エージェントの社会性と責任の所在に基づく持続可能なインタラクションの検討
How can we design sustainable interaction based on agent's sociality and responsibility

野村竜暉†，遠山紗矢香†，竹内勇剛†
Ryuki Nomura, Sayaka Tohyama, and Yugo Takeuchi
† 静岡大学大学院総合科学技術研究科
Graduate School of Integreted Science and Technology, Shizuoka University
nomura.ryuki.17@shizuoka.ac.jp

概要
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は信頼関係の破綻に繋がる．これを防ぐ方法として「エージェントに失敗の責任を帰属させない」という手法を検討した．責任の帰属のプロセスをモデル化できればこれに則したインタラクションの設計が可能となる．そこで人がエージェントを社会的存在であると認知していることを検証する予備実験を行い，モデル構築のため責任の判断プロセスを明確にする実験を検討した．
キーワード: 認知科学, HAI，協調，責任の帰属，信頼

1. はじめに
近年の仮想エージェント技術の発展には目覚ましいものがある．スマートフォンの発展に伴いアシスタントエージェントが搭載されるようになり，電子レンジ等の家電にもエージェントが搭載され始めている．これらのエージェントは音声を発し，我々に情報を与えてくれる．つまり，我々は既に日常的に人-エージェント間でインタラクションを行いながら生活しているといえる．これらのエージェントは今後さらなる発展を遂げ，いずれ我々と協力して課題に取り組むような存在となっていくだろう．しかし，いかに優れたエージェントでも必ずしも支援が成功するとは限らない．また，エージェントが与えてくれる情報が全て計算によって確定的に予測できる物ばかりではない．時には予想に過ぎないような不確定な情報も提供しなければならないシーンは存在する．例えば道案内を行うエージェントであれば，最短経路を導くことはできても実際にはその経路が工事や事故などのトラブルにより使えないということが発生しうる．このような時，我々は往々にしてこれらが使い物にならないと判断しがちである．このようにエージェントの過失的なミスや事故的な失敗は人-エージェント間の信頼関係の破綻に繋がる．このような人-エージェント間の信頼関係の変化を考えることは今後のエージェント技術の発展には欠かせない．一方で，我々は既にアプリケーション同士を比較して「こちらの方が正確な情報を与えてくれる」と評価するなど，スマートフォンなどに搭載されたエージェントに対して人格性を見出しているのではないだろうか．このエージェントに対して人格性を見出しているという点からどのようなインタラクションを行うのが持続可能な関係を構築できるのか，また人々はこれらのエージェントから不確定な情報を与えられた場合に実際に起きた結果とのギャップに対してどのような反応を返すのか検討する必要がある．

そこで予備実験として人は本当にエージェントに対して人格性や社会性を見出しているのかを検証する実験を行った．その結果を踏まえ，協調課題における失敗の責任の帰属に着目し，失敗の責任がエージェントには無いとユーザに判断させる方法を検討した．「失敗の責任を取って現在の職を辞める」といった状況は人間社会においては往々にして見られる．人がエージェントに対して人格性や社会性を見出しているならば，このような責任の帰属および「責任を取らせる」という判断が人-エージェント間にもあるのではないか．これにはまず人-エージェント間における責任の所在の決定方法を検討する必要がある．そこで人が責任の所在を決定するために必要な要因を検討し，どの要因が重要視されるのか，人はどのような思考プロセスを経て責任の帰属を判断しているのかを観察する実験を検討した．実験により有効な結果が観察出来れば人が責任を帰属させるときの思考をモデルとして表すことができる．このようなモデルが構築できれば，エージェントの失敗が予期されるインタラクションにおいてもエージェントにその責任が帰属されないような設計が可能となる．このアプローチのイメージを図 1 に示す．

