

人間とエージェントの繰り返し交渉における Misrepresentation の効果

吉岡 幸輝†

藤田 桂英‡

†東京農工大学 工学部 情報工学科

‡東京農工大学大学院 工学研究院 先端情報科学部門

1 はじめに

マルチエージェントシステムに関する研究において、自律エージェント同士で協調した動作を可能とする技術として、自動交渉が注目されている [1]. その中でも、最近では人間と交渉を行うエージェントへの関心が高まっている。人間とエージェントの交渉では情報の交換が重要である。情報の交換を通して、交渉者は交渉タスクへの理解を深め、効率的な解決案を模索する。しかし、交渉者は自身の利益を高めるために、交換する情報に嘘 (misrepresentation) を用いる可能性がある。既存の研究として、統一的な交渉における “fixed-pie lie” に対して嘘が有効であることが示されている [2, 3]. また、人間の交渉者が交渉中に嘘に気づくことは困難であると述べられている。これらの研究では、単一の交渉における嘘の有効性は示されているが、繰り返し交渉における有効性については検証が不十分である。

本研究では、単一の交渉で有効であった “fixed-pie lie” を用いた戦略の他に、より度合いの小さい嘘を用いた新たな戦略を提案する。さらに、正直な戦略と比較を行うことで、それぞれの嘘の有効性を、人間とエージェントの獲得効用から評価し、繰り返し交渉における嘘の有効性やどのような嘘が有効かについて示す。また、嘘を用いる戦略と、正直な戦略を繰り返し交渉中に使い分けることで、連続して同じ戦略を使用した場合と比較し、結果に影響を及ぼすか調査する。繰り返し交渉は同じ交渉相手と複数回交渉を連続して行うため、単一の交渉よりも交渉相手に嘘をついていると悟られてしまう可能性が高いと考えられる。そこで、繰り返し交渉では人間の交渉者が嘘に気づくことが可能であるか、どのような戦略を用いた場合に人間は交渉相手が嘘をついたと判断するか、人間の交渉者へのアンケートを行い検証する。

2 関連研究

2.1 IAGO

人間とエージェントの交渉プラットフォームとして、IAGO [4, 5] が提供されている。IAGO では、交渉に用いるドメインや人間とエージェントに与えられる選好を決定することができる。また、Best Alternative To Negotiated Agreement (BATNA) と呼ばれる、時間内に交渉が合意に達しなかった場合に人間とエージェントが受け取る効用の値が設定されている。

人間の交渉者は、図 1 のような画面で交渉を行う。交渉画面の右上には、交渉で送受信されるメッセージがチャット形式で表示される。交渉画面の右下では、交渉相手であるエージェントに対して送るメッセージをあらかじめ設定された選択肢の中から決定することができる。送信できるメッセージには様々な種類があり、相手の選好に関する質問や自分の選好の送信などがある。交渉画面の左下には、エージェントに対して送る提案を決定するテーブルが表示されている。決定した提案を送信すると、右上のチャットボックス内に決定した提案が文章として表示される。また、画面左上にはエージェントが示している表情を表す画像や、交渉の制限時間などが表示される。

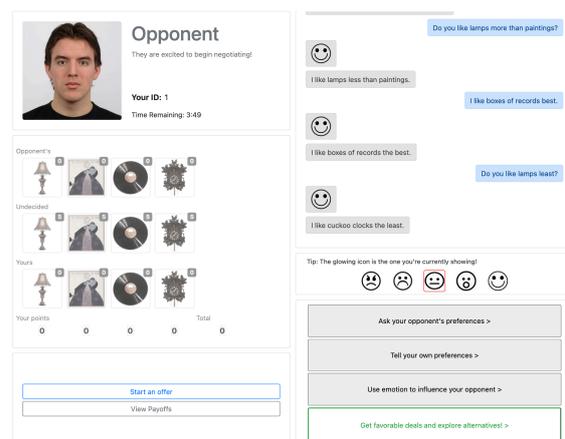


図 1: 交渉プラットフォーム IAGO

IAGO では、基本となるエージェントが提供されている。このエージェントは積極的に提案やメッセージを送信することはなく、人間のプレイヤーの行動

Effect of Misrepresentation over Human-Agent Repeated Negotiations

†Department of Computer and Information Sciences, Faculty of Engineering, Tokyo University of Agriculture and Technology

