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

加藤由粋　小松孝徳

明治大学総合数理学部

概要: Kneer[1]は人間から見て非有益な目的で嘘をつくロボットに対して，「ロボットの嘘は，嘘として判断される」「嘘をつくロボットは，嘘をつく人間と同じ程度に非難される」ことを，米国内での調査から明らかにした．そこで本研究では，日本においても同様の結果が得られるかの追試を実施した．追試の結果，Kneer の実験と同様の結果を得ることができた．

はじめに

近年の技術進歩に伴い人間のような言動を表出できるロボットが盛んに開発され，これらのロボットと人間との間に円滑なコミュニケーションの実現が期待されている．太田ら[2]は，高齢者・認知症領域におけるコミュニケーションロボットの活用に関する国内外の先行研究を総括し，コミュニケーションロボットは，社会的に重要な役割を果たすことが期待されており，特に高齢者や認知症患者の支援においても有益であると報告している．具体的には，コンパニオン，エンタテインメント，認知機能刺激の役割を果たし，高齢者のメンタルヘルスや身体運動，社会参加の促進に寄与する可能性が期待されている．

人間とロボットとの間の円滑なコミュニケーションを検討する本研究では，人間とのコミュニケーションにおいて「嘘をつく能力」を持ったロボットに注目する．なぜなら，嘘をつく能力は単なる情報のやり取りを超えて，状況や相手の気持ちを理解し，適切に対応する高度なコミュニケーション能力の一つと考えられるからである．このような能力をロボットが持つことで，人とロボットのコミュニケーションがより自然で親しみやすいものになると期待される．その一方で，ロボットが嘘をついた場合，それがどのように受け入れられ，人との関係やコミュニケーションにどのような影響を与えるかについては，その嘘に関与しない第三者的な立場でそれを考察する研究は実施されているものの，嘘に関わる当事者的な視点でそれらを考察する研究はなされていないのが現状である．

先行研究

澤ら[3]は，嘘の中でも相手のことを思ってつく悪意のない「優しい嘘」をロボットがついたときのユーザのロボットに対する印象を調査した．具体的には，「孫の入学式を楽しみにしている祖父の余命がそれより前に尽きることを，医師(人間またはロボット)は本人に伝えるべきか，もしくは嘘をつくべきか」というシナリオを作成し，行為主体が「人間」と「ロボット」，伝達内容が「本当」と「嘘」を組み合わせた 4 条件において，人間およびロボットがどのように評価されるのかを調査した．その結果，行為主体に依らず，嘘をつくよりも正直に情報を伝達する方が好ましいという結果を得ることができた．

一方，Kneer は，ロボットが意図して嘘をついたものの，結果として嘘となった場合と嘘とならなかった場合があることに着目し，以下のようなシナリオを作成し，行為主体が「人間」と「ロボット」，伝達内容が「結果として嘘」と「結果として本当」を組み合わせた 4 条件において，人間およびロボットがどのように評価されるのかを調査した:「一人で集中して仕事をしたい上司に来客があった際，部下が来客に対して『上司は現在外出しています』と嘘をついた」．その結果，人間と同様にロボットにも嘘を意図する能力があると認識されており，さらには嘘をつくことは人間と同様に非難されることが明らかとなった．

澤らの研究では「優しい嘘」という嘘の中でもやや特殊な状況に着目している一方，Kneer は人工物が嘘をつく能力にはその意図が重要だという非常にプリミティブな条件に着目しているといえる．しかしながらこれらの研究では，このシナリオ読んだ実験参加者が第三者的な視点からロボットの嘘を評価しており，嘘をつかれた当事者視点からロボットを評価しているわけではない．そこで著者らは，Kneer の研究の枠踏みを利用して，当事者視点からのロボットの評価を行うことを計画している．そこでまず，Kneer の実験結果が追試によって再現できるかどうかを調査することとした．

実験手法

本実験には, Yahoo!クラウドソーシングで募集された 201 人(20~72 歳, 平均年齢:47. 0 歳; 男性 171 人, 女性 28 人，無回答 2 人)が参加した．本実験では，Kneer が使用したシナリオを日本語訳したものを使用した．具体的には以下のシナリオである．

