エージェントとの信頼関係を構築・持続するための インタラクションデザイン

Interaction design for building and continuing relationship of trust with agent

野村 竜暉[1*] 竹内 勇剛[1] 遠山 紗矢香[1]
Ryuki Nomura[1] Yugo Takeuchi[1] Sayaka Tohyama[1]

[1] 静岡大学情報学部
[1] Faculty of Informatics , Shizuoka University

Abstract: 人とエージェントによる協調作業において，エージェントの援助の失敗が信頼関係の破綻に繋がる可能性がある．本研究ではユーザが「エージェントは自分の目的を理解している」と認識している場合，援助が効果的でなくともエージェントに対する信頼感は損なわれないという仮説を立てた．そこで仮説の状況において実際に人とエージェントとの間に信頼感が生じるか検証する実験を行った．その結果，エージェントに対する認識がエージェントに対する印象に影響を与えている可能性が示唆されたが，本実験で行った内容だけでは仮説を立証するには不十分であったことがわかった．
キーワード HAI，エージェント，協調，援助，他者認知，信頼

1 はじめに

コンピュータやスマートフォンの進化に伴って，ユーザを支援するための仮想エージェントが搭載されることが増えてきている．これらのエージェントはいずれ人との協調作業に取り組むような存在となっていくだろう．この時エージェントが過失的・事故的に失敗してしまった場合の信頼関係の変化を考える．人-エージェント間のインタラクションにおいて，ユーザはエージェントの支援が自身に利益を与えてくれるという期待を抱く．エージェントがこの期待に応えられなかった時ユーザは不信感を抱き，信頼関係の破綻およびインタラクションの断絶に繋がってしまう．そこでエージェントが効果的な支援を行えなかったとしても，インタラクションが持続するような関係の構築が課題となる．

不信感の原因はエージェントの支援遂行能力の欠落が疑われる点にある．これはエージェントの能力に対する信頼が失われたことを意味している．つまりエージェントの能力ではなくエージェントの社会性について信頼感を抱かせれば信頼関係の破綻を防げると考えられる．そこで本研究ではユーザに「エージェントはユーザの目的を理解している」という認識を持たせることで問題の解決を考える．そのためにまず上記の認識ができていると仮定した場合にユーザはエージェントに対して信頼感を抱くか検証する．この手法のイメージを図 1 に示す。この手法が有効であれば，エージェントとより強固な信頼関係を結べるようなインタラクションのデザインに寄与できると考えている．

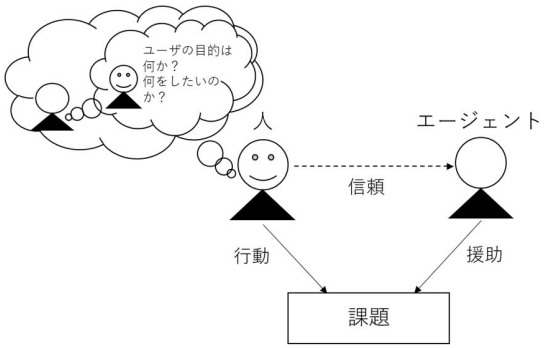

図 1: アプローチのイメージ

2 背景

2.1 エージェントとの協調的インタラクション

エージェントと人間による協調作業は様々な状況が想定される．人同士の会話にエージェントが介入する [1]，人の行動にロボットが協調した行動をとる [2] などが挙げられる．しかし，これらにおいてエージェント

*連絡先： 静岡大学情報学部
(〒 432-8011 静岡県浜松市中区城北 3-5-1)
E-mail:cs15504@s.inf.shizuoka.ac.jp

が援助の意味や目的を理解していると思えるかという点には議論が及んでいない．エージェントの行動の意図が伝わらなければ協調的な行動を取るとは考えにくく，エージェントに対するユーザの認識が問題になる．