‡Division of Advanced Information Technology and Computer Science, Institute of Engineering, Tokyo University of Agriculture and Technology

に応じて反応を返すように設計されている。本研究では、作成するエージェントの基盤として、このエージェントを用いる。これを本論文では基本エージェントと呼ぶ。基本エージェントは、相手から選好に関する質問が送られた場合、正直に答えるように設計されている。また、提案が送られた場合にはその提案で自分が獲得する効用とあらかじめ設定されたマージンの和と、相手が獲得する効用を比較し、自分の効用が相手の効用よりも大きかった場合に提案を受け入れる。このマージンの値は交渉開始時は0に設定されており、エージェントが提案を拒否するたびに1ずつ大きくなる。これにより、エージェントは提案を拒否するたびに次の提案の受け入れ条件が緩和されることになる。

このIAGOを用いた研究として、人間と交渉するエージェントの有効性を調べる研究 [6] や交渉の分析に機械学習を用いた研究 [7] などがある。

2.2 The Misrepresentation Game

先行研究である The Misrepresentation Game[2] では、Multi-issue bargaining task を交渉タスクとして扱っている。Multi-issue bargaining task では、各論点はいくつかのレベルで構成されており、交渉者はそれぞれの論点についてどのようにレベルを分け合うかを交渉し、決定する。各交渉者は交渉が合意に達した場合、報酬を受け取る。例えば、図 2.2 のような交渉では論点はリンゴとオレンジの 2 種類であり、レベルはどちらも 3 となる。

この研究では、人間の交渉者の特徴として、いくつかの仮定を行なっている。以下にその仮定を示す。

- 公平さ: 人間は交渉者間の報酬の差を最小化するように努力すると仮定する。
- 効率性: パレート効率的な取引を好むと仮定している。
- 情報の交換: 一方的に相手に情報を提示することは自分が交渉で不利な立場に陥ることにつながるため、互恵的な情報交換を行う。
- fixed-pie バイアス: 人間はよく相手の利益が自分の利益と反対であると仮定する。

これらの仮定に基づき、人間の交渉人は公平で効率的な解決策を交渉を通して模索する。しかし、このような交渉では、選好を偽ることでより多くの利益を獲得できる可能性がある。例えば、図 2 の上のような交渉では、エージェントは正直に選好を述べて

いるため、公平な交渉となっている。一方で下のような交渉では、エージェントは嘘の選好を述べることで、人間に公平に見せかけながらより多くの利益を獲得している。このような仮定のもとで、交渉相手が公平で効率的だと考え、かつ自分の利益が最大となるような嘘の選好情報を分析した結果、自分の選好を相手と同じ選好と偽る、すなわち分配的な選好であると装うことで、最大の利益を得ることができると示された。このような嘘を“fixed-pie lie”と呼ぶ。また、嘘の選好を用いた場合の方が正直に選好を伝えた場合よりも受け入れられやすく、より公平だと評価されることが示された。

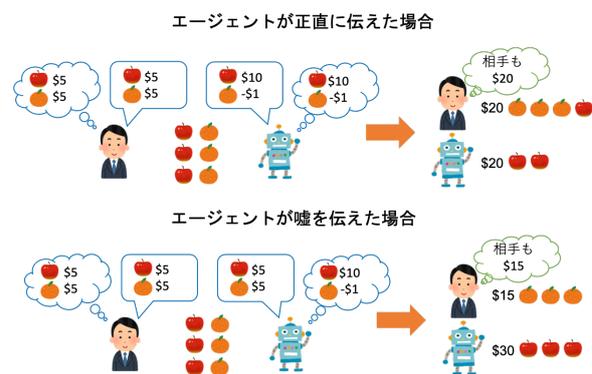


図 2: 嘘を用いた交渉の例

上記の研究ではいくつかの仮定に基づいて検証が行われていた。この結果から以下の 3 つの仮説について、既存の大規模な交渉コーパスを用いて、限定的な仮定がない場合での検証が行われた [3]。

