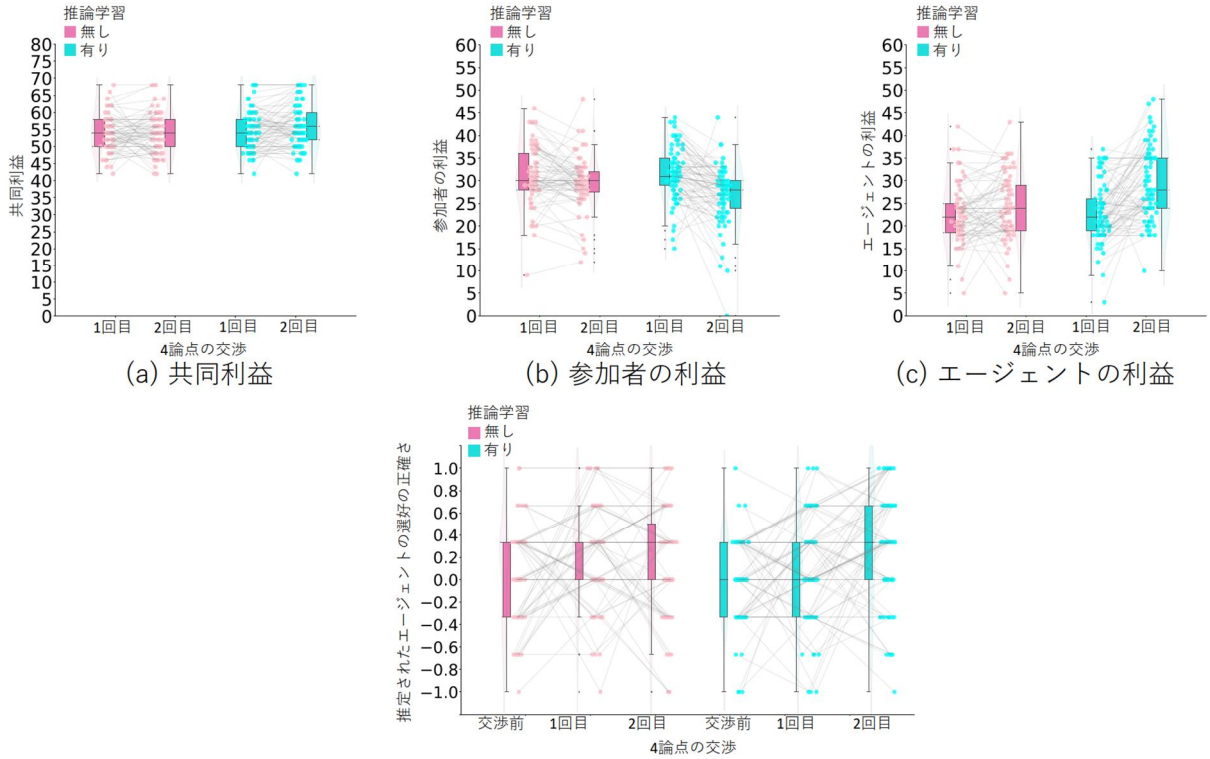


図2　交渉パフォーマンス、選好推論性能を測定するために用いたインタフェース

このトレーニングインタフェースの有効性を検証するためのオンライン実験を 187 人（平均年齢 46.9 歳、男性70%）の参加者を対象に実施し、すべての参加者に 4 論点（スイーツ4種類）の交渉（複数論点最後通牒ゲーム）を 2 回行ってもらいました（用いたインタフェースは図 2 参照）。実験では、参加者をトレーニング経験群（2 回の交渉の間にトレーニングインタフェースを使った選好推論トレーニングを行うグループ）と、トレーニングを行わない非経験群に分けました。そして、それぞれ 1 回目の 4 論点交渉に比べて 2 回目の 4 論点交渉のパフォーマンスと選好推論の正確さが向上しているかどうかを測定しました。

分析の結果、トレーニング経験群にのみ選好推定性能の向上が見られ、トレーニングインタフェースが選好推定能力の向上に有効であることが確認されました（図 3d）。このことは、実験参加者がトレーニングインタフェースを使うことで、交渉中に交換される非言語情報から正しく相手の選好をデコードする能力（相手の心情を読み取る能力）を向上させたことを示唆します。交渉のパフォーマンスについては、共同利益を減ずることなく、自分の利益を低く、相手の利益を高くするようになることが確認されました（図 3a、3b、3c）。