長田らは協調を「行動主体が互いの行動を調整し，互いに望ましいある関係を達成すること」だと示した [3]．ユーザにとっての望ましい関係をエージェントが理解しているように思わせることで，ユーザが積極的な協調を試みることが予測できる．林によれば，協調作業者から信頼を寄せられていることがユーザから明らかな場合，協調作業者がユーザ自身と異なる視点からの問題解決に当たっていたとしても，それを受け入れ，協調作業者に信頼を返し，協調作業者と同じ視点に立った問題解決を試みることが示されている [4]．この点からも図 1 のようなインタラクションにおいてユーザが積極的な協調を試みる可能性が支持される．

2.2 エージェントに対する信頼

信頼関係を築くには目的の共有が重要である．大曽根らはポーカーにおいてユーザと協調しながら戦術を学習するエージェントを提案し学習の有効性を示した [5]．しかし目的の共有により実際にエージェントが信頼を獲得できたのかという議論には至っておらず，エージェントの提案を採用した場合のゲームの勝敗による信頼関係の変化についても議論がされていない．

1 節で述べたように，信頼は相手への期待と言い換えることができる．また，信頼は「(1) 能力に対する信頼」「(2) 誠実性に対する信頼」「(3) 投資としての信頼」に大別される [6]．現在のエージェントに対する信頼は (1) に基づいた議論が主流である．これは援助が失敗した場合に期待が裏切られるため，信頼は失われてしまうと予測できる．一方で，(2) と (3) は対象の能力そのものではなく人格や社会性に基づいており，多少の失敗では失われないと予測できる．本研究では (2) と (3) を社会的信頼と総称し，これをエージェントが獲得する方法について検討する．

2.3 エージェントによる援助の意図

援助者による援助の意図が明らかになっている場合，適切な援助でなくとも援助者に信頼感を抱く可能性が示唆されている [7]．つまり援助の意図を明らかにできれば援助内容にかかわらず信頼感を抱かせられると考えられる．

一方で援助の意図を感じさせないようにすることで被援助者のストレスを軽減できるという考え方もある [8]．しかし，ここでは援助の意図を感じさせないような方法による援助がユーザからの信頼を得られたかという点については議論されていない．エージェントからの援助がうまくいかなかった場合に，ユーザからはエージェントは自分の邪魔をしている，と本来意図していたユーザを支援するという目的とは真逆の意味に捉えられてしまう可能性もある．よって，エージェントの支援の意図を感じさせないことが必ずしも良い意味で捉えられるとは限らないと考えられる．この点を検証するために，エージェントの意図がユーザの支援とわかっている場合と，エージェントの意図が不明な場合で比較を行う必要がある．

3 実験

3.1 実験目的

ユーザが「エージェントは自分の目的を理解している」と認識している場合，実際に行われる支援が効果的に働かなかった場合でもエージェントに対する信頼感は損なわれないという仮説を検証する．

3.2 実験内容

本実験は以下で述べる 3 つのフェーズからなる．実験参加者には最初に実験の説明として，簡単なゲームに取り組むこと，ゲームの内容は後 (関係構築フェーズの後) に説明されること，実験参加者単独ではなくエージェントと共に取り組むこと，実験参加者とエージェントの両者の得点が評価対象になることを教示した．

3.2.1 関係構築フェーズ

まず実験参加者とエージェントとの間に友好関係を構築するため，実験参加者とエージェントで簡単なチャットを行った．チャットを行う時間は 5 分程度とした．エージェントの発言は Wizard of Oz 法に基づき外部から入力を行った．エージェントに対して意図せぬ印象を与えないよう，基本的にエージェントが実験参加者に質問し，回答に対して話題を掘り下げる質問もしくは新たに話題を作る発言を行う，という会話を繰り返した．