1. 交渉者が分配的な選好を伝えることで、利益が増加する
2. 交渉が統合的で、相手に伝える嘘が分配的な選好ならば、利益が増加する
3. 嘘をつく交渉者は嘘の選好と自分が送る提案との間に矛盾が生じる

さらに、3 の仮説が正しい場合、生じた矛盾に気づくことで嘘をついていると疑う可能性がある。そのため、嘘は正直さの評価に影響を与えるかどうか、研究課題として設定された。

まず、1 の仮説についての検証では、伝えられた選好が分配的な選好に近いほど、より多くの報酬を獲得したことが示された。また、この結果に関して交渉の構造は関連性がなかった。この結果から、1 の仮説が支持された。2 の仮説についての検証では、嘘をついた交渉者と正直な交渉者の獲得した報酬を t 検定

によって比較した結果、統合的な交渉では嘘つきの方が有意に多くの報酬を獲得し、分配的な交渉では有意な差がないことが示された。また、統合的な交渉における嘘の88%が分配的な選好を装う嘘であった。さらに3の仮説について、伝えた選好に基づく順位推定と、提案に基づく順位推定の距離の差について、嘘つきと正直者で比較した結果、有意な差があることが示された。最後に、この結果から3の仮説が正しいと仮定し、研究課題について検証が行われた。交渉相手に対する正直さについて7段階の評価が行われており、その結果から統合的な交渉では嘘つきと正直者に有意な差が生じず、分配的な交渉では嘘つきの方が正直者よりも有意に正直さが低いと評価されたことが示された。この結果から、人間の交渉者が分配的な選好を装う嘘を見破ることは困難であると見ることができる。

3 提案手法

3.1 エージェントの用いる偽の選好情報の決定

本研究では、エージェントの用いる嘘の選好情報の戦略として2種類提示する。第一の戦略として、分配的な選好を装う嘘を用いる。既存の研究 [2, 3] では、人間の交渉者に有効な嘘として、単一の統合的な交渉において分配的な選好を装う嘘が有効であると示されている。これらの研究で有効と示された分配的な選好を装う嘘が、繰り返し交渉においても有効であるか、獲得した効用について正直なエージェントと比較し、検証する。以降この第一の戦略を戦略Mと呼ぶ。ある交渉において論点数が n で、任意の k 番目の論点の人間の交渉者側の選好が p_k 番目に好ましいものであるとき、戦略Mを用いた場合にエージェントが人間の交渉者に伝える選好の順序 o_k を式(1)のように定義する。ただし、 $1 \leq k \leq n$ を満たし、複数の論点と同じ選好をもつことはない。

$$o_k = p_k \quad (1)$$

また、第二の戦略として、相手の交渉者が最も好ましいアイテムに焦点を当て、自分も同様に最も好ましいと主張し、他のアイテムについては選好の順序を変更しないような嘘を用いる。同じ交渉相手と連続して交渉を行う繰り返し交渉では、単一の交渉に比べて交渉相手に嘘をついていることが見破られる可能性が高いと考えられる。これを考慮し、第二の戦略は嘘を一部分のみに限定し正直な情報と合わせることで、相手に嘘をついていることを見破られにくくし、全ての選好を正直に伝える場合よりも多くの

効用を獲得することを目的として提案する。以降この第二の戦略を戦略mと呼ぶ。ある交渉において論点数が n で、任意の k 番目の論点の人間の交渉者側の選好が p_k 番目に好ましいもので、エージェント側の選好が a_k 番目に好ましいものであるとする。ここで、 l 番目の論点の人間の交渉者にとって最も好ましいものであるとき、戦略mを用いた場合にエージェントが人間の交渉者に伝える選好の順序 o_k を式(2)のように定義する。ただし、 $1 \leq k \leq n$ かつ $1 \leq l \leq n$ を満たし、複数の論点と同じ選好をもつことはない。

$$\begin{cases} o_l = p_l & (p_l < p_k) \\ o_k = a_k & (a_l < a_k) \\ o_k = a_k + 1 & (a_k < a_l) \end{cases} \quad (2)$$

