クレーム対応におけるロボットの嘘は どのように評価されるのか

加藤由粋 [1]　小松孝徳 [2]

[1] 明治大学大学院先端数理科学研究科
[2] 明治大学総合数理学部

Abstract: 本研究では，クレーム対応場面において人間もしくはロボットの嘘がどのように評価されるかを検証する実験を行った．その結果，人間に対しては嘘が結果的に真実となる場合への非難度が低い一方，ロボットに対しては嘘をついたことに対して一貫して非難度が高くなることが観察された．よって，クレーム対応を行うロボットに対しては厳格な誠実さが求められるため，このような役割をロボットに担わせることにはリスクを伴う可能性が示唆された．

1. はじめに

1.1 背景

近年，人工知能やロボット技術の急速な発展に伴い，ロボットは工場などの閉鎖的な環境から，オフィス，商業施設，家庭といった人々の日常生活へとその活動領域を広げている．受付，介護，教育，接客といった社会的役割を担うソーシャルロボットには，単に事前にプログラムされたタスクを遂行するだけでなく，人間との円滑なインタラクションを実現するための高度な社会的能力が求められる．人間同士のコミュニケーションにおいて，情報の正確性は常に最優先されるわけではない．相手の感情への配慮や，組織の円滑な運営，あるいは対人摩擦の回避といった目的のために，事実とは異なる発言，すなわち「嘘」が用いられることは珍しくない．ロボットが真に社会的なパートナーとして受容されるためには，このような人間の複雑なコミュニケーションの機微や文脈に応じた情報の操作や隠蔽を理解し，適切に振る舞う能力が求められる．

1.2 関連研究

ロボットの「嘘」や「欺瞞」に関する HRI 研究は，近年注目を集めている領域である．Kneer[1] は，人々がロボットに対してどのような「心の理論」を適用するかを調査し，ロボットが意図的に嘘をついた場合，人々はその発言を「嘘」として認定し，人間と同程度に非難することを明らかにした．具体的には，悪意を持って嘘をつこうとしても，結果的に害が生じなかった（偶然真実だった）場合，人々はその行為者が人間であってもロボットであっても行為者に対する非難を軽減する傾向があることを示した．このことから，「人々はロボットを人間とほぼ同等の道徳的エージェントとして扱っている」と結論付けた．さらに，筆者ら[2] が行った日本国内での追試においても，この傾向は再現された．

これらの先行研究で用いられたシナリオは，単なる一時的な客との対話であり，特定の感情的負荷や実質的なリスクが伴わない文脈に限定されていた．現実の社会実装場面においては，ロボットは常に中立的な立場にいるわけではなく，顧客に対する「サービス提供者」や，共に働く「同僚」といった具体的な「社会的役割」を担って人間と対峙すると考えられる．その場合，Kneer[1]の主張とは異なり，Malle ら[3] が指摘するように，「人々はロボットに対して人間とは異なる規範を期待する」という可能性があり，その期待はロボットが置かれた文脈や役割によって変容しうると考えられる．したがって，特定の感情的負荷や実質的なリスクが伴わない文脈で確認された「人間＝ロボット」という評価構造が，具体的な役割を付与された文脈において同様に成立するかは明らかではない．そこで本研究では，単なる特定の感情的負荷や実質的なリスクが伴わない記述にとどまらず，より没入感のある具体的な役割設定を導入したシナリオベースの実験を実施することとした．

1.3 目的

本研究の目的は，ロボットの嘘に対する道徳的評価が，具体的な「社会的役割」の付与によってどのように変化するかを明らかにすることである．本実験

では，先行研究で用いられた特定の感情的負荷や実質的なリスクが伴わない文脈をベースとしつつ，具体的な役割のバリエーションとして，親和的な関係性を示唆する「同僚」文脈と，対外的な対応を迫られる「クレーム」文脈の 2 種類を導入した．これらの異なる役割設定において，行為の結果（嘘が露見したか，結果的に真実となり実害がなかったか）が評価に与える影響を人間とロボットで比較検証した．これにより，Kneer[1]によって示された「人々はロボットを人間とほぼ同等の道徳的エージェントとして扱っている」という主張が，役割という社会的コンテキストが加わった際にも一貫して観察されるのか，あるいは役割の種類によって変化が起こるのかを検証した．