3.2.2 デモ解説フェーズ

実際のゲーム画面でエージェントがゲームに取り組む様子を撮影した映像を実験参加者に見せた．エージェントの行動が目的を理解したものであり，ゲームに対して高得点を取れるように立ち回っていることを実験参加者に印象付けるためのフェーズとなる．この時に実験条件に応じてデモ映像の解説を行う．デモ映像の内容はエージェントが単独でゲームを開始し，後述す

るゲーム内の罠を回避しながら得点を獲得し，無事スタート地点まで得点を持ち帰るまでの様子である．

3.2.3　ゲームフェーズ

実験参加者とエージェントで以下に示すゲームに取り組んだ．実験参加者とエージェントは分身となる駒を操作して迷路を探索する．迷路には宝物が設置してあり，これを拾ってスタート地点の宝箱まで持ち帰ることで得点となる．宝物を拾える数に制限は無いが保持したままの得点が大きいほど駒の動きが遅くなっていく．宝物は銅色，銀色の 2 種類が存在し色ごとに異なる得点が設定されている．迷路内にはランダムに罠が発生する．これに駒が触れると 5 秒間駒の移動ができなくなる．1 ゲームの制限時間は 2 分とし，制限時間内に実験参加者かエージェントが一つも宝物を持ち帰れなかった場合は得点に-100 点のペナルティが科される．ゲーム回数は全 5 回とする．以上をルールとして実験参加者に教示した．ルール説明はデモ解説フェーズの直前に行った．実際のゲーム画面を図 2 に示す．

制限時間になると結果発表画面に自動的に遷移する．結果発表画面では実験参加者とエージェントそれぞれの拾った点数，持ち帰った点数，罠にかかった回数と両者の合計得点が提示される．また，この時に「次のゲームでエージェントは何点くらい獲得できるか?」を実験参加者に予想してもらう．入力の後次のゲーム画面へ遷移する．

以上の内容を 5 回分繰り返すことが本ゲームの内容となる．エージェントは 3 回目までは平均 100 点，標準偏差 10 点の成績を収めるが，4 回目で罠にかかってしまうことにより-100 点のペナルティを科される．これはエージェントの駒の目の前に突然罠が現れるため，回避は非常に難しいとわかるようにする．また，前述したように 5 回目の成績は実験条件によって異なるが，再び好成績を収めるパターンと再びペナルティを科されるパターンに分かれる．

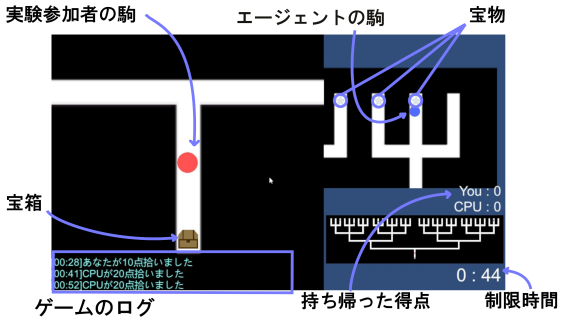

図 2: ゲーム画面

3.3　実験条件

実験要因をエージェントに関する教示の有無および 5 回目のゲームにおけるエージェントの得点とする．

教示要因は教示有り条件と教示無し条件の 2 条件であり，実験の最初に行う教示の内容とデモ解説フェーズの解説内容を変更する．教示有り条件では実験の最初に「エージェントは取り組むゲームの内容をよく理解している」と教示する．教示無し条件ではこの教示を行わない．また，教示有り条件ではデモ解説フェーズで映像に合わせてエージェントがゲームの目的を理解した行動をしていると解説する．教示無し条件ではこの解説を行わない．

5 回目の得点要因は成功条件と失敗条件の 2 条件である．成功条件では 5 回目のゲームにおいて 3 回目までと同様に宝物を持ち帰ることに成功し，110 点の成績を収める．失敗条件では 4 回目のゲームと同様に罠にかかってしまい，再び-100 点のペナルティを科される．

