言語理解と説明がエージェントに対する信頼と共感に与える影響

How language comprehension and explanation affect trust and empathy toward agents

津村賢宏 [1] *　　Takahiro TSUMURA[1]

[1] 東洋大学
[1] Toyo University

Abstract: 人工エージェントは意思決定支援に用いられる一方で，利用者は課題内容や推論を十分に理解できない状況でも，エージェントを信頼するか判断しなければならない。本研究では，言語理解が制限された状況におけるエージェントへの信頼と共感の形成，および説明の影響を検討した。オンラインのクイズ実験（$N = 400$）において，課題言語（理解可能／理解不能）と説明の有無を操作し，インタラクション前後の信頼と共感を測定した。その結果，信頼は主に理解不能条件で増加し，説明によって強化されたが，正答率には影響しなかった。共感はいずれの条件でも増加したが，その変化は課題言語によって異なっていた。これらの結果は，意味的理解が制約される状況においても，社会的評価としての信頼と共感が形成されうることを示唆する。

1　はじめに

人工知能（AI）やロボットは，情報探索や推薦，タスク支援など，人の意思決定に直接関与する存在として日常的に用いられている。このようなエージェントに対する信頼は，助言を受け入れるか否かを左右し，システムの有効性や社会的受容性にとって重要な要因である。これまでの研究では，性能や信頼性，透明性，説明可能性，利用者特性などが信頼形成に影響することが示されてきた [1, 2, 3]。しかし，多くの研究は，利用者が課題内容やエージェントの説明を十分に理解できることを前提としている。

一方，現実の利用状況では，外国語環境や専門知識の不足，時間的制約などにより，利用者が課題内容や推論を十分に理解できない場合も少なくない。それにもかかわらず，利用者はエージェントを信頼するかどうかを判断しなければならない。こうした言語理解が制限された状況における信頼形成については，実証的研究が十分に行われていない。

説明は信頼を促進する手段として重視されてきたが，近年では，説明が必ずしも理解の向上を通じて信頼を高めるとは限らず，誠実さや配慮を示す社会的シグナルとして機能する可能性が指摘されている [4, 5, 6, 7]。また，Human-Agent Interaction (HAI) 研究では，信頼に加えて共感も重要な心理的要因とされており，人は人工的存在に対しても社会的・感情的反応を示すことが知られている [8]。しかし，意味理解を伴う認知的共感が，課題内容を理解できない状況でどのように形成されるのかは明らかでない。

本研究の目的は，課題言語の理解可能性と説明の有無が，エージェントに対する信頼と共感の形成にどのように影響するかを明らかにすることである。課題言語が理解可能か理解不能かを直接操作し，インタラクション前後の評価変化を測定することで，意味的理解が制約された状況においても生じる社会的評価のメカニズムを検討する。

2　関連研究

HAI において，信頼は意思決定支援の受容や依存行動を左右する中心的要因であり，性能や信頼性，透明性，説明可能性，利用者特性などが信頼形成に影響することが示されてきた [1, 2, 3]。一方で近年の研究は，信頼が必ずしも意味的理解や正確性評価のみに基づくものではなく，対話の流暢さや社会的フレーミング，説明の提示といったヒューリスティックな手がかりによって形成されることを示している [9, 5, 7]。特に，利用者が正誤を独立に検証できない状況では，説明は理解を高めるというよりも，誠実さや配慮を示す社会的シグナルとして機能する可能性が指摘されている [6]。

信頼と並び，共感もエージェントの受容や協調行動に関わる重要な要因である。メディア方程式理論が示すように，人は人工的存在に対しても社会的・感情的反応を示しうるが [8]，共感は情動的共感と認知的共感に区別

*連絡先：東洋大学情報連携学部
東京都北区赤羽台 1 丁目 7 − 11
E-mail：takahiro.tsumura@iniad.org

され，後者は意味理解や視点取得に依存する [10, 11]。HAI/HRI 分野では，関係的・インタラクション的な振る舞いを通じて，利用者がエージェントに共感を形成することが示されており，内部推論の理解が十分でない場合でも共感が生じうることが報告されている [12]。とくに Tsumura らは，言語的・関係的手がかりやインタラクション構造が，エージェントに対する信頼や共感の評価に影響することを一連の実証研究で示している [13, 14, 15, 16, 17, 18]。また，共感的表現は信頼を高め，失敗後の信頼回復にも寄与することが示されている [19, 20, 15]。