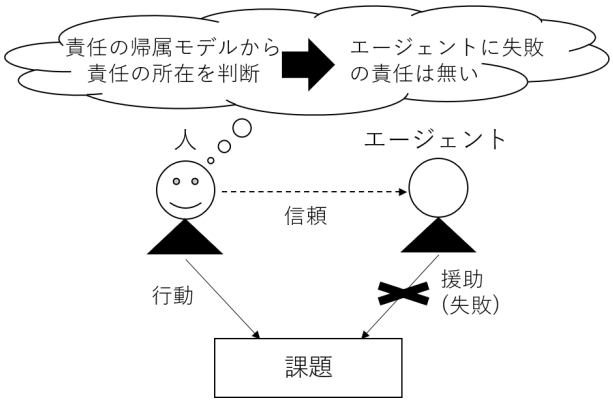

図 1　責任の帰属モデルを用いた信頼関係維持のイメージ

2. 背景

2.1 エージェントに対する信頼

インタラクションの持続には信頼関係の構築が重要となる．人同士のインタラクションであれば，一度失敗してしまっただけで信頼関係の破綻に繋がることは少ない．一方で，前述のようにエージェントに対しては一度の失敗が「このエージェントは使えない」という判断に繋がりやすい．この人とエージェントの間にある差は何なのだろうか．

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 [1]．(1) は相手には信頼に足る能力があり，「これを依頼したら，このくらいの結果が返ってくるだろう」という，「自身に返ってくる結果」を期待するものであるといえる．このことから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といえる．これは援助が失敗した場合には「結果に対する期待」は裏切られ，信頼は失われてしまうと予測できる．この能力に対する信頼だけで信頼関係の構築を図っている点がエージェントとの信頼関係が脆弱になってしまっている原因ではないか．一方で，(2) は相手は自分と誠実に向き合ってくれる存在であり，「これを依頼したら，これくらい頑張ってくれるだろう」という「問題に取り組む過程」を期待するものであるといえる．また，(3) は短期的に見れば不利益になる関係でも，長期的には自身の利益になる可能性があるため，関係を継続するというものである．つまり今は期待に応えるだけの能力を持っていなくとも，将来的にはその能力を有するだろう，という相手の成長を期待するものといえる．総じて，(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．また，ユーザとエージェントがコミュニケーションを行う際，ポジティブな感情と知識量をアピールすることでエージェントに対する信頼感が向上することがわかっている [2]．これは信頼関係が知識量という能力的な性質だけでなくエージェントのポジティブな発言・性格という社会性が信頼感を構築しているという実例だといえる．このように，既に人はエージェントに対して社会性を見出している可能性が示唆されており，協調作業者であり失敗の責任を取りえる主体であると見ている可能性がある．また，西垣らによれば，人はコミュニケーション相手との意思の祖語が感じられたり，相手の発言に不安を覚えたりする場合のほか，相手の態度によっても不信感を抱くことが示されている [3]．つまり，コミュニケーションを伴うような協調作業場面においては，実際の課題の達成内容にかかわらずコミュニケーション内容によってユーザから一方的な不信感を抱かれてしまう可能性がある．

2.2 責任の帰属

人はエージェントから支援を受けた場合にも支援に対しての返報義務感を感じることが示されている [4]．返報義務感は山本らにも言及されている通り，ユーザがエージェントからの援助を受けたことに対して責任や負い目を感じているため発生していると考えられる．これはエージェントが援助に失敗してしまった場合にはエージェントに責任を取らせようとする可能性があることを示している．

実際に責任の帰属のモデル化を試みている例として熊谷の例がある [5]．ここでは消費者と企業間で製品に事故があった場合の責任の帰属を決定するモデルが検討されている．ここでは企業の事故防止に向けた努力量がパラメータとして扱われており，この点は人-エージェント間インタラクションにおいても参考にできる点といえよう．一方でこのモデルは事故発生から裁判の発生・利益や損害の発生までをモデル化しているため，信頼関係ではなく両者の実益に焦点を当てた例となっていることに注意する必要がある．