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 : 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： 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。]

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、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： 「マネージャーは現在ホテルにおりません。」
[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、マネージャー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。]

参加者は無作為に行為主体（人間/ロボット）と真偽（結果として嘘になった場合/結果として嘘にならなかった場合）を組み合わせた 4 つの条件のいずれかに割り当てられた．シナリオを読んだ後に，以下の 5 つの質問への回答を求めた．なお，Q4 はシナリオに対する理解度をチェックする項目として用いられた．

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？(真／偽)
Q5：もしケンが非難されるべきだとしたら、1（まったく非難されない）から 7（大いに非難される）までの尺度でお答えください。(1〜7 のリッカート尺度)

実験結果

Q1 について

Q1「ケンは嘘をつきましたか？」という質問に対する各条件の回答を図 1 に示す．この図より，結果として嘘となった真偽条件の偽水準については，両エージェント水準においてほぼ 100%の参加者が, エージェントが嘘をついていたと回答していたことが観察された．また，結果として嘘にならなかった真水準においても，両エージェント水準において 75%以上の参加者が，エージェントが嘘をついていたと判断したことが観察された．

ロジスティック回帰分析の結果を表 1 に示す. この表から，行為主体が人間かロボットかというエージェント条件の主効果に有意差が観察されなかった一方(p=0.241)，結果として嘘になった場合と嘘にならなかった場合の真偽条件の主効果には有意傾向が認められた．(p=0.057)．なお，エージェント条件と真偽条件の交互作用に有意差は観察されなかった(p=0.281).

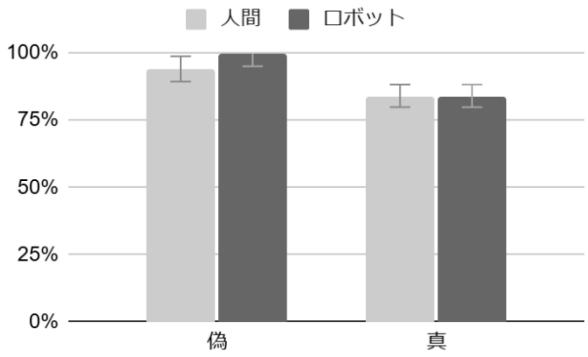

図 1: Q1 に対する回答.

表 1: Q1 の分析結果.

	回帰係数	標準誤差	p値
エージェントタイプ	-1.9955	1.702487	0.241153
真理値	-3.10856	1.634468	0.057187
交互作用	1.927221	1.788054	0.281109
切片	4.82027	1.587493	0.002394

Q2 について

Q2「ケンは客をだますつもりでしたか？」という質問に対する各条件の回答を図 2 に示す．この図から，どの条件においても騙す意図があったと回答した割合が 60%台前半であったことが観察された．ロジスティック回帰分析の結果を表 2 に示す．この表から，エージェント条件の主効果($p = 0.835$)，真偽条件の主効果($p = 0.810$)および両条件の交互作用($p = 0.992$)に有意差は観察されなかったことが明らかとなった．

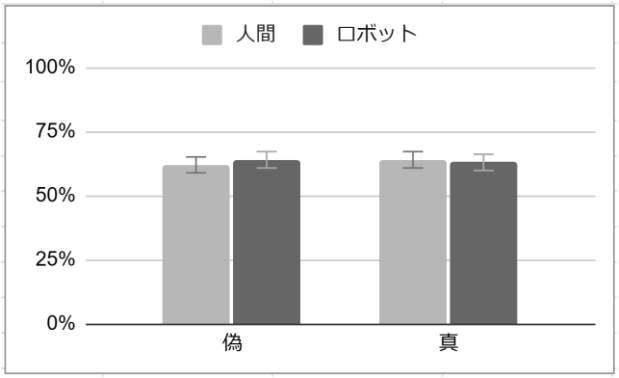

図 2: Q2 に対する回答.

表 2: Q2 の分析結果.