3　方法

3.1　仮説

本研究は，課題言語の理解可能性および説明の有無が，エージェントに対する信頼と共感の形成にどのように影響するかを検討することを目的とする。課題言語（日本語／ドイツ語）と説明の有無を操作し，以下の仮説を設定した。

H1 課題言語が理解不能な場合（ドイツ語），理解可能な場合（日本語）よりもエージェントへの信頼は高くなる。

H2 課題言語が理解不能な場合，エージェントへの共感は低くなる。

H3 エージェントが説明を提示することで，信頼は高まる。

H4 エージェントが説明を提示することで，共感は高まる。

3.2　実験手順

オンラインのクイズ課題を用いた実験を実施した。参加者はエージェントに対する事前評価を行った後，3 問の四択クイズに回答した。各問題では，エージェントが提示する回答を確認した後，参加者自身が回答を選択した。エージェントは 1 問目と 3 問目で正答，2 問目で誤答を提示した。図 1 はクイズタスクの一例である。

課題言語は日本語またはドイツ語とし，質問文および選択肢のみを操作した。エージェントの提示はすべて無音の動画とし，音声的手がかりの影響を排除した。全課題終了後，参加者は事後質問紙に回答した。

実験は，課題言語（日本語／ドイツ語）と説明（なし／あり）を参加者間要因，時間（事前／事後）を参加者内要因とする混合計画で実施した。

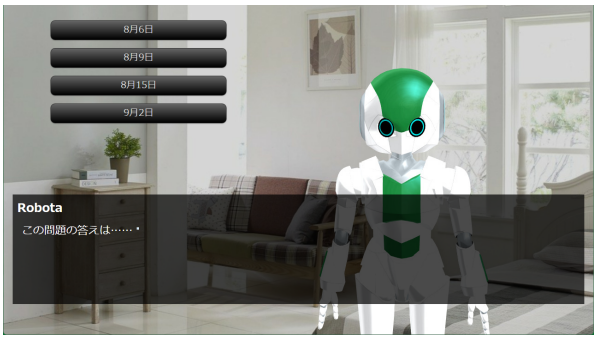


図 1: クイズタスクの一例

3.3　参加者

参加者は Yahoo!クラウドソーシングを通じて募集され，400 名が実験を完了した。平均年齢は 49.41 歳（$SD = 11.98$）で，男性 204 名、女性 196 名であった。参加者は事前にドイツ語の知識がないことを自己申告しており，条件に合致しない者は分析から除外した。

3.4　アンケート

エージェントに対する評価として，信頼および共感を測定した（Table 1）。信頼は認知的信頼および感情的信頼から構成され，7 件法で評価した。共感は情動的共感および認知的共感から構成され，5 件法で評価した。いずれの尺度も高い内的一貫性を示した（信頼：$\alpha = .93$–$.98$，共感：$\alpha = .75$–$.91$）。

3.5　要因とクイズ内容

説明条件では，エージェントが回答選択後に簡潔な説明文を提示した。説明は正答を示すものではなく，設問の主題を一般的に述べる内容とした。説明文およびその他の操作文はすべて日本語で提示され，選択肢の言語のみが課題言語条件によって異なった。説明なし条件では，回答提示のみを行った。

課題言語として，日本語（理解可能）またはドイツ語（理解不能）を操作した。質問文および選択肢のみを言語操作の対象とし，説明文や指示文は両条件で共通とした。これにより，意味理解の有無のみが評価に影響するよう統制した。

クイズの内容は以下の 3 つであった。**歴史** (日本のポツダム宣言受諾) **科学** (動物細胞には存在しない小器官)、**地理** (最大の国土面積を持つ国)。