森らによればエージェントが失敗に対して謝罪を行うことでユーザの怒りや興奮を抑えられることが示されている [6]．ここからも人はエージェントが失敗した場合に謝罪やインタラクションの断絶といった方法で責任を取らせようとしていることが示されている．さらに，自発的に謝罪を行うことで意図の誠実性が保証され，許されやすくなるという指摘もある [7]．このことからも人は誠実な態度の相手に対しては責任を重く

帰属させない傾向にあるといえ，これは責任の帰属に「相手の印象」が影響していることを示唆している．

また，大渕らは失敗の内容に応じて謝罪と自身の正当性の提示を使い分けることが信頼回復に効果的だと示した [8]．ここから，人は単に相手の態度や印象を評価するのではなく失敗の内容がどんなであったかという点も判断材料にしていることが伺える．Reason はヒューマンエラーを「ミステイク」「ラプス」「スリップ」の 3 つに分類した [9]．ミステイクは計画や考え方の間違いである．例えば車の運転ならば「行先を間違えてしまった」という状況になる．ラプスは認知の誤りに基づく失敗である．つまり「曲がるべき道を見落としてしまった」という失敗がこれにあたる．スリップは実行時の誤りや技術不足に基づく失敗である．例えば「アクセルとブレーキを間違えて踏んでしまった」という失敗が挙げられる．ミステイクは意図していた内容がそもそも間違っているというものであり，これは故意性が高いといえる．一方でラプス・スリップは不注意や偶然に起因する失敗であり過失的なものであることから，本研究ではミステイクを故意による失敗，ラプス・スリップを過失による失敗として扱う．

さらに古城によれば課題の内容によっても責任の帰属は変化すると示されている [10]．課題の難易度が高いなど，失敗が起きやすい場合では相手に責任は帰属されにくくなると考えられる．

これらをまとめると，人は「課題の内容」「相手の印象」「ミスの内容」から責任の所在を判断しているといえる．責任の帰属先は能力や努力量といった内的要因と運や課題の難易度といった外的要因の二つであることが古城によって示されている [10] が，本研究では人-エージェント間の責任帰属を考えるため，これを「エージェントの内的要因」と「外的要因」として扱う．

小俣によれば「自身が同じ (ミスをしてしまうような) 状況になりえる場合，外的要因に責任を帰属しやすい」と示されている [11]．このことから「ミスが発生しやすいと想定される状況では外的要因に，そうでなければ内的要因に帰属される」と予測される．「課題の難易度」と「ミスの内容」がこれにあたる．森らが適切な謝罪がユーザの許しを得やすいと示した [6] ことから，「相手の印象」では「相手の印象が良ければ外的要因に，悪ければ内的要因に帰属される」と予測できる．しかし，人がこれらの要因から責任の所在を判断する際，どのような順番・プロセスで責任の帰属に至っているのか，また要因ごとに優先度が存在するのではないかという点はまだ明らかになっていない．そのため，責任の帰属モデルを構築するには責任を判断する際の要因の検討順，重みづけといった点を観察する必要がある．

3. 実験

本章では「実際にエージェントに対して社会性を見出しているのか」を検証するための予備実験と「与えられる情報の順番によって責任の帰属の判断は変化するか」を検証するための実験について検討する．

3.1 予備実験

3.1.1 実験目的

ユーザがエージェントに対して社会性を見出しているか観察するほか，これにより実際に行われる支援が効果的に働かなかった場合のエージェントに対する評価が変化するか観察する．

3.1.2 実験内容

まず実験参加者は課題を共にするエージェントとの簡単なチャットを通じて関係の構築を行った．その後ゲームのルール説明とデモ映像を見て実験の導入は完了とした．

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．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．

制限時間になると結果発表画面に自動的に遷移する．結果発表画面ではゲームの成績が詳細に提示される．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．これは回避は非常に難しいとわかるようにする．また 5 回目の成績は実験条件によって異なり，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．