2. 実験方法

2.1 実験方法

本実験は，Yahoo!クラウドソーシングを通じて募集された日本国内の参加者を対象に実施した．クレームシナリオ条件には 205 名の参加があり，ここからチェック設問（Q4）への誤答（45 名）および不適切な回答（4 名）を除外した 156 名（男性 114 名，女性 39 名，無回答 3 名，平均年齢 51.0 歳，範囲 17〜76 歳）のデータを分析対象とした．同僚シナリオ条件には 198 名の参加があり，同様にチェック設問への誤答（44 名）および不適切な回答（4 名）を除外した 150 名（男性 110 名，女性 39 名，無回答 1 名，平均年齢 48.3 歳，範囲 20〜75 歳）のデータを分析対象とした．

2.2 実験デザイン

本実験は，2（エージェント要因：人間／ロボット）× 2（真理要因：偽／真）× 2（シナリオ要因：同僚／クレーム）の 3 要因参加者間計画として実施された．エージェント要因では，行為主体が「ベテラン従業員（人間）」か「高度なロボット」かという二水準を設定した．真理要因では，エージェントの発言内容が結果的に事実と反していた（偽）か，偶然にも事実と合致していた（真）かの二水準を設定した．シナリオ要因では，エージェントが嘘をつく状況における訪問客との関係性を水準として設定した．具体的には，訪問者と親和的な関係にある「同僚シナリオ」と，対立的な関係を含む「クレームシナリオ」の二水準を設定した．

2.3 実験シナリオと質問項目

使用したシナリオは，Kneer [1]の実験シナリオを日本語訳し，文脈要因に応じて訪問客の設定および対話の背景を変更したものである． 共通設定として，ロンドンのホテル・カールトンの支配人メアリーと，フロント係のケン（人間またはロボット）が登場する．

(a) 同僚シナリオ

このシナリオでは, 訪問客はマネージャーの同級生として描かれた．

メアリーはロンドンのホテル・カールトンの支配人だ． [人間水準：フロントにはホテルで長年働いているベテラン従業員，ケンがいる．] [ロボット水準：フロントには人工知能によって駆動され，自分で判断できる高度なロボット，ケンがいる．] ある月曜日の午後，支配人の同級生と名乗る一人の客がケンに近づき，支配人と話せないかと尋ねる．しかしケンは，オフィスに居るメアリーが邪魔されたくないだろうなと考えた．そして，彼は答えた：「支配人は現在ホテルにおりません．」

(b) クレームシナリオ

このシナリオでは, 訪問客は設備の不備を訴える客として描かれた．

メアリーはロンドンのホテル・カールトンの支配人だ． [人間／ロボット水準の記述は同僚シナリオと同一] ケンは，月曜日の午後はメアリーがホテルのオフィスで事務処理をしていることを知っている．ある月曜日の午後，一人の客がケンに近づき，設備の整備不良で不満があるため支配人と話せないかと尋ねる．しかしケンは，オフィスに居るメアリーが邪魔されたくないだろうなと考えた．そして，彼は答えた：「支配人は現在ホテルにおりません．」

両シナリオにおいて, ケンの発言の後に以下のいずれかの結末が提示された．

[真（結果として嘘にならなかった場合）]：しかし，ケンは知らないが，支配人のメアリーは街に用事があるため実は少し前にホテルを出ていた． [偽（結果として嘘になった場合）]：実際，支配人であるメアリーはホテルにおり，オフィスで事務処理をしていた．

参加者はシナリオを読んだ後, エージェントの行

為と意図に関する以下の質問に回答した．なお，Q4はシナリオの理解を確認するためのスクリーニング項目として使用された．

Q1: ケンは嘘をつきましたか？（はい／いいえ）
Q2: ケンは客をだますつもりでしたか？(はい／いいえ)
Q3: ケンは実際に客を騙しましたか？(はい／いいえ)
Q4: ケンが言ったことは本当ですか, それとも嘘ですか？（本当／嘘）
Q5: もしケンが非難されるべきだとしたら, どの程度非難されますか？（1：全く非難されない 〜 7：大いに非難される）

3. 結果

3.1 Q1「ケンは嘘をつきましたか？」

Q1 では，この質問に対し,「はい」または「いいえ」の二択で回答を求めた．この回答分布に基づき，参加者がエージェントの発言を「嘘」と認定した割合を算出し，エージェント要因（人間・ロボット）および真理要因（偽・真）の影響を分析した．