(a) 共同利益　(b) 参加者の利益　(c) エージェントの利益

(d) 参加者の選好推定性能

図3　実験結果。(a)、(b)、(c)は交渉パフォーマンス。(d)は選好推論性能。

【今後の展開】

本研究グループの提案する AI エージェントとのインタラクションで交渉スキルが向上すれば、共同利益の向上により、人々のウェルビーイングが向上する可能性があります。それだけでなく、相手の選好や価値観の読み取り能力の向上によって、単一の価値観にとらわれることなく、個人の選好に従った資源の分配や適材適所の人やタスクの配置が可能になり、ウェルビーイングの高い社会が実現されることが期待されます。

また、他者の選好の正確な理解は対立の解決に寄与します。双方で重視する論点が異なる可能性があるからです。国家間の対立から、家族の夏休みの過ごし方の議論に至るまで、選好の正確な推論に基づき Win-Win の解を見出す能力は対立を解消し、人々のウェルビーイングを高めることが期待されます。

今後は本研究の成果を応用し、道徳と算数をハイブリッドした教育プログラムを AI エージェントとのインタラクションの中に実装することで、コミュニケーションに課題を抱える子どもたちも含めた、社会の未来を担う多くの子どもたちが、相手の心情を理解し、多様な価値観を尊重し、複雑な人関関係にうまく対応する能力を獲得できるシステムを開発する予定です。

【論文情報】

雑誌名：IEEE Transactions on Affective Computing
論文タイトル：Teaching Reverse Appraisal to Improve Negotiation Skills
著者：Motoaki Sato, Kazunori Terada, Jonathan Gratch
DOI: 10.1109/TAFFC.2023.3285931

【用語解説】

[1]心の状態：

心の状態として目的、価値観、選好、信念があり、それらは状態という概念を使って情報学的に定義されます。目的は世界の状態における特定の領域もしくは一点、価値観は世界の状態に割り当てられた評価値（効用値）、選好は価値観の相対化によって得られる順序関係、信念は世界の状態について知っていることです。

[2]AI エージェント：

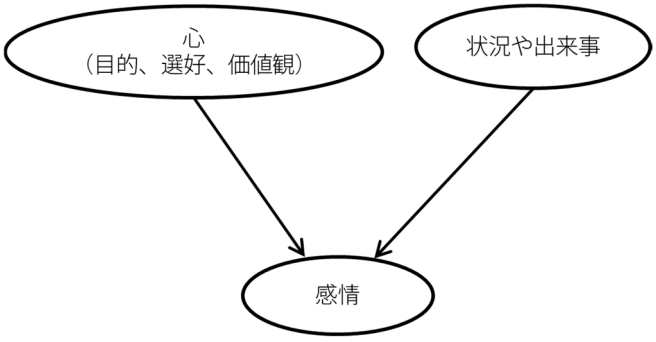
AI エージェントとは、一般的に学習能力を持ち特定の目的を達成するために適応的かつ自律的に行動するソフトウエアまたはハードウェアシステムのことを指しますが、本研究における AI エージェントは、人に交渉を教える目的を達成するために、適応的に動作する、人に類似した外観を有するソフトウエアシステムのことを指します。

[3]認知バイアス（固定パイバイアス、アンカリング、過信）：

認知バイアスとは一般的に、人の意思決定や行動に影響を与える、経験則の重視や直観、先入観によって発生する認知的な誤りや偏りのことを指します。固定パイバイアス以外の認知バイアスは交渉に特有のものではありません。固定パイバイアスは、交渉特有の認知バイアスで、交渉の利益が固定されている、すなわち、一方の利益が他方の不利益であるというゼロサムの視点をとるバイアスです。アンカリングは、一般的に事前に与えられた特定の情報に引きずられた評価が行われることを指し、交渉においては、最初に提示された価格や条件がその後の議論や判断に影響を及ぼすことを指します。過信は一般的に、自分の能力、知識、または判断を過大評価する傾向を指し、交渉においては、自分の立場の強さや自分の提案が相手に受け入れられる可能性を過大評価することを指します。