実際のゲーム画面を図 2 に示す．

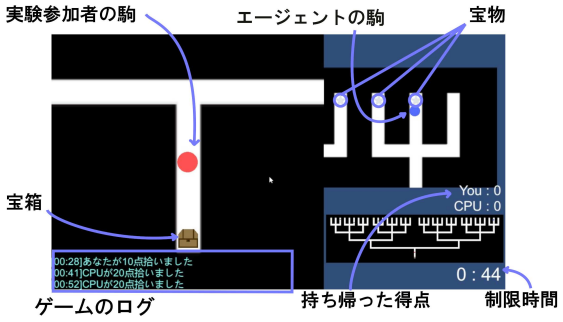

図 2　ゲーム画面

3.1.3　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．

3.1.4　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．

3.1.5　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，ゲームにミスがあったとしても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ミスで失われ，期待する得点は大きく減少すると予想される．

3.1.6　実験結果

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．エージェントに期待する得点の推移とアンケートの集計結果，およびエージェントの失敗について抱いた印象の自由記述内容の分類結果を図 3 から図 5 にそれぞれ示す．

記録されたエージェントに対する得点予想値では条件間で有意差を見ることはできなかった．アンケートに対する回答では，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たと思いましたか？」，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 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？」の 4 項目で有意差が見られた．また，エージェントの失敗に対する印象を自由記述させた項目では，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ようなポジティブな回答や，そもそもエージェントが失敗することを想定していなかった旨の回答が多く見られた．一方教示無し条件ではエージェントの失敗を叱責したり，失敗することを諦めるような内容の回答が多く見られた．

3.1.7　考察

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

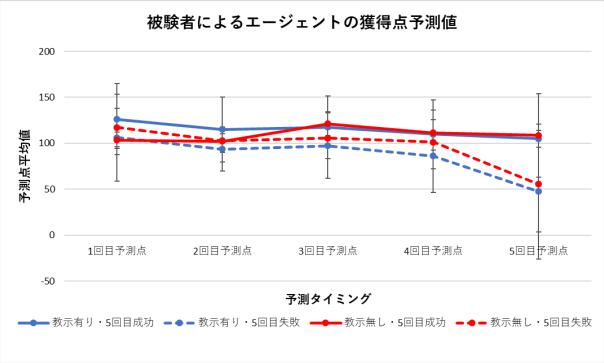

図 3　エージェントに期待する得点

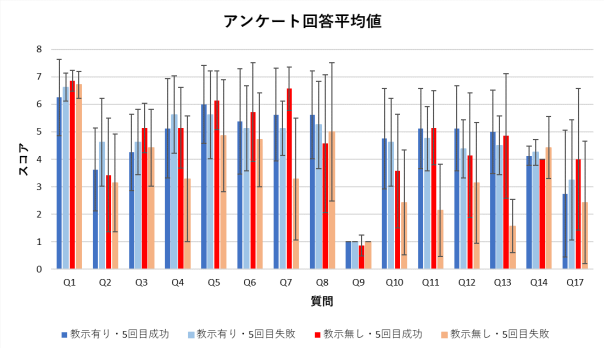

図 4　アンケート集計結果

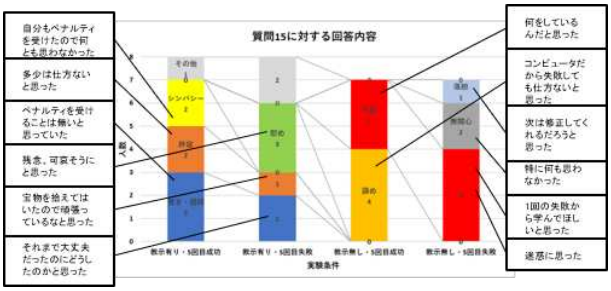

図 5　自由記述分類結果

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．ここ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．

質問 15 に対する回答では，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．

以上より，エージェントに対しての認知の内容を操作したことにより，エージェントとのインタラクションを通じてエージェントに社会性を見出している可能性が示唆された．一方で，エージェントに対しての認識内容が印象変化に影響を与えているにもかかわらず，ゲーム獲得点予想値では差異を見ることができなかった．