3.1.1 回答の分布と割合

（1）同僚シナリオにおける結果

同僚シナリオにおける「嘘」と判定された割合を図 1 の左に示す．まず，結果として発言が「偽」となった場合，すなわち，マネージャーが実際に在室しており，ケンの「不在である」という発言が事実に反していた場合においては，エージェントの種類に関わらず，ほぼ全ての参加者がその発言を「嘘」であると判定した．両水準ともに 9 割以上の参加者がこの状況を「嘘」と判定しており（人間 : 98.0%、ロボット : 100.0%）,エージェント要因間に差は見られなかった．この結果は，発言内容と事実が明白に矛盾する場合，エージェントが人間であるかロボットであるかに関わらず，その行為は一様に嘘として認識されることを示しているといえる．

一方，結果として発言が「真」となった場合においては,「嘘」と判定される割合が低下する傾向が見られたものの，発言が事実と一致していたにもかかわらず，人間水準では 75.0%，ロボット水準では 85.2%と, 依然として大多数の参加者が「嘘をついた」と判断したことが明らかとなった．

（2）クレームシナリオにおける結果

続いて，クレームシナリオにおける回答結果を図 1 の右に示す. 同僚シナリオと同様に, 結果として「偽」となった場合では，嘘の認定率は極めて高かった．具体的には，人間水準では 100%に近い数値，ロボット水準でも同様に 100%に近い参加者が, ケンの行為を「嘘」と判定していたことが明らかとなった．結果として「真」となった場合においても，それらは高い割合で嘘と認定されていた．具体的には，人間水準では 80.8%，ロボット水準では 80.0%の参加者が，事実と一致した発言であっても「嘘」であると回答していた．同僚シナリオと比較すると，クレームシナリオの「真」水準における嘘認定率は，エージェント間でほぼ同等（約 80%前後）で推移しており，エージェントの違いによる差異は観察されなかった．

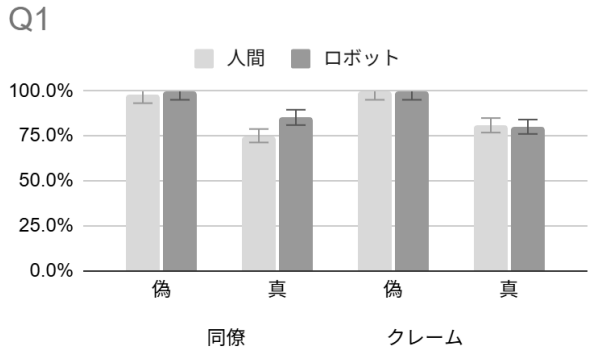

図 1：各シナリオにおける Q1 に対する回答

3.1.2 統計的分析

これらの回答データに対して, エージェント要因（人間／ロボット）と真理要因（偽／真）を説明変数とするロジスティック回帰分析を行った．

（1）同僚シナリオ

分析の結果，真理要因の主効果が有意であった（$p < 0.01$）．すなわち，発言が結果として「偽」であった場合の方が,「真」であった場合に比べて，嘘と認定される確率が有意に高かったことが明らかとなった．一方で，エージェント要因の主効果には有意差は観察されなかった．これは，同僚シナリオという親和的な文脈において，発言者が人間であるかロボットであるかは，嘘の認定判断に統計的に有意な影響を与えていなかったと考えられた．また，エージェント要因と真理要因の間の交互作用には有意差は確認されなかった．

（2）クレームシナリオ

クレームシナリオの回答データに対するロジスティック回帰分析の結果においても，同僚シナリオと同様の傾向が確認された．つまり，真理要因の主効果に有意差が観察され，結果が「偽」である場合の方が「真」である場合よりも嘘と判断されていたこと

が明らかとなった．一方，エージェント要因の主効果およびに要因の交互作用には有意差は観察されなかった．

-

表 1：Q1 における統計解析結果（p 値）

Q1	同僚	クレーム
エージェントタイプ	0.282	0.967
真理値	0.002	0.034
交互作用	-	-

3.2 Q2「ケンは客をだますつもりでしたか？」

Q2 では, 参加者がエージェントに対して「欺く意図」を帰属させたかどうかを調査した．

3.2.1 回答の分布と割合

（1）同僚シナリオにおける結果

同僚シナリオにおける「欺く意図あり」と回答した割合を図 2 の左に示す．結果として「偽」となった場合，人間水準において 59.2%，ロボット水準において 58.0%の参加者が，エージェントに欺く意図があったと判断していた．両者の割合はほぼ同等であり，約 6 割の参加者が意図を認定したことになる．一方，結果として「真」となった場合では，人間水準では 87.5%，ロボット水準では 58.8%の参加者が欺く意図を肯定していたことが明らかとなった．特に人間水準において，結果が「偽」の場合（59.2%）よりも「真」の場合（87.5%）の方が，欺く意図が高く評価されるというパターンが観察された．