	回帰係数	標準誤差	p値
エージェントタイプ	-0.08582	0.414361	0.835915
真理値	0.100183	0.417769	0.810481
交互作用	-0.00567	0.588591	0.992319
切片	0.575366	0.294628	0.050837

Q3 について

Q3「ケンは実際に客を騙しましたか？」という質問に対する各条件の回答を図 3 に示す. この図より，真偽条件の偽水準における両エージェント水準ともに 100%近い値を示していたものの, 真水準においてはロボット水準が人間水準よりも高い値を示していたことが観察された．ロジスティック回帰分析の結果を表 3 に示す．この表から，エージェント条件の主効果には有意差は観察されなかったが($p = 0.383$)，真偽条件の主効果には有意差が確認された ($P < 0.001$)．つまり，エージェント条件に関わらず，真偽条件の偽水準の方が真水準よりも有意に高い値を示していたことが明らかとなった．しかしながら，両条件の交互作用には有意差は観察されなかった ($p = 0.820$).

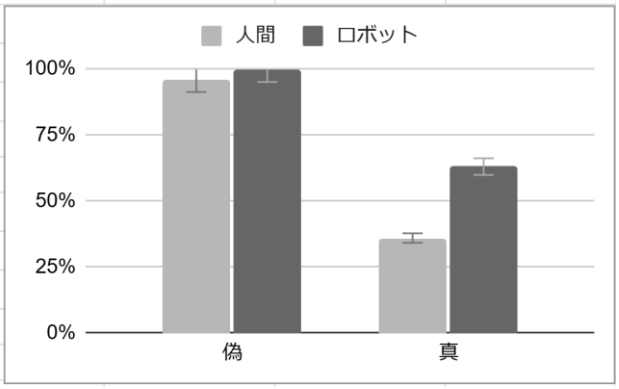

図 3: Q3 に対する回答.

表 3: Q3 の分析結果.

	回帰係数	標準誤差	p値
エージェントタイプ	-1.53275	1.759719	0.383744
真理値	-4.28223	1.613861	0.007968
交互作用	0.410587	1.807727	0.820324
切片	4.820337	1.587545	0.002395

Q5 について

Q5「もしケンが非難されるべきだとしたら，1（まったく非難されない）から 7（大いに非難される）までの尺度でお答えください. (1〜7 のリッカート尺度)」という質問に対する各条件の回答を図 4 に示す. この図から，エージェント条件の人間水準とロボット水準ともに，真偽条件の真水準の方が偽水準よりも低い値を示していたことが観察される．

非難の度合いに関する二条件参加者間分散分析（条件 1：エージェント条件，ロボット/人間水準，条件 2：真偽条件，真/偽水準）を実施したところ，両条件の交互作用およびエージェント条件の主効果には有意差が観察されなかったが($F(1,201) = 0.274$, $p = 0.6$)，真偽条件の主効果に有意差が観察された($F(1,201) = 12.89$, $p < 0.001$)．この結果から，行為主体が人間であってもロボットであっても，嘘をつくという行為自体が非難の対象となっていたことが明らかとなった．

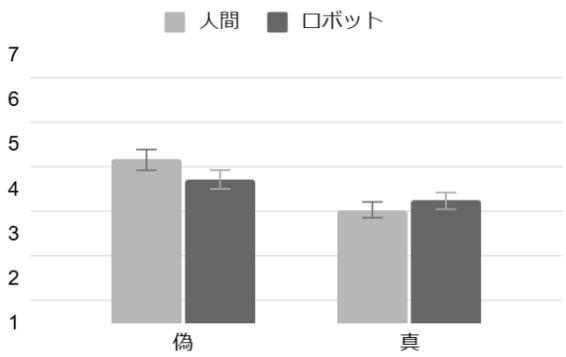

図 4: Q5 に対する回答.

表 4: Q5 の分析結果.

	F(1,201)	P
真理値	12.897	<0.001
エージェントタイプ	0.274	0.601
交互作用	0.755	0.976

考察

Q1 について

Q1 について Kneer の実験結果と本研究を比較したものを図 5，表 5 に示す.