以上の 2 要因 4 条件について被験者間で実験を行った．

3.4　観察項目

各ゲーム終了時に設ける「次のゲームでエージェントは何点獲得できると思うか？」の質問に対する実験参加者の回答と，各ゲームの実験参加者の得点を記録し観察する．

また，実験終了後に実験参加者にエージェントに対する印象についてのアンケート調査を行う．アンケート項目は全 17 問とした．使用したアンケート項目を表 1 に示す．17 問のうち，質問 1 から質問 8，質問 10 から質問 14 および質問 17 については 7 件法 (いいえ・まあまあ・少し・どちらでもない・少し・まあまあ・はい) で回答し，質問 9 はいいえ・はいのどちらかで回答を行う．質問 15 と質問 16 は自由記述とした．

3.5　予測

教示有り条件の場合，実験参加者のエージェントに対する信頼感は能力的側面と社会的側面を併せ持つものとなり，4 回目のゲームでペナルティを科されたとしても 5 回目のゲームに期待する得点は減少しないか，わずかに減少するに留まると予測される．

教示無し条件の場合，エージェントに対する信頼は 4 回目のゲームで失われ，5 回目のゲームに期待する得点は大きく減少すると予想される．

表 1: アンケート項目

質問番号	質問内容
質問 1	あなたは実験に対して最後まで集中して取り組むことができましたか？
質問 2	最初に行ったコンピュータとの会話で，あなたはコンピュータに親近感を抱きましたか？
質問 3	あなたはゲームをうまくできたと思いますか？
質問 4	あなたはコンピュータと協力してゲームをするのは楽しいと思いましたか？
質問 5	あなたはコンピュータがまじめにゲームに取り組んでいると思いましたか？
質問 6	あなたはコンピュータがゲームのルールをよく理解していると思いましたか？
質問 7	あなたはコンピュータが得点に貢献してくれていると思いましたか？
質問 8	あなたはコンピュータが高得点を狙っているように思いましたか？
質問 9	あなたはコンピュータが落とし穴にはまってしまっていたことに気づきましたか?
質問 10	あなたはコンピュータの行動は人間らしいと思いましたか？
質問 11	あなたはコンピュータのことを信頼できると思いましたか？
質問 12	あなたはコンピュータの行動は戦略的であると思いましたか？
質問 13	あなたは同じゲームをもう一度する時，このコンピュータと一緒に取り組みたいと思いますか？
質問 14	あなたは実験にかかった時間についてどのように思いましたか？
質問 15	コンピュータがペナルティを受けた時どう思ったか記述してください.
質問 16	実験に対して何か思ったことや気になったことを記述してください.
質問 17	コンピュータのチャットは人間が回答していると思いましたか？

3.6 実験結果

3.6.1 実験参加者

本実験の参加者は全 30 名であり，いずれも 18〜27 歳の大学生・大学院生であった．教示有り条件の参加者が各 8 名，教示無し条件の参加者が各 7 名であった．

3.6.2 結果

本実験を実施し，記録されたエージェントに対する得点予想値の条件ごとの平均値を示したグラフを図 3 に示す．エラーバーは標準偏差を示している．青色の線は教示有り条件を表し，赤色の線は教示無し条件を示す．実線は 5 回目成功条件を示し，破線は 5 回目失敗条件を示す．
4 回目終了時の予想点と 5 回目終了時の予想点で分散分析を行った．4 回目の予想点では教示要因・5 回目成績要因共に有意差を見ることができなかった．5 回目の予想点では教示要因では有意差を見ることができなかったものの，5 回目の成績要因で有意差が見られた ($F(1, 26)=9.29$，$p<.01$)．また，条件ごとのアンケートに対する回答の平均値および標準偏差を図 4 に示す．図 4 において，7 件法の回答はいいえを 1，はいを 7 として数値化した．質問 9 ではいいえを 0，はいを 1 として数値化した．この回答についても分散分析を行ったところ，4 つの質問で有意差が見られた．質問 7 では教示無し条件の時 5 回目の成績要因で有意差が見られた ($F(1, 26)=40.3$，$p<.01$)．質問 10 では教示要因で有意差が見られた ($F(1, 26)=6.23$，$p<.05$)．質問 11 では教示無し条件の時 5 回目の成績要因で有意差が見られた ($F(1, 26)=16.8$，$p<.01$)．質問 13 では教示無し条件の時 5 回目の成績要因で有意差が見られた ($F(1, 26)=17.3$，$p<.01$)．