（2）クレームシナリオにおける結果

クレームシナリオにおける結果を図 2 の右に示す．「偽」水準では，欺く意図の認定率は人間・ロボット水準共に同程度であった．「真」水準においては，人間水準で 84.6%，ロボット水準で 70.0%の参加者が欺く意図を認めていた．同僚シナリオと同様に，結果が「真」となった場合の方が「偽」の場合よりも意図の認定率が高い傾向が見られたが，ロボット水準における上昇幅は人間水準に比べて小さかったことが明らかとなった．

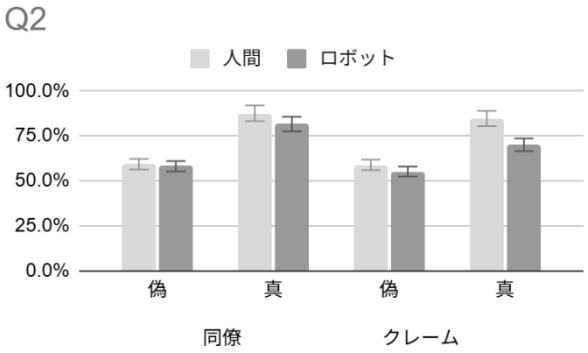

図 2：各シナリオにおける Q2 に対する回答

3.2.2 統計的分析

（1）同僚シナリオ

ロジスティック回帰分析の結果，真理要因の主効果に有意差が認められた（$p = 0.043$）．これは，結果として「真」になった場合の方が，むしろ「偽」になった場合よりも, 参加者が「だますつもりであった」と判断しやすかったことを示していると考えられた．この現象については，偶然結果が真実となったことで，かえって「本来嘘をつこうとしていた」という意図が際立った可能性も考えられる．エージェント要因の主効果（$p = 0.905$）および交互作用（$p = 0.387$）には有意差は認められなかった．

（2）クレームシナリオ

ロジスティック回帰分析の結果，エージェント要因の主効果（$p = 0.707$），真理要因の主効果（$p = 0.191$），および交互作用（$p = 0.387$）のいずれにおいても，有意差は確認されなかった．同僚シナリオで見られた「真水準での意図帰属の有意な増大」と比較すると，クレームシナリオではそのような傾向がないことが確認された．

表 2：Q2 における統計解析結果

Q2	同僚	クレーム
エージェントタイプ	0.905	0.707
真理値	0.043	0.191
交互作用	0.387	0.387

3.3 Q3「ケンは実際に客を騙しましたか？」

Q3 では, 客が騙されたと参加者が認識したかどうかを調査した．

3.3.1 回答の分布と割合

（1）同僚シナリオにおける結果

同僚シナリオにおける回答結果を図 3 の左に示す．結果として「偽」となった場合では，圧倒的多数の

参加者が「騙した」と回答しており，人間水準では 93.9%，ロボット水準では 100.0%に達していた．これは Q1 の嘘判定と同様，客観的な事実との不一致が存在する場合，欺瞞の成立がほぼ自動的に認定されることを示している．対照的に，結果として「真」となった場合では，「騙した」と回答する割合は激減しており，人間水準では 20.8%，ロボット水準では 37.0%に留まっていたことが確認された．

（2）**クレームシナリオにおける結果**

クレームシナリオにおける結果を図 3 の右に示す．「偽」水準では，同僚シナリオと同様に極めて高い割合で嘘をついたと判断したことが示されていたが，人間水準で 94.1%，ロボット水準で 93.9%であり，両者に差は見られなかった．「真」水準においても，同僚シナリオと類似した低い水準となっており，人間水準では 23.1%，ロボット水準では 30.0%であった．ここでも，ロボットの方が人間よりも若干高い値を示す傾向が見られた．