各問題には 4 つの選択肢が表示され、歴史の問題については、第二次世界大戦終結に関連する 4 つの候補日が選択肢として選ばれました。科学の問題では、選択

表 1: 使用したアンケートの一覧

信頼			
認知的信頼			
Qt1: 信頼できるか？	Qt2: 予測できるか？	Qt3: 頼りになるか？	Qt4: 一貫しているか？
Qt5: 有能であるか？	Qt6: 熟練しているか？	Qt7: 能力があるか？	Qt8: 細心であるか？
感情的信頼			
Qt9: 安心するか？	Qt10: 快適であるか？	Qt11: 満足できるか？	
感情的共感			
個人的苦痛			
Qe1: もしキャラクターに非常事態が起こった場合，不安で落ち着かなくなる．			
Qe2: もしキャラクターが感情的になっていた場合，何をしたらいいか分からなくなる．			
Qe3: もし差し迫った助けが必要なキャラクターを見た場合，混乱してどうしたらいいかわからなくなる．			
共感的関心			
Qe4: もしキャラクターが困っているのを見た場合，気の毒に思わない．			
Qe5: もしキャラクターが他人にいいように利用されているのをみた場合，その相手を守ってあげたいような気持ちになる．			
Qe6: キャラクターの話や起こった出来事に心を強く動かされる．			
認知的共感			
視点取得			
Qe7: キャラクターの立場と人間の立場の両方に目を向ける．			
Qe8: もしキャラクターのことをよく知ろうとした場合，相手からどのように物事がみえているか想像する．			
Qe9: 自分が正しいと思った時に，キャラクターの言い分を聞かない．			
空想			
Qe10: キャラクターの話や起こった出来事に引き込まれることはなく，客観的である．			
Qe11: キャラクターに起こった出来事が自分の身に起こったらどんな気持ちになるだろうと想像する．			
Qe12: キャラクターの気持ちに深く入り込んでしまう．			

肢は 4 種類の細胞小器官で構成されていました。地理の問題については、4 つの国名の選択肢がありました。

3.6 分析方法

分析には 3 要因混合分散分析を用いた。参加者間要因は課題言語および説明，参加者内要因は時間であった。信頼および共感を従属変数とし，効果量には部分 η^2 を用いた。分析は R（ver. 4.1.0）で実施した。

4 結果

Table 2 に，各条件における信頼および共感の平均値と分散を示す。以下では，分散分析（Table 3）および有意な交互作用に対する単純主効果分析（Table 4）の結果を要約する。

4.1 信頼

信頼については，説明×時間，および課題言語×時間の交互作用が有意であった（Table 3）。単純主効果分析の結果，信頼はいずれの説明条件においても課題前後で有意に増加したが，その増加量は説明あり条件でより大きかった。課題前および課題後の時点をそれぞれ比較した場合，説明条件間に有意差は認められなかった。

また，課題言語との交互作用については，ドイツ語条件では課題前後で信頼が有意に増加した一方，日本語条件では有意な変化は認められなかった。この結果は，課題内容を理解できない状況において，エージェントに対する信頼がより顕著に形成されることを示している（Figure 2）。

4.2 共感

共感については，課題言語×時間の交互作用が有意であった（Table 3）。単純主効果分析の結果，共感はいずれの言語条件においても課題後に増加したが，その変化の大きさおよび推移は課題言語によって異なっていた。説明と時間の交互作用は有意ではなく，説明は共感の変化に対して限定的な影響しか示さなかった（Figure 3）。