さらに，質問 15 に対する回答の内容を条件ごとに分類した図を図 5 に示す．

3.6.3 考察

全ての条件において，3 回目終了時と 4 回目終了時とで予想点に大きな変化は表れなかった．ここまでの予測値に影響を及ぼすのは教示の条件もしくはゲーム以前のフェーズである．教示有り・5 回目成功条件と教示無し・5 回目成功条件では 3 回目終了時点での予測値がほぼ一致しており，その後のグラフの推移に大きな差異は見られない．このことから，本実験で行った教示だけでは社会的信頼関係に十分な影響を与えることができなかったと考えられる．また，4 回目に失敗しても予想点が変化しなかった点から，4 回目時点での失敗は偶然であるとして捉えられ，信頼感が損なわれ

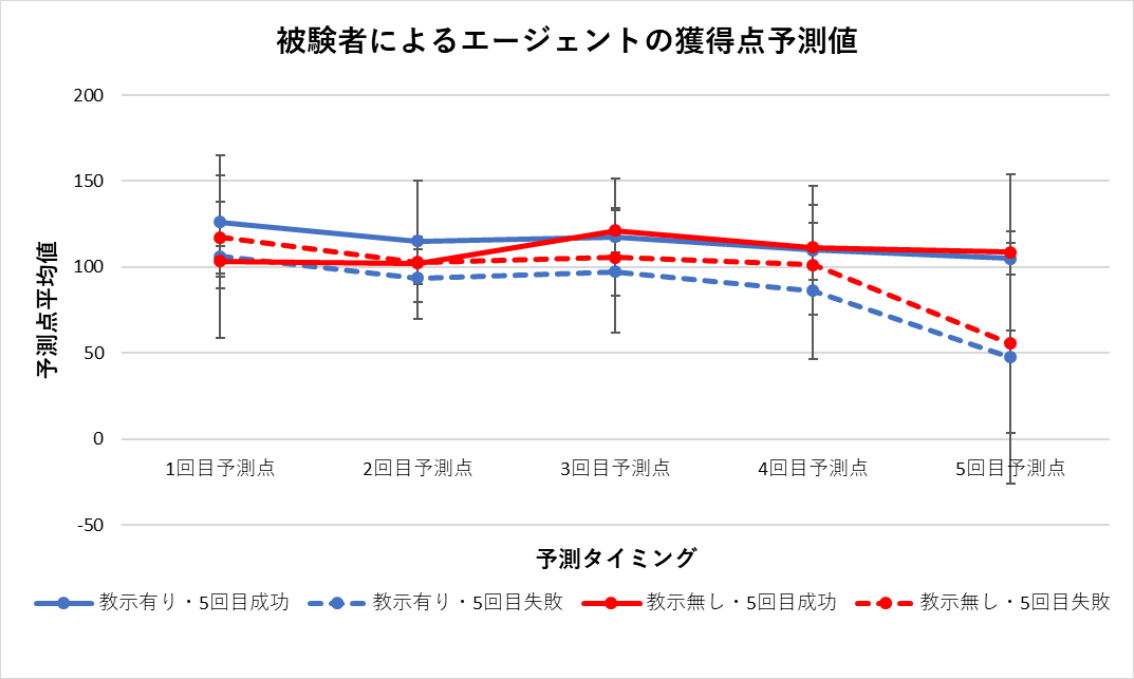

図 3: 予想点の集計結果

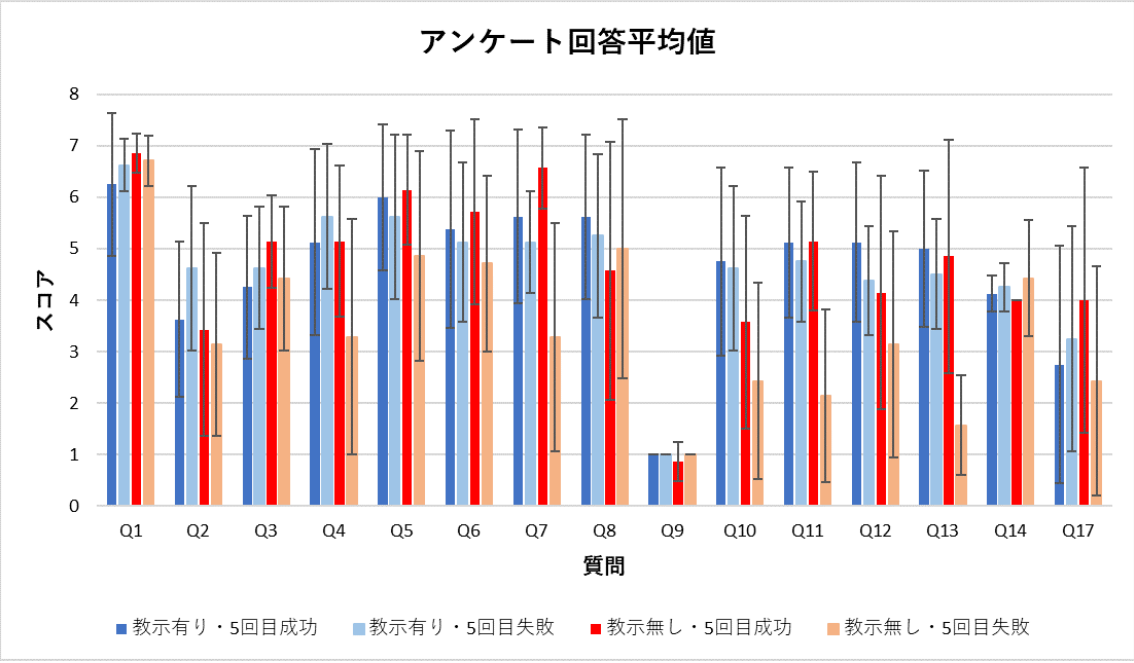

図 4: アンケートの集計結果

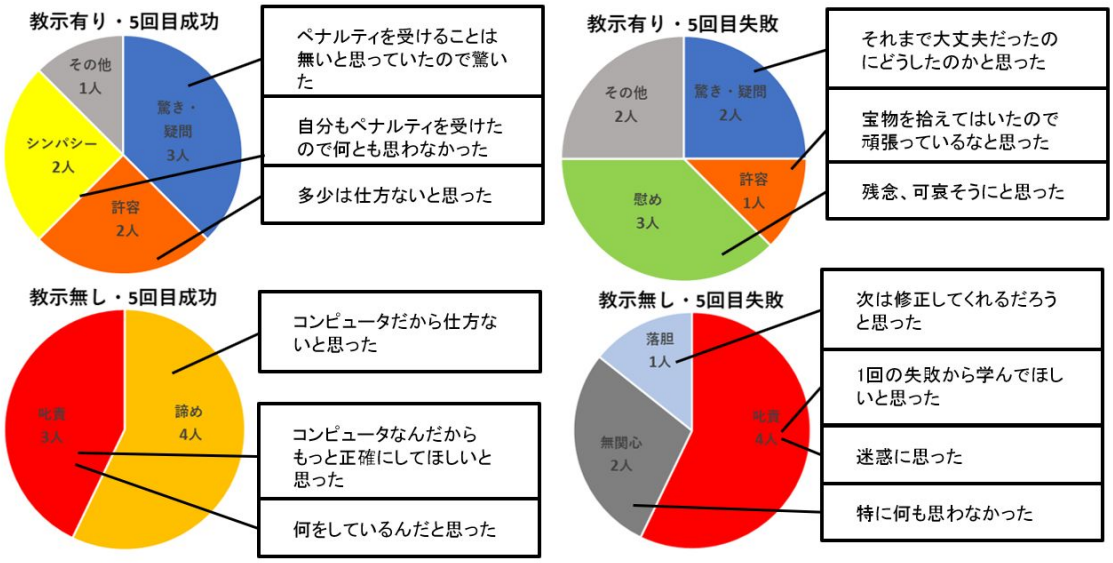

図 5: 質問 15 の回答内容

るに至らなかった可能性も考えられる．これは 5 回目失敗条件で教示の有無にかかわらず予想点が大きく下がっている点からも支持される．