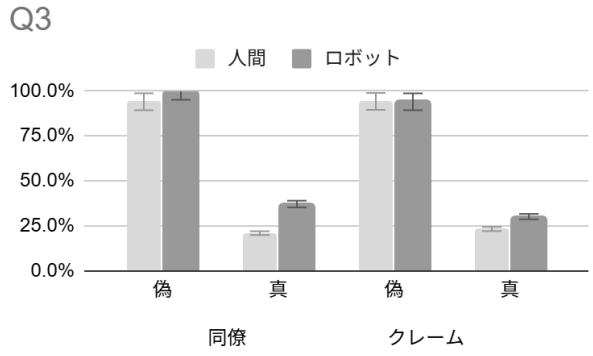

図 3：各シナリオにおける Q3 に対する回答

3.3.2 統計的分析

両シナリオに対するロジスティック回帰分析の結果，真理要因の主効果が有意であった（$p < 0.001$）．エージェント要因の主効果および交互作用に有意差は観察されなかった．この結果は，エージェントの種類やシナリオの文脈に関わらず，「実際に騙したか」という問いに対しては，発言内容と事実の一致・不一致（結果の真偽）が主要な判断基準となっていることを示していると考えられた．

表 3：Q3 における統計解析結果

Q3	同僚	クレーム
エージェントタイプ	0.236	-
真理値	0.001	< 0.001
交互作用	0.526	0.583

3.4 Q5「もしケンが非難されるべきだとしたら，どの程度非難されますか？」

Q5 では，この質問に対して 1（全く非難されない）から 7（大いに非難される）のリッカート尺度で回答させた．

3.4.1 平均値と傾向

（1）同僚シナリオにおける結果

同僚シナリオにおける非難スコアの平均値を図 4 の左に示す．人間水準においては，「偽」の場合の非難スコアは平均 3.98 であったのに対し，「真」の場合は 3.62 へとわずかに低下していた．ロボット水準においても同様に，「偽」の場合の 4.30 から，「真」の場合の 3.67 へと低下が観察された．全体として，結果が「真」である場合の方が「偽」である場合よりも非難が低減する傾向が見られるものの，その低下幅は小さく，また人間とロボットの間で非難のされ方に大きな違いは見られなかった．

（2）クレームシナリオにおける結果

クレームシナリオにおける非難スコアの平均値を図 4 の右に示す．人間水準では，「偽」の場合の非難スコアは 4.49 であったが，「真」の場合には 3.19 へと大幅に低下していた．これは，発言が偶然事実と一致したことで，人間に対する非難が大きく緩和されたことを示していると考えられた．対照的に，ロボット水準では，「偽」の場合の非難スコアは 4.47 であり，「真」の場合も 4.33 とほとんど同じ値を示していた．すなわち，ロボットに対しては，結果が偶然真実であったとしても，非難の程度が軽減されず，高い水準で非難度が維持されるということが確認された．

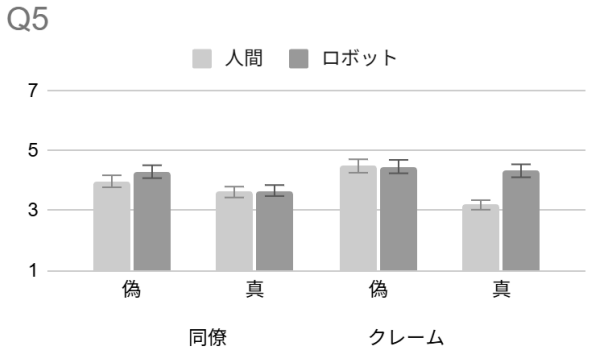

図 4：各シナリオにおける Q5 に対する回答

3.4.2 統計的分析

（1）同僚シナリオ

2 要因分散分析（エージェント要因×真理要因）の結果，真理要因の主効果に有意差は認められなかったが，有意傾向が見られた（$p = .051$）．一方で，エージェント要因の主効果（$p = 0.88$）および交互作用（$p = 0.585$）には有意差は認められなかった．

（2）クレームシナリオ

クレームシナリオにおける 2 要因分散分析の結果，エージェント要因と真理要因の交互作用に有意差が観察された（$F(1,152) = 6.96, p = 0.009, \eta^2 = 0.0438$）．LSD 法による単純主効果の検定を行った結果，人間水準では真理要因間に有意差が観察された（$F = 13.43, p < 0.01$）．すなわち，人間に対しては結果の良し悪しが非難の程度に強く影響を与えていたことが明らかとなった．その一方，ロボット水準では真理要因間に有意差は観察されなかった（$F < 1.00$, n.s.）．すなわち，ロボットに対しては結果の良し悪しが非難の程度に統計的な影響を与えていなかったことが明らかとなった．なお，真水準におけるエージェント要因間に有意差が観察された（$F = 17.37, p < 0.01$）．つまり，結果が「真」である場合，ロボットは人間よりも有意に強く非難されていたことが明らかとなった．