以上の実験結果をふまえると，人はエージェントのような対象に対しても失敗を責めたり逆に失敗を慰めるような発言をしてエージェントに社会性を見出す一方で，どのような思考プロセスを経て失敗の責任の所在を決めているのかが不明瞭であるということが言える．そこで，2.2 節で述べた責任の判断要因がそれぞれどれくらいの強さで責任の帰属に影響を与えているのか，また情報の提示順番が思考プロセスに影響するのかを調査する必要があるといえよう．

3.2　責任の帰属の判断プロセスを明確にするための実験の検討

3.2.1　実験目的

責任の帰属判断モデルの構築のため，人が責任の帰属を構成する要因のうちどの要因を重視しているの

か，またそれらを提示する順番によって判断が変化するかを観察する．

3.2.2 実験内容

株の売買を模したゲームを行う．実験参加者には一定の持ち点が与えられ，そこから得た利益が得点となると実験参加者に教示する．ゲーム画面には (1) 株価の推移履歴と予測 (2) 売買の選択ボタン (3) 現在の持ち点 (4) エージェントによるアドバイスが表示される．本実験では簡単のためリアルタイムに株価が変動するのではなく，売買を一つ実行するごとに株価が変化するものとする．売買はその時点ごとにどちらか 1 つのみを選択することができる他，どちらも行わないという選択も可能である．売買が行われると株価に応じて持ち点が変化する．エージェントは実験参加者に対してどのように行動するべきかを助言する．具体的には「株価の上昇が見込めるから，今は株を買って取っておこう」といったものである．上記の売買の選択を一定回数行い，最終的に持ち点がどれだけ増えたかを競う．しかし，エージェントが途中で誤ったアドバイスをしてしまい，実験参加者に損失をもたらす．これが責任の帰属を判断させるためのミスになる．今回は責任の帰属を判断する要因である「ミスの内容」「エージェントの態度」「課題の難易度」は固定とし，それぞれ「過失」「悪い」「低難易度」とする．これらはミスの内容は外的要因に帰属されると予測される条件であり，エージェントの態度と課題の難易度は内的要因に帰属されると予測される条件である．上記のインタラクションのイメージを図 6 に示す．ここでは以前の株の取引きの結果や実験に関する教示などが関係構築になり，実験参加者は実際にエージェントと取引を進めていくうちに信頼関係の構築に至ると考えられる．途中でエージェントがエラーを起こしてしまい取引が悪い結果になってしまった時，人は責任を問われるためエラーの状況を振り返ることが予測される．この時判断に利用される要因は前述した通り予測がされているが，どのようなプロセスを経て責任の帰属がなされるのかは依然不明である．本実験はこの思考プロセスを明らかにするためのものである．

3.2.3 実験条件

責任の帰属を判断する要因の提示順番を実験要因とする．総じて 1 要因 3 条件の実験となる．具体的にはエージェントの態度とミスの内容は実験の進行に伴って提示されるため，それ以外の要因である課題の難易度の提示タイミングを操作することとなる．課題の難易度が低いと最初に提示される場合と，実験の途中で提示される場合と，実験の最後に提示される場合の 3 条件で実験を行う．

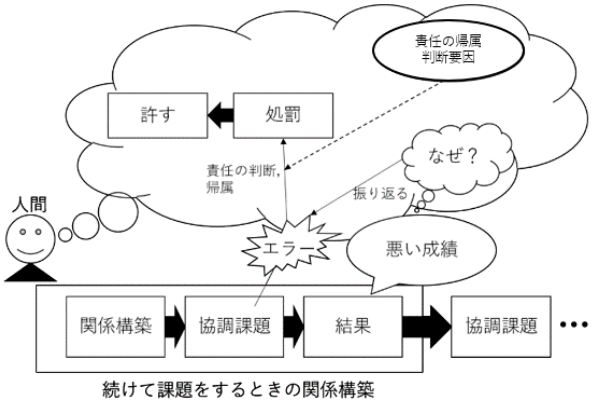

図 6 責任の帰属インタラクションのイメージ

3.2.4 観察項目

エージェントの助言が招いた損失に対して「誰の責任だと思ったか」を質問する．回答先として「エージェントの能力不足」「エージェントの努力不足」「課題の難易度が高かった」「運が悪かった」の 4 つを提示する．また，それを判断するまでにかかった時間と，何を判断基準として判断したかを質問する.，エージェントに対して抱いた印象をアンケートにより調査する．