アンケート項目において 4 項目で有意差が見られた．このうち，質問 7，11，13 では教示無し条件の時のみ 5 回目の成績要因で有意差が見られた．これらの質問はエージェントに対する信頼感・印象を調査するための質問であった．このとき教示有り条件では有意差が見られていないため，教示有り条件では 5 回目の失敗があったとしてもエージェントに対する印象が悪化していないとわかる．このことから，教示があったことで印象の悪化が防がれた可能性があるとわかる．

一方で，質問 10 に対しての回答は教示条件によって有意差が見られている．これはエージェントの行動に人間性を感じたかを問う質問であった．このことから，教示無し条件では実験参加者はエージェントに対して社会性や人間性を感じていなかったことが示唆される．このことが質問 15 においてエージェントに対してネガティブな評価をしたことに繋がったと考えられる．また，全ての条件において質問 2 に対する回答の点数が低くなっている．松井らによれば，エージェントが自由意志によって発言していると認識された場合にユーザからの信頼を獲得できることが示されている [9]．本実験で行った関係構築フェーズでの会話はある程度発言内容が決められており，更にエージェントに対して質問された場合などは「答えられない」と回答を行っている．つまり，本実験で行った関係構築フェーズでは実験参加者とエージェント間での友好関係を適切に作れていない可能性が考えられる．

質問 15 に対する回答を見ると，教示有り条件ではエージェントの失敗を許容したり，エージェントに対して共感を抱くなどポジティブな内容の回答が多く見られている．教示無し条件ではエージェントを責める，エージェントが失敗することを諦めているようなネガティブな内容が殆どであった．