さらに、選好を正直に伝える戦略を戦略Hと呼ぶこととする。ある交渉において論点数が n で、任意の k 番目の論点のエージェント側の選好が a_k 番目に好ましいものであるとき、戦略Hを用いた場合にエージェントが人間の交渉者に伝える選好の順序 o_k を式(3)のように定義する。ただし、 $1 \leq k \leq n$ を満たし、複数の論点と同じ選好をもつことはない。

$$o_k = a_k \quad (3)$$

3.2 偽の選好に応じた感情の表出

人間の交渉者に嘘の選好情報を伝達していることを見破られないための方法の一つとして、相手の提案に対して偽の選好に応じた感情を表出することを考える。基本エージェントは、交渉相手の提案を受け入れる際に喜びを示す感情(happy)を表出し、断る際には悲しみを示す感情(sad)を表出するため、提案を断る際にその提案がどの程度受け入れ難いのか、人間の交渉者に伝わりにくい。したがって、人間の交渉者の提案で自分が得られる報酬と人間の交渉者が得られる報酬を比較し、その割合に応じて感情を変化させることで、より人間の交渉者に自分の伝えたい選好情報が伝わりやすくなると考えられる。さらに、嘘を用いるエージェントの場合にはこの感情の表出を嘘の選好情報をもとに行うことで、より相手に嘘の選好情報を印象付ける。

表出する感情として、3段階の感情を使い分ける。基本エージェントは基本的に自分の獲得効用が相手の獲得効用以下の場合に提案を断る。すなわち、相手の獲得効用が社会的余剰の5割以上のとき、提案を拒否する。この割合が大きければ大きいほど、エージェントにとって不利な提案であるため、より厳しい感情を表出とする。まず、相手の獲得効用が社会的余剰の5割以上で7割に満たない場合、悲し

みを示す感情 (sad) を表出する。次に、7割以上で9割に満たない場合、怒りを示す感情 (angry) を表出する。さらに9割以上の場合には、うんざりした感情 (disgusted) を表出する。このように感情を使い分けることで、人間の交渉者からの提案に対してエージェントがどのように評価したかを明確にする。嘘の戦略を用いる際には、提案の受け入れは本来の選好情報をもとに判断を行い、感情を表出する際に用いる自分の獲得効用の計算に嘘の選好情報を用いれば良い。

4 実験

4.1 実験設定

実験には、2種類のドメインを用いる。一方のドメインは4種類のアイテム各5個の合計20個について分配する交渉とする。各アイテムには種類ごとに異なる1から4の効用値を設定し、交渉は全て統合的な交渉とする。この交渉では、効率的な交渉の解はエージェントと人間ともに0.7となる。このような解を kalai-smorodinsky 解 [8] と呼ぶ。もう一方のドメインでは、5種類のアイテム各5個の合計25個について分配する交渉を行う。各アイテムには種類ごとに異なる1から5の効用値を設定し、こちらも全て統合的な交渉とする。この交渉では、効率的な交渉の解は(0.68, 0.72)の組み合わせとなる。交渉は3回で1つの繰り返し交渉とし、これを1ラウンドとする。この2種類のドメイン両方について5種類のエージェントで交渉を行うため、合計10種類の繰り返し交渉について実験を行うこととなる。被験者は全ての種類の交渉を1回ずつの合計10ラウンドの交渉を行い、実験を繰り返すことによる操作の習熟度によって差を生み出さないようにするため、被験者ごとに実験を行う順番は異なるようにする。

本研究の実験では以下の5種類のエージェントで交渉を行う。各エージェントの名称は、3回の繰り返し交渉のうち、各回の交渉で用いる戦略を示している。例えば、エージェントMHMは1回目の交渉で戦略M、2回目の交渉で戦略H、3回目の交渉で戦略Mを用いる。