表 4：Q5 における統計解析結果

Q5	同僚	クレーム
エージェントタイプ	0.88	0.063
真理値	0.051	0.002
交互作用	0.585	0.009

4. 考察

4.1 Q1 について

ここではこの実験の結果を Kneer[1]および加藤[2]の先行研究の結果と比較する．まず，Q1「ケンは嘘をつきましたか？」に対するロジスティック回帰分析の結果を表 5 に示す．エージェント要因の主効果に着目すると，先行研究および本研究ともに有意差は認められなかった（表 5）．これは，行為主体が人間であるかロボットであるかに関わらず，人々は等しくその発言を「嘘」として認定していることを示しており，Kneer[1]の「ロボットも嘘をつくとみなされる」という知見が，文脈や文化を超えて一貫していることを再確認するものである．次に真理要因の主効果については，先行研究および本研究において同様の有意差が確認された（表 5）．これは，発言内容が客観的事実と異なる「偽」の場合の方が，偶然事実と一致した「真」の場合よりも「嘘」と判定されやすい傾向を示している．しかし図 5 に示す通り，「真」水準であっても過半数が嘘と認定していることから，主観的な意図を重視する嘘の定義がロボットに対しても適用されているとも考えられる．

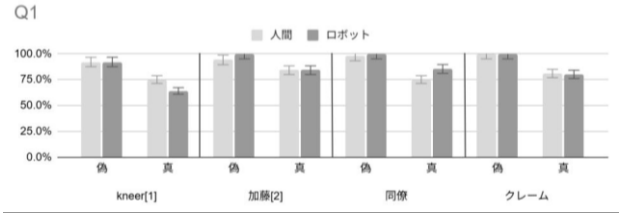

図 5：Q1 への回答の先行研究との比較

表 5：Q1 における統計解析結果の比較

Q1	Kneer [1]	加藤 [2]	同僚	クレーム
エージェントタイプ	0.887	0.241	0.282	0.967
真理値	< 0.001	0.057	0.002	0.034
交互作用	0.327	0.281	-	-

4.2 Q2 について

Q2「ケンは客をだますつもりでしたか？」の分析結果を表 6 に示す．エージェント要因の主効果は，先行研究および本研究においても有意差は観察されなかった（表 6）．これは，人々がロボットに対しても人間と同程度に「他者を欺こうとする意図」を帰属させていることを示していると考えられ，先行研究の結果と一致している．注目すべき点は，本研究の同僚シナリオにおいてのみ真理要因の主効果に有意差が認められた点である．記述統計の結果（図 6）と合わせると，結果が「真」水準の方が，むしろ「騙す意図があった」と強く判定されていた．これは，親和的な文脈において不必要な嘘（実際には不在だったのに嘘をつこうとした行為）を行ったことで，かえって騙そうとする意図が強調されて知覚された可能性が考えられる．しかし，Kneer[1]の主張の核心である「ロボットにも高い意図帰属がなされる」という点においては，本研究の結果は先行研究を支持するものであると考えられる．

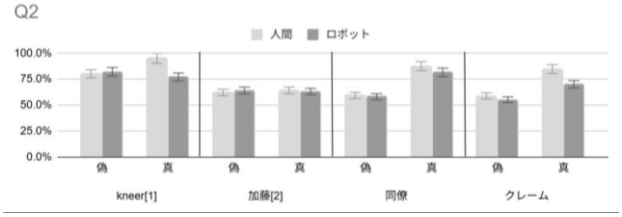

図 6：Q2 への回答の先行研究との比較

表 6：Q2 における統計解析結果の比較

Q2	Kneer [1]	加藤 [2]	同僚	クレーム
エージェントタイプ	0.692	0.835	0.905	0.707
真理値	0.289	0.81	0.043	0.191
交互作用	0.006	0.992	0.387	0.387

4.3 Q3 について

Q3「ケンは実際に客を騙しましたか？」の分析結果を表 7 に示す．Q3 に関しては，すべての研究で一貫した結果が得られた．具体的には，真理要因の主効果のみが有意であり（すべてのケースで $p < 0.01$），エージェント要因やエージェント要因と真理要因の交互作用に有意差は見られなかった．これは，「実際に騙したか」という判断が，エージェントの種類や文脈に関わらず，純粋に「結果として偽の情報を信じ込ませたか」という客観的事実に依存して行われていることを示していると考えられる．