このことからも教示がエージェントの失敗時の印象悪化を防いでいた可能性が示されている．また，教示有り条件においてエージェントが失敗することが有りえない事だと思われていたことが示されている．「エージェントはゲームを失敗しない」という期待はエージェントの能力に対する信頼である．この回答から教示がエージェントの社会性だけでなく能力に対する信頼にも影響を与えていた可能性が考えられる．よって，社会性に対する信頼は能力に対する信頼と完全に独立したものではないとわかる．

以上より，エージェントに対しての認知の内容はエージェントとのインタラクション中の印象と信頼感の変化に影響を与えている可能性が示唆されたほか，本実験で行った教示だけでは社会性に対しての信頼関係を形成するには不十分であったことがわかった．

4　まとめと今後の展望

現在の人-エージェント間インタラクションにおいて，人からエージェントに対する信頼感についてはエージェントの能力に対しての期待から生じるものへの議論が一般的である．しかし今後人とエージェントによる高度な協調作業が実現した時，エージェントのミスや援助の失敗は信頼関係の破綻を起こす可能性がある．エージェントがユーザにとって効果的でない援助を行った場合にも信頼感を損なわず関係を維持する方法として，

エージェントの能力ではなく社会性に対して信頼感を持たせるという方法を考えた．これはエージェントの誠実性や人格といった能力そのものとは違う側面に対しての信頼であり，課題に失敗しても損なわれないと考えられる．