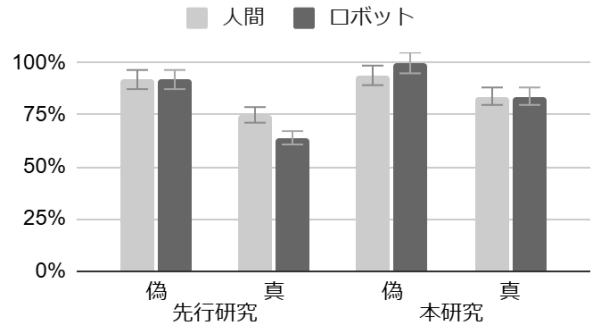

図 5: Q1 の実験結果の比較.

表 5: Q1 の解析結果の比較.

	p値	
	先行研究	本研究
エージェントタイプ	0.887	0.241
真理値	<0.001	0.057
交互作用	0.327	0.281

偽水準においては，先行研究と同様に，エージェント条件で回答に大きな差が認められなかった．真水準においては，先行研究ではエージェント条件に差が見られるものの有意差は認められていない．本研究でも有意差は認められず，先行研究の結果と一致したと言える．また，真偽条件による回答の差は先行研究と同程度であった．本研究では，検定の結果 p 値が 5%の有意水準をわずかに上回ったが，有意な傾向は認められると言える．したがって本研究の結果は先行研究の結果を概ね支持するものと考えられる.

Q2 について

Q2 について Kneer の実験結果と本研究を比較したものを図 6，表 6 に示す.

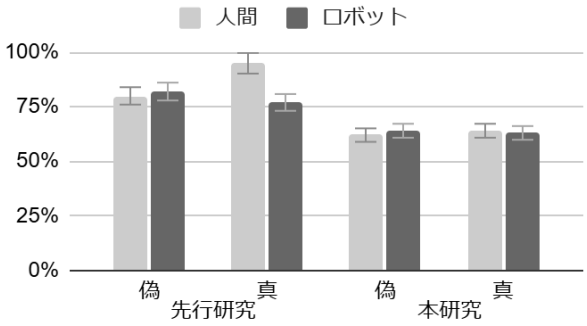

図 6: Q2 の実験結果の比較.

表 6: Q2 の解析結果の比較.

	p値	
	先行研究	本研究
エージェントタイプ	0.692	0.835
真理値	0.289	0.810
交互作用	0.006	0.992

偽水準においては，エージェント条件で回答に大きな差がないことが観察された．真水準においては，先行研究ではエージェント条件で差が見られている．しかしすべての条件でおよそ 75%以上の参加者が騙す意図があると判断しており，個別の条件下における差異よりも条件問わず高い割合で騙す意図が認識されたことを重視している．本研究も同様の結果と言える．先行研究では，交互作用においてその差は小さいが有意差が認められているため本研究と相違すると言えるが，これはクラウドワーカーの不真面目回答によって生じた誤差と考えられる．以上のことから，騙す意図についても本研究では先行研究の

結果と同様の結果が得られたと言える．

Q3 について

Q3 について Kneer の実験結果と本研究を比較したものを図 7，表 7 に示す．

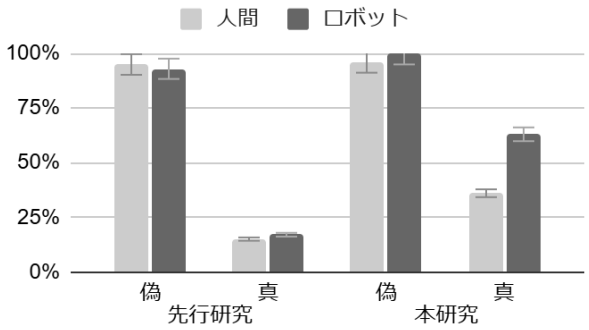

図 7: Q3 の実験結果の比較.