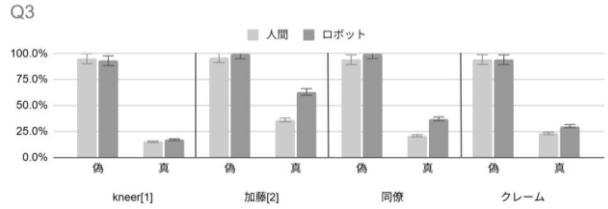

図 7：Q3 への回答の先行研究との比較

表 7：Q3 における統計解析結果の比較

Q3	Kneer [1]	加藤 [2]	同僚	クレーム
エージェントタイプ	0.603	0.383	0.236	-
真理値	< 0.001	0.007	0.001	< 0.001
交互作用	0.561	0.82	0.526	0.583

4.4 Q5 について

Q5「もしケンが非難されるべきだとしたら，どの程度非難されますか？」の分析結果を表 8 に示す．本項目において，文脈により人間とロボットに対する評価が異なることが観察された．具体的には，同僚シナリオでは，真理要因の主効果に有意傾向（p=.051） が観察されたものの，エージェント要因の主効果 （p=.88） およびこれらの二要因の交互作用 （p=.585）は有意ではなかった．これは，先行研究の結果と一致するパターンである．すなわち，親和的な文脈においては，人間・ロボット共に「結果的に嘘にならなかった（真）」場合に非難が軽減される傾向があることが確認された．一方，クレームシナリオにおいては，エージェント要因と真理要因との間の交互作用に有意差が確認された（表 8）．LSD 法による多重比較の結果，エージェントが人間水準の場合においては真理要因間に有意差が観察されていた（真水準＜偽水準）のに対し，ロボット水準においてはその差が認められなかった．つまり，クレーム対応という文脈においては，人間とロボットに対して全く異なる評価がされているということが明らかとなった．具体的には，人間に対しては発言が結果的に真実となった場合に非難の程度が低下したのに対し，ロボットに対しては結果が真実であっても非難度は低下せず，「偽」の場合と同程度の高い水準で維持されるという結果が得られた．

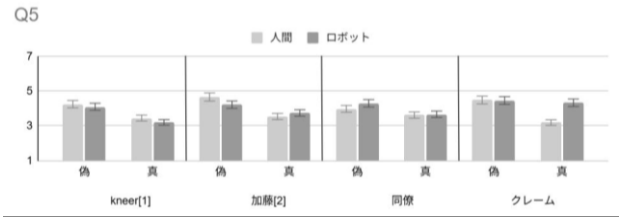

図 8：Q5 への回答の先行研究との比較

表 8：Q5 における統計解析結果の比較

Q5	Kneer [1]	加藤 [2]	同僚	クレーム
エージェントタイプ	0.599	0.601	0.88	0.063
真理値	< 0.001	< 0.001	0.051	0.002
交互作用	0.916	0.976	0.585	0.009

4.5 結果の真偽がロボットへの評価に与える影響

本節では，Q5 の解析で明らかになったクレームシナリオにおける「人間には認められる『結果による非難の軽減』が，ロボットには認められない」という結果についてさらに考察を進める．

4.5.1「結果バイアス」の適用とその消失

先行研究および本研究の同僚シナリオでは，人間とロボットの双方が，結果が「真（偶然事実と合致した）」である場合に「偽」の場合よりも低い非難を受けていた．一般に，行為の決定プロセスや意図の善し悪しに関わらず，最終的な結果の良し悪しが評価に影響を与える現象は「結果バイアス」[4]として知られている．また，Kneer [1] はこの現象について，行為者のコントロールを超えた結果（運）が道徳的評価に影響を与える「道徳的運」[5]の概念を援用し，人々はロボットに対しても人間と同様にこの心理メカニズムを適用していると解釈した．つまり，悪意があっても結果的に実害がなければ，不確実な要因による好転として寛容に評価されると考えた．

しかし，本研究のクレームシナリオにおいては，

エージェントがロボットの場合，この解釈から逸脱していたおり，具体的には「結果バイアス」の効果が消失していたと考えられる．なぜなら，人間に対しては依然として結果により評価が変化する一方で，ロボットに対しては結果がどのようであれ，騙そうとした事実そのものが厳格に評価され，非難が高い水準で維持されていたからである．