1. エージェントMMM
2. エージェントmmm
3. エージェントHHH
4. エージェントMHM
5. エージェントMHH

本実験では、エージェントと人間のそれぞれについて正規化された個人効用値の平均の他に、アンケートを行い、比較に用いる。アンケートは、各交渉後に7段階でエージェントの好感度を尋ねるアンケートと、各ラウンド終了後に交渉相手が嘘をついているように感じたかについて5段階で尋ねるアンケートの2種類を行う。さらに、全ての交渉終了後には、どのような場面でエージェントが嘘をついていると感じたかについて質問を行う。

4.2 実験結果

まず、各エージェントごとに人間とエージェントそれぞれの正規化された個人効用値は図3のようになった。

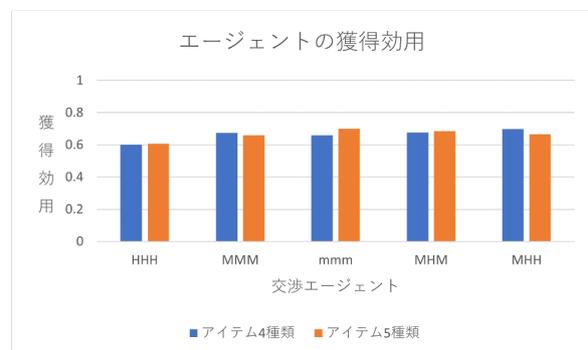


図3: 正規化された個人効用値

アイテム4種類の場合、人間とエージェントともにエージェントHHHを用いた場合の効用が最小であった。嘘を用いたどのエージェントも、正直な戦略のみを用いたエージェントHHHよりも高い平均獲得効用を示したが、エージェントHHHとそのほか4種類のエージェントの間で有意水準5%のt検定を行ったところ、有意差は生じなかった。また、人間とエージェントの個人効用の和である社会的余剰についても、嘘を用いたエージェントの平均の方が高い値を示したが、有意差はなかった。このことから論点数を4としたとき、嘘をつくことでより多くの効用を獲得できるとは言えない。一方、この交渉の効率的な解である(0.7, 0.7)から最も離れた平均値を示したエージェントがエージェントHHHであったことから、正直に選好を伝えても必ずしも効率的な解を生み出せるとは限らないと推測できる。

さらにアイテム5種類の場合でも、アイテム4種類の時と同じく、人間とエージェントともにエージェントHHHを用いた場合の効用が最小であった。エージェントHHHとそのほか4種類のエージェントの間で有意水準5%のt検定を行ったところ、有意差は

生じなかった。また、社会的余剰についても嘘を用いたエージェントの平均の方が高い値を示したが、有意差はなかった。このことから論点数を5としたときについても、論点数4の時と同様に嘘をつくことでより多くの効用を獲得できるとは言えない。また、この交渉の効率的な解である(0.68, 0.72)から最も離れた平均値を示したエージェントがエージェント HHH であり、正直に選好を伝えることが効率的な解を生み出すとは限らないことがこの結果からもわかる。

次に、各交渉後に行ったエージェントに対する7段階評価の結果について示す。エージェントごとの評価の平均値は図4のようになった。

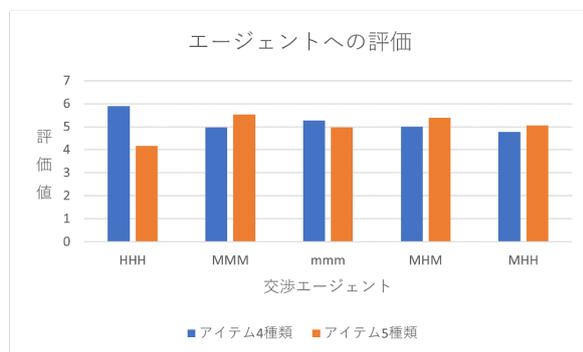


図4: 7段階評価

アイテムが4種類の場合、正直なエージェントの評価値が最高であった。このことから、エージェント HHH は嘘をつかなかったことによって評価値が最も高くなっていると推測できる。

その一方で、アイテム5種類の場合にはエージェント MMM の評価が最大であり、エージェント HHH が最小であった。このことから、エージェント HHH は嘘を用いたエージェントよりも低い評価を受けたことがわかる。これは、アイテムが4種類の場合と異なる結果となった。