5 議論

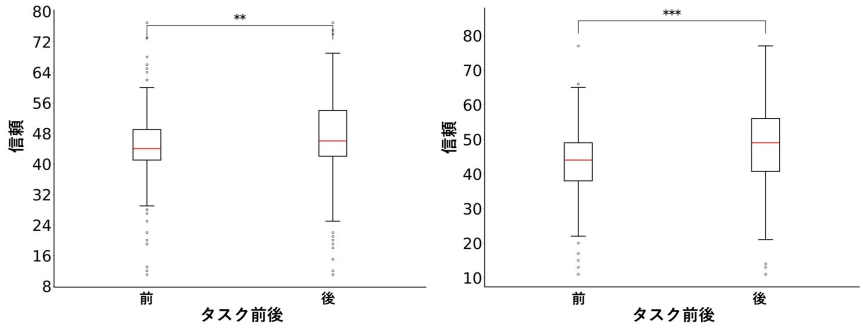
5.1 仮説に関する考察

本研究では，課題言語の理解可能性および説明の有無が，エージェントに対する信頼と共感の形成に与える影響について，4 つの仮説を検証した。

まず，**H1（課題言語と信頼）**に関して，課題言語と時間の交互作用が有意であり，信頼はドイツ語条件においてのみ課題前後で有意に増加した。この結果は，課

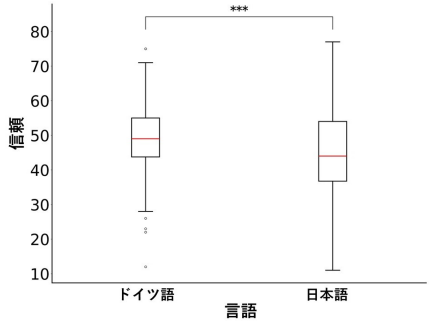
表 2: 本研究の統計情報

要因			信頼		共感	
説明	言語	前後	平均 (S.D.)	95% 信頼区間	平均 (S.D.)	95% 信頼区間
なし	ドイツ語	前	43.47 (8.575)	[42.15, 44.79]	37.81 (7.314)	[37.25, 38.37]
なし	ドイツ語	後	47.89 (10.28)	[46.57, 49.21]	38.11 (8.415)	[37.55, 38.67]
なし	日本語	前	44.43 (10.92)	[42.68, 46.18]	36.39 (6.757)	[35.80, 36.98]
なし	日本語	後	45.00 (14.64)	[43.25, 46.75]	35.01 (7.815)	[34.42, 35.60]
あり	ドイツ語	前	43.32 (8.681)	[41.68, 44.96]	36.98 (7.085)	[36.35, 37.61]
あり	ドイツ語	後	51.35 (10.69)	[49.71, 52.99]	38.46 (7.984)	[37.83, 39.09]
あり	日本語	前	42.21 (11.22)	[40.39, 44.03]	36.35 (6.180)	[35.69, 37.01]
あり	日本語	後	44.14 (13.60)	[42.32, 45.96]	34.95 (7.457)	[34.29, 35.61]

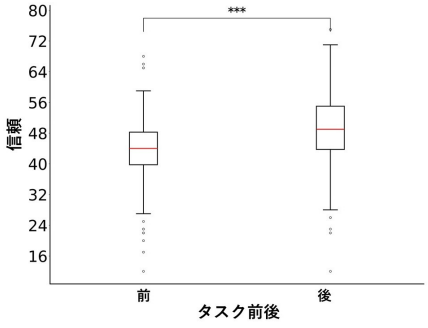


(a) 前後関係の説明なし条件

(b) 前後関係の説明あり条件

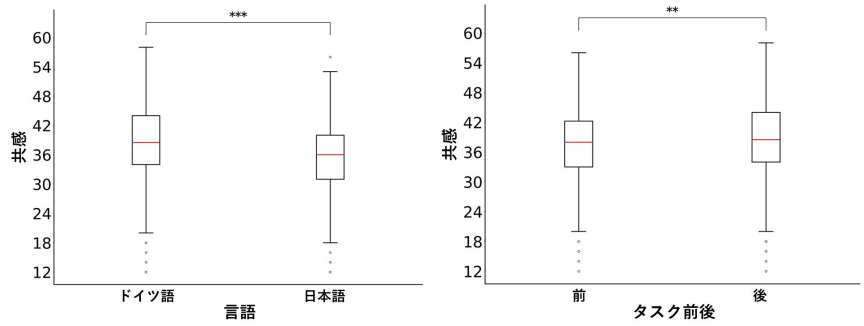


(c) 言語を用いたタスク後条件

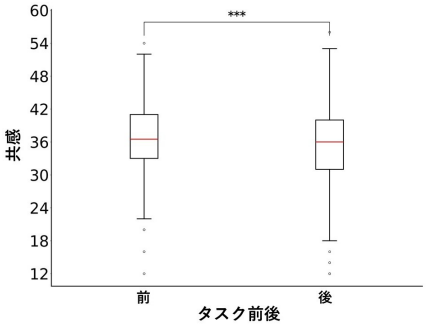


(d) ドイツ語の前後比較条件

図 2: 信頼条件の分布。赤い線は中央値、丸は外れ値を示す。



(a) 言語を用いたタスク後条件　　(b) ドイツ語の前後比較条件



(c) 日本語の前後比較条件

図 3: 共感条件の分布。赤い線は中央値、丸は外れ値を示す。

表 3: ANOVA の分析結果