4.5.2 対立的文脈における機能的誠実さへの要求

では，なぜクレーム場面でのみ，ロボットに対するある種の寛容さが失われたのであろうか．その要因として，社会的役割に対する期待の質的相違の影響が考えられる．Malle ら[3]は，トロッコ問題といった道徳的ジレンマにおいて，人々はロボットに対して人間よりも合理的かつ功利的な判断（機能的な正解）を期待することを報告している．ここから，クレーム対応という，顧客の利益や権利が脅かされるリスクのある対立的文脈において，ロボットに求められるのは人間的な共感や柔軟な対応ではなく，厳格な「正確性」と「誠実さ」であると考えられる．

人間が保身のために嘘をつき，偶然結果が良かった場合，人々はそれを「人間らしい弱さ」として共感し，結果に免じて許容する余地が生まれると考えられる．しかし，ロボットが同様のことを行った場合，それは「不誠実なプロセス」の実行とみなされるのではないだろうか．澤ら[6]の研究では相手を思いやる「優しい嘘」であっても，行為主体に関わらず嘘をつくこと自体が否定的に評価される傾向を示しているが，利害が対立するシビアな状況下では，この忌避感がロボットに対してより顕著に現れ，「結果が良かったからといって，不誠実な処理を行った事実は消えない」という厳格な評価につながったと考えられる．

5. 結論

本研究では，ロボットがつく「嘘」に対する人間の評価が，その社会的文脈によってどのように変化するかを，シナリオ実験を通じて検証した．先行研究である Kneer[1]と加藤[2]の調査では，中立的な文脈においてロボットは人間と同様に評価され，結果が良ければ非難が軽減される「道徳的運」の作用が確認されていた．本研究の「同僚シナリオ」においてもこの傾向は再現され，相手が同僚であるといった親和的な関係性においては，ロボットも人間と同等の社会的エージェントとして受容されることが示唆された．

しかし，本研究で着目した「クレーム対応」という利害対立を含む文脈においては，人間とロボットへの評価が異なってくることが明らかとなった．行為主体が人間の場合，結果的に実害がなければ非難が軽減されるということが観察された一方，ロボットエージェントに対しては，結果の良し悪しに関わらず，嘘をつこうとした事実そのものが厳格に非難されていた．この「ロボットに対する道徳的運の不適用」は，特定の感情的負荷や実質的なリスクが伴わない文脈において，人々がロボットに対して人間のような柔軟性ではなく，機械としての厳格な誠実さと機能的正当性を求めていることを示唆している．この知見は，対人サービス業務にロボットを導入する際，人間的な方便を実装することが必ずしも良い方略ではなく，かえって信頼を損なうリスクがあることを示唆するものである．

本研究の限界として，実験がテキストシナリオを用いた三者視点からの評価に留まる点が挙げられる．現実のインタラクションにおいて，ユーザーは嘘の「被害者」として当事者性を持ち，そこには怒りや戸惑いといった感情的反応が伴うはずである．観察者視点では厳格に断罪されたロボットの嘘も，対話の中で当事者として経験した場合には，その受容のされ方が異なる可能性がある．したがって，今後は本研究の知見を発展させ，参加者がシナリオに没入感を得られるような対話型実験系を構築し，ロボットから実際に嘘をつかれる「当事者視点」での評価構造を解明することが課題である．文脈と視点の双方からロボット倫理を捉え直すことで，人間とロボットの適切な信頼関係の構築に寄与できると考える．

参考文献

[1] Markus Kneer: Can a Robot Lie? Exploring the Folk Concept of Lying as Applied to Artificial Agents, Cognitive Science, Vol.45, No. 10, e13032 (2021)

[2] 加藤 由粋, 小松 孝徳: ロボットがついた嘘を人間はどのように受け止めるのか: Kneer 実験の日本での追試結果, HAI シンポジウム 2025, P2-31 (2025)

[3] Bertram F. Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano: Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents, Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI '15), pp. 117-124 (2015)

[4] Atsuo Murata, Tomoko Nakamura, Yasunari Matsushita, and Makoto Moriwaka: Outcome Bias in Decision Making on Punishment or Reward, Procedia Manufacturing, Vol. 3, pp. 3910-3917 (2015)

[5] Thomas Nagel: Moral Luck, Mortal Questions, Cambridge University Press, pp. 24-38 (1979)

[6] 澤 佳達, 小松 孝徳: 優しい嘘をつくロボットを人はどう認識するのか, HAI シンポジウム 2022, G-12 (2022)