さらに、交渉相手が嘘をついているように感じたかについての評価の結果を示し、分析する。得られた結果は図5のようになった。

アイテム4種類の場合、エージェント mmm が最も嘘をついたと評価され、エージェント HHH が最も嘘をついたと評価されなかった。この結果から人間の交渉者は繰り返し交渉の中でエージェントの嘘を見破り、正直なエージェントが嘘をついているとはそれほど考えなかったと考えられる。

また、アイテム5種類の場合にもエージェント MMM が最も嘘をついたと評価され、エージェント HHH が最も嘘をついたと評価されなかった。アイ

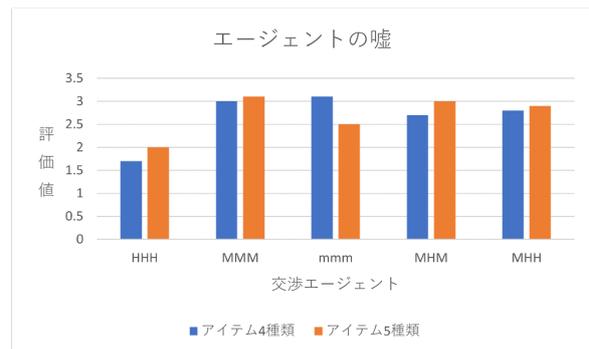


図5: 嘘をついているように感じたかの評価

テム4種類の結果と同じようにエージェント HHH の評価値が他のエージェントと比較して大きく低くなっており、この結果からも人間の交渉者が繰り返し交渉中にエージェントの嘘を見破ったと考えられる。

以上の実験結果から、今回の実験では嘘の戦略を用いることが繰り返し交渉でより利益を生み出すとはいえないことがわかる。嘘の戦略が利益を生み出せなかった原因として、繰り返し交渉では人間の交渉者に嘘が見破られてしまったことがあげられる。このことは、図5で示した結果によって人間の交渉者が嘘を用いたエージェントについて、正直なエージェントよりも嘘をついていると判断したことから支持される。この結果は、単一の交渉で示された結果とは異なるものであった。このように嘘が見破られた要因について考える。人間の交渉者がどのような場合にエージェントが嘘をついていると感じたか、全ての交渉終了後に質問した結果、以下のような回答が得られた。

表1: 嘘をついていると感じた状況

同じ選好と主張されたにもかかわらず、一番好ましくないものをエージェント、一番好ましいものを人間側が受け取るような提案が受諾されたとき
両者ともに一致している一番好ましいものを同数分け合うような提案が拒否されたとき
相手がより獲得ポイントが高くなる提案が拒否されたとき
好ましくないはずのアイテムを要求するような提案が行われたとき
明らかに相手側にとって不利な提案がされたとき

一番目から三番目の回答については、エージェントが人間の交渉者の提案に対して受諾するか、拒否

するか行動について触れられている。一番目は嘘をついていることによって、エージェント視点では互いに好ましいものを受け取る公平な交渉になっているが、人間の交渉者視点ではエージェントにとって不利な提案が受諾されることになり、疑念を抱かされている。同様に二番目は、人間の交渉者視点では公平な交渉であるが、実際はエージェントにとって不利な交渉であるため提案を拒否しており、それが不自然であったと考えられる。三番目では、嘘によって人間の交渉者が本来のエージェントの選好を把握できていないため、人間の交渉者視点ではエージェントにとって有利であっても実際には有利でない可能性があるため、提案が拒否されたのだと考えられる。また、四番目、五番目の回答についてはエージェントが行った提案について触れられている。四番目では、エージェント視点では好ましいアイテムを獲得するような提案になっているが、人間の交渉者視点ではエージェントが好ましくないはずのアイテムを要求しているようにみえるため、嘘をついていると判断されたと考えられる。五番目については、統合的な交渉においてエージェントが好ましいアイテムを多く獲得し好ましくないアイテムを相手に譲るような提案をする場合、人間の交渉者も好ましいアイテムを多く獲得するような提案となる。この提案が、人間の交渉者視点ではエージェントが好ましくないアイテムを獲得しているようにみえるため、不自然であったと考えられる。