[4]評価理論（appraisal theory）：

評価理論は、人が体験する感情が、個々の状況や出来事をどのように解釈し、それが自分の目的や価値観にどのように関わるかという「評価」に基づいていると考える理論です。人が状況や出来事をどのように評価するかは、目的や価値観に依存するために、目的や価値観と状況や出来事が与えられたときに感情を出力する関数として表現することができます。これは、心と状況が与えられた場合に感情が発生する条件付き確率$p(\textit{感情}|\textit{心}、\textit{状況})$で表現されます。図示すると以下のようになります。たとえば、同じ出来事でも、それが個人の目的にとって有益だと評価すれば喜びを感じ、逆に損害だと評価すれば怒りや悲しみを感じます。なお、喜怒哀楽等の感情そのものも推定するべき心の状態として考えることも可能ですが、感情を機能としてとらえる立場では、それらの感情状態ではなく、目的、価値観、選考、信念を推定するべき心の状態として考えます。



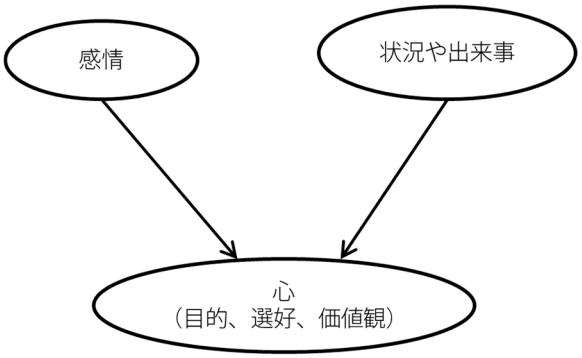
p(感情|心,状況)

[5]デコード、エンコード：

エンコードとデコードは、情報伝達の基本的なプロセスを指します。これらの概念は、情報通信だけでなく、人同士のコミュニケーションを説明するためにも用いられます。エンコードは情報を効率的に送信、保存するために、情報を特定の符号（形式）に変換するプロセスです。デコードはエンコードされた情報をもとの形式や理解可能な形式に戻すプロセスを指します。人同士のコミュニケーションにおいて、エンコードは、思考や感情を言葉や非言語的な表現（身振りや表情）に変換（符号化）することを指します。デコードは符号を受け取った人が言葉や非言語的の表現を思考や感情に変換して理解するプロセスを指します。

[6]逆評価（reverse appraisal）：

感情の逆評価は感情の評価過程の確率的生成モデル（どのような心的状態がどのような状況でどのような感情を生成するかについての知識）を用いて、観測された状態と感情から目的や価値観を逆に推論することです。下図のように、評価過程の確率的生成モデルを尤度関数として用いたベイズ推論としてモデル化されます。



$$p(\text{心}|, \text{感情}, \text{状況}) \propto p(\text{感情}|\text{心}, \text{状況})p(\text{心})$$

[7]尤度、ベイズ推論：

ベイズ推論とは、確率統計における一つの推論手法で、ベイズの定理に基づいています。ベイズ推論は事前確率、尤度、事後確率の概念によって構成されます。事前確率は新たなデータが得られる前の、事象が起きる確率です。尤度はその事象が起きたと仮定した場合に、得られたデータが観測される確率です。事後確率は、新たなデータが得られた後に、その事象が起きる確率です。本研究における事象は心的な状態（選好）、データは表情です。一般的に尤度は事象と観測の対応関係を統計的に学習することで得られますが、本研究では、感情の評価（appraisal）理論にもとづいて、理論的に導出した心的な状態と表情の関係（合理的エージェントの生成モデル）を尤度として用いています。

[8]複数論点最後通牒ゲーム：

最後通牒ゲームは、2 人のプレイヤーが一定の報酬を分割する経済ゲームです。プレイヤー1 は最初に報酬を分ける方法を提案し、プレイヤー2 はそれを受け入れるか拒否します。もし受け入れれば、その通りに分配されます。しかし、拒否した場合、両者とも何も得られません。通常の最後通牒ゲームは論点が一つです。単一論点の最後通牒ゲームは分配的ゲームであるためにゼロサムゲーム的ですが、複数論点に拡張することで、互換的、統合的なゲームとなり得るために、協力的要素が強くなります。