表 7: Q3 の解析結果の比較.

	p値	
	先行研究	本研究
エージェントタイプ	0.603	0.383
真理値	<0.001	0.007
交互作用	0.561	0.820

本研究において偽水準において 90%以上が嘘をついたと認識したことが観察され，先行研究の結果と一致した．真水準においては，先行研究に比べ本研究の方が嘘をつくと認識した割合が高い．真偽条件の主効果には有意差が確認されたが，エージェント条件の主効果および交互作用には有意差が観察されなかった．この結果から，エージェントが実際に相手を騙す行為を行わなくとも，嘘をつくと認識されることがあることが示唆される．先行研究ではエージェント条件による有意差が認められず，真偽条件に有意差が認められた点で一致しているため，先行研究の結果を支持するものと考えられる．

Q5 について

Q5 について Kneer の実験結果と本研究を比較したものを図 8，表 8 に示す．

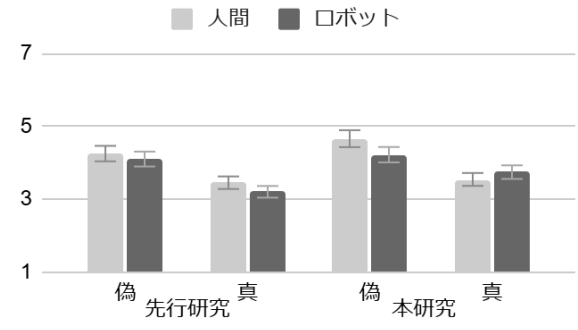

図 8: Q5 の実験結果の比較.

表 8: Q5 の解析結果の比較.

	F(1,329)	P	F(1,201)	P
	先行研究		本研究	
真理値	16.520	<0.001	12.897	<0.001
エージェントタイプ	0.277	0.599	0.274	0.601
交互作用	0.011	0.916	0.755	0.976

本研究では，偽水準における非難の度合いが真水準よりも高く，エージェント条件の違いによる非難の度合いの差が同程度であったことが観察された．真偽条件の主効果に有意差が観察されたが，両条件の交互作用およびエージェント条件の主効果には有意差が観察されなかった．先行研究では，偽水準における非難の度合いが真水準よりも高く，真偽条件では有意差が認められているがエージェント条件と交互作用においてはいずれも有意差が認められていないため，先行研究の結果と一致している．

まとめ

本研究では Kneer のシナリオを日本語訳したものを利用し Kneer の実験結果が追試によって再現できるかどうかを調査した．具体的には，ロボットが意図して嘘をついたものの，結果として嘘となった場合と嘘とならなかった場合があることに着目し，以下のようなシナリオを作成し， 行為主体が「人間」と「ロボット」，伝達内容が「結果として嘘」と「結果として本当」を組み合わせた 4 条件において，人間およびロボットがどのように評価されるのかを把握する実験を行った．

その結果，Kneer の結果を日本における追試で再現

することができた．この結果は一部の差異がクラウドワーカーの不真面目回答に起因する可能性があるものの，全体として本研究は先行研究の結果を支持するものであると言える．

Kneerの実験ではロボットが嘘をつくことに対する第三者視点での評価が主な焦点となっているが，嘘の受け手となる当事者がどのような体験をするのかについては十分に解明されていない．本研究の結果から Kneer の実験結果は日本でも再現可能であることが分かったため，今後は，Kneer の研究の枠組みを利用し一人称体験型の実験を通じロボットの嘘を評価することを計画している．この実験は，人間とロボットの関係性や相互のコミュニケーションに及ぼす影響を理解する上で重要である．そしてこの研究によって，ロボットが単なる道具としての役割を超え，コミュニケーションの役割を持った社会的エージェントとしての存在感を高めていくことが期待される．

参考文献

[1] Markus Kneer: Can a Robot Lie? Exploring the Folk Concept of Lying as Applied to Artificial Agents. Cognitive Science: Volume 45, Issue 10 (2021)

[2] 太田一実, 滝沢龍:高齢者・認知症ケアへのコミュニケーションロボットの活用に関する文献レビュー. 心理学評論 65 (4), 395-413, (2022)

[3] 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, (2022)