5 まとめ

本論文では、繰り返し交渉における嘘を用いたエージェントを提案し、繰り返し交渉において嘘を用いることで、人間とエージェントの獲得効用や人間に与えるエージェントの印象にどのような影響を及ぼすかを分析した。

実験に用いるエージェントの戦略として、既存の研究で単一の統合的な交渉で有効であると示された“fixed-pie lie”と呼ばれる、分配的な交渉を装う嘘を用いる戦略、相手が最も好ましいアイテムについてのみエージェントも同様に最も好ましいと主張する戦略の2つを提案した。これに加えて、正直に選好を伝える戦略の3種類をエージェントの戦略として用いた。実験では、3種類のそれぞれの戦略を1ラウンドの繰り返し交渉内の全ての交渉で連続して使用するエージェント3つと、戦略を使い分けて使用するエージェント2つの合計5つのエージェントで交渉を行い、正直なエージェントと嘘を用いたエー

ジェントの結果にどのような差が生まれるかを確認した。

エージェントの獲得効用については、嘘を用いたエージェントの平均獲得効用が正直なエージェントの平均獲得効用を上回ったものの、有意差は生じなかった。また、人間の獲得効用についても、有意差は生じなかったものの嘘を用いたエージェントの方が正直なエージェントよりも高い効用が得られ、嘘を用いることで効率的な交渉解に近づいた。

人間に与えるエージェントの印象としては、人間の交渉者は、嘘を用いたエージェントが正直なエージェントと比較してより嘘をついたという印象を強く受けた。このことから、繰り返し交渉では単一の交渉と異なり、人間の交渉者が交渉相手の嘘を見破ることが示された。

今後の課題としては、以下のものが挙げられる。

- 繰り返し交渉において有効な、新たな嘘の戦略の提案
- 繰り返し交渉において人間の交渉者がどのように嘘を見破ったかの検証
- 異なる構造の繰り返し交渉での分析

新たな嘘の戦略としては、“fixed-pie lie”に基づかない嘘を用いた戦略の提案や、今回の実験では検証しなかった戦略の組み合わせについての検証などがあげられる。

人間の交渉者がどのように嘘を見破ったかの検証方法としては、エージェントの行動に矛盾が生じていないかの確認が考えられる。特に今回の実験では、人間の交渉者がエージェントが嘘をついたと判断した理由から、エージェントの行動と提案との間に人間の交渉者が違和感を覚えることが予測できる。

さらに異なる構造の繰り返し交渉については、1ラウンドあたりの交渉回数や制限時間、論点数、被験者が行う交渉のラウンド数の変更などがあげられる。

参考文献

- [1] 産業競争力懇談会 COCN. 産業競争力懇談会 2017 年度プロジェクト最終報告人工知能間の交渉・協調・連携. <http://www.cocn.jp/report/theme98-L.pdf>, 2018. (Accessed on 01/17/2021).
- [2] Jonathan Gratch, Zahra Nazari, and Emmanuel Johnson. The misrepresentation game: How to win at negotiation while seeming like a nice guy. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 728–737, 2016.

- [3] Zahra Nazari and Jonathan Gratch. Predictive models of malicious behavior in human negotiations. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 855–861, 2016.
- [4] Johnathan Mell and Jonathan Gratch. Iago: interactive arbitration guide online. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems*, pages 1510–1512, 2016.
- [5] Johnathan Mell and Jonathan Gratch. Grumpy & pinocchio: answering human-agent negotiation questions through realistic agent design. In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems*, pages 401–409, 2017.
- [6] Johnathan Mell, Jonathan Gratch, and Gale M Lucas. The effectiveness of competitive agent strategy in human-agent negotiation. In *Orally Presented at the 2018 American Psychological Association’s Technology, Mind, and Society conference*, 2018.
- [7] Johnathan Mell, Markus Beissinger, and Jonathan Gratch. An expert-model & machine learning hybrid approach to predicting human-agent negotiation outcomes. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 212–214, 2019.
- [8] Ehud Kalai and Meir Smorodinsky. Other solutions to nash’s bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 513–518, 1975.