[9]3つの交渉（分配的交渉、互換的交渉、統合的交渉）：

分配的（distributive）交渉：双方の選好が完全に一致しているときに分配的になります。固定され

た資源(パイ)を双方で争うことになります。一方の利益が他方の損失になります。
互換的(compatible)交渉:双方の選好が正反対の場合に互換的になります。双方ともに自分の利益になる論点を取得し、不利益な論点を放棄することで、利害が全く対立しない分配が可能です。
統合的(integrative)交渉:分配的交渉と互換的交渉の中間に位置する交渉です。どの論点が対立していてどの論点が互換的かを適切に見分けることによって、双方ともに満足度の高いWin-Winの分配が可能です。

【研究サポート】

本成果は、以下の事業・研究領域・研究課題によって得られました。
科学技術振興機構(JST) 未来社会創造事業 探索加速型
研究領域:「個人に最適化された社会の実現」
(運営総括:和賀 巌 NEC ソリューションイノベータ株式会社 シニアフェロー)
重点公募テーマ:「他者とのインタラクションを支えるサービスの創出」
研究開発課題名:「数理的社会情動能力の発達を促進する AI エージェントシステムの開発」
(JPMJMI22J3)
研究開発代表者:寺田 和憲 岐阜大学 教授
研究開発期間:令和4年10月～令和7年3月

JST はこの重点公募テーマで、多様な個人が他者と相互につながり合うインタラクションを介して、より良い社会関係性を構築するような社会的適応を支援する技術(製品やサービス)や仕組みを開発することを目標とします。そのために、これまでの発達心理学や認知科学と、脳科学・生理学・情報工学等の統合により、社会情動的能力の成長と発達を促し、これにより、誰もがウェルビーイングを実感できる社会の実現を目指します。上記研究課題では、ゲーム理論、進化心理学、社会心理学、認知科学の知見に基づいて構成した、道徳と算数をハイブリッドした教育プログラムを AI エージェントとのインタラクションで実装することで、社会の未来を担う、発達障害児を含む子どもたちが学校教育の中で「数理的社会情動能力」を獲得できるシステムを開発しています。社会情動能力は、IQによって計測される「認知能力」と対比し「非」認知能力とされるが、対人関係の軋轢(搾取、所得格差、いじめ、パワーハラスメントなど)は、自他の価値を明示的に相対化し、数理最適化することで認知的に解くことが可能であり、我々はその能力を「数理的社会情動能力」と呼びます。数理的社会情動能力のコアは、A)見えない状態である「相手の心」や「相手との関係」を推論し(心の理論)、B)関係を数理最適化する能力です。

科学技術振興機構(JST) 戦略的創造研究推進事業 CREST
研究領域:「信頼される AI システムを支える基盤技術」
(研究総括:相澤 彰子 情報・システム研究機構 国立情報学研究所 教授)
課題名:「納得感のある人間-AI 協調意思決定を目指す信頼インタラクションデザインの基盤構築と社会浸透」(JPMJCR21D4)
研究代表者:山田 誠二 情報・システム研究機構 国立情報学研究所 教授
主たる共同研究者:寺田 和憲 岐阜大学 教授
研究開発期間:令和3年10月～令和9年3月

【問い合わせ先】
＜研究に関すること＞
東海国立大学機構 岐阜大学工学部電気電子・情報工学科　教授 寺田 和憲（テラダ カズノリ）
電話：058-293-2792
E-mail：terada2023@ai.info.gifu-u.ac.jp

＜JSTの事業に関すること＞
科学技術振興機構 未来創造研究開発推進部 内田 信裕（ウチダ ノブヒロ）
電話：03-6272-4004
E-mail：kaikaku_mirai@jst.go.jp

＜報道に関すること＞
東海国立大学機構 岐阜大学　総務部広報課広報グループ
電話：058-293-3377
E-mail：kohositu@t.gifu-u.ac.jp

科学技術振興機構 広報課
電話：03-5214-8404
E-mail：jstkoho@jst.go.jp