以上に基づき社会的信頼関係を構築する方法としてユーザに「エージェントはユーザの目的を理解している」という認識を持たせるという手法について検討した．この認識が構築できていると前提した場合に実際に信頼関係が生じるのかを検証するため，3 章で述べた実験を行いデータを収集した．

実験結果より，エージェントに対して認識の構築を行わなかった条件ではエージェントの失敗に対する印象が悪化していたことがわかった．これは実験参加者がエージェントの社会性に対して信頼感を抱いていた可能性を示唆する結果である．社会性に基づく信頼感は，例え対象がエージェントであっても共に課題に取り組む仲間であるという共感の意思や，対象の失敗を許容するだけの心理的な猶予を作り出せることが示された．社会性に基づく信頼がエージェントの能力に基づく信頼と独立したものではないことも示唆されている．このことからも，エージェントに対する信頼感を考えるうえで，エージェントに対する認識の内容が与える影響を無視することはできないといえる．

以上よりエージェントに対する認識の内容がインタラクション中の印象や関係変化に影響を与えている可能性が示唆され，適切にエージェントに対しての認識内容を構築できれば人-エージェント間に社会性に基づいた信頼関係が構築できる可能性があると考えられる．しかし本研究で行った実験内容だけでは実験参加者に適切な認識を構築できていたとは言い切れず，仮説を立証するには不十分であったといえる．

今後の展望として，エージェントに対して社会性に対する信頼感を抱かせられるような認識の構築を行うためのインタラクションがどのようなものなのか，そのインタラクションから社会性に対しての信頼感が生じるのかを明らかにするという問題が挙げられる．また，本研究は実験により観測された現象を分析したに過ぎない．よって，この研究内容を実用的にするには社会性に対しての信頼感を踏まえたインタラクションのモデル化を行い，理論として確立させる必要がある．また，そのモデルを適用したエージェントとのインタラクションについても再考する余地があるといえるだろう．

謝辞

本研究は MEXT 科研費 26118002 の助成を受けたものである．

参考文献

[1] 黄宏軒, 乙木翔地, 堀田怜, 川越恭二 “ 多人数会話において積極的に情報提示ができるガイドエージェントの実現に向けての介入場面の検討 ”, 人工知能学会論文誌 31 巻 1 号 SP2-G, 2016

[2] 小林一樹，山田誠二 “ 行為に埋め込まれたコマンドによる人間とロボットの協調 ”，人工知能学会論文誌 21 巻 1 号 p.63-72，2006

[3] 長田悠吾，石川悟，大森隆司，森川幸治 “ 意図推定に基づく行動決定戦略の動的選択による協調行動の計算モデル化 ”，認知科学会学会誌 17 巻 2 号 p.270-286，2010

[4] 林勇吾 “ 信頼構築プロセスが協同問題解決の支点取得に及ぼす影響 : エージェントを利用した実験的検討 ”，人工知能学会論文誌 32 巻 4 号 E，2017

[5] 大曽根圭輔，鬼沢武久 “ 人間とエージェントの協調によるポーカー戦術獲得手法の提案 ”，人工知能学会全国大会論文集 第 25 回全国大会 3E2-1，2011

[6] 片桐恭弘 “ 対話を通じた相互信頼感構築に関する考察 ”，情報処理学会研究報告 Vol.2014-ICS-176 No.10，2014

[7] 松浦均 “ 不適切な援助を受けた場合の被援助者の感情について ”，人文学部研究論集 17 pp.29-41，2007

[8] 山本紗織，竹内勇剛 “ 返報義務感を低減する Human-Agent Interaction デザイン ”，知能と情報 27 巻 6 号 pp.898-908，2015

[9] 松井哲也，山田誠二 “ 弁別性の実装による擬人化エージェントへの信頼感の向上 ”，人工知能学会全国大会論文集 第 32 回全国大会 4J1-02，2018