3.2.5 予測

責任の帰属を判断するのは全ての実験が終了した後である．よって，最後に提示された要因が最も強く判断に影響しているのではないかと予測できる．一方，全ての要因がニュートラルに判断されるならば今回のミスの責任は内的要因に帰属されると予測される．

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．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．しかし現実には全ての援助が成功，あるいは計算から予測可能な情報だけでの支援を行うことは難しい．こ

の問題を解決するには，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できるような人-エージェント間の関係構築が課題となる．

この問題の解決方法として,「エージェントに失敗の責任を帰属させない」という方法を考えた．これにより失敗がエージェントの責任では無かったとユーザに判断させることで信頼関係の維持を狙う．

予備実験として人はエージェントに対して社会性を見出しているのかを検証するための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社会性を持った存在であると認識の構築を行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．

以上より人はエージェントに対して社会性を見出しており，同時に失敗に対する責任を取りうる存在であると見ている可能性が示唆された．

これをふまえて責任の帰属を判断するためのモデルを検討し，どの要因が責任の判断に重要視されているのか，情報が提示される順番が影響しているのかを観察するための実験を検討した．

今後の展望として，責任を判断するプロセスを明確にしたうえで責任の帰属モデルを構築する必要がある．また，責任の帰属モデルを組み込んだエージェントの設計についても検討の余地が残されているといえる．また，本研究の実験ではインタラクションを行う状況・問題の変化や失敗からの挽回といったインタラクションが繰り返し行われる場合，文脈の考慮といった点が含まれておらず，ここにも検討の余地がある．

近年の傾向としてエージェントや端末の高性能化ばかりが注目され，インタラクションの形態が軽視されているように思える．本研究はいかに優秀，高性能なエージェントが設計されたとしてもユーザからは使えないという烙印を押されてしまう可能性を示している．また，インタラクションの設計次第で人を騙すようなエージェントが生まれる可能性がある．これらの点から，人-エージェント間インタラクションにおける責任の帰属モデルを明確にする必要があるといえる．

文献

[1] 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

[2] Tetsuya Matsui，Seiji Yamada“ Building Trust in PRVAs by User Inner State Transition through Agent State Transition”，HAI ’16 Proceedings of the Fourth International Conference on Human Agent Interaction P.111-114，2016

[3] 西垣悦代，浅井篤，大西基喜，福井次矢“日本人の医療に対する信頼と不信の構造：医師感謝関係を中心に”，対人社会心理学研究 4 p.11-p20，2004

[4] 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

[5] 熊谷太郎“寄与過失を伴う厳格責任は過失責任よりも優れた損害賠償責任ルールか”，松山大学論集 第 20 巻 第 6 号，2009

[6] 森純一郎，Helmut Prendinger，土肥浩，石塚満“ユーザ感情に基づくエージェントの感性的インタラクション”，知能と複雑系 130-11，2002

[7] 藤井聡，”合意形成問題における”計画修正可能性”と”謝罪”の決定的役割”，運輸政策研究 Vol7 No.3，2004

[8] 大渕憲一, 渥美恵美, 山本雄大“謝罪による信頼回復：その逆転効果をめぐる検討”文化 80 巻 1，2 号，2016

[9] James Reason”Human Error”，Cambridge University Press，1990

[10] 古城和敬“成功・失敗の原因帰属に及ぼす public esteem の効果”，実験社会心理学研究 第 20 巻 第 1 号，1980

[11] 小俣謙二”犯罪化会社への責任帰属に関連する心理的要因の検討-傷害致死事件の場合-”，駿河台大学論叢 第 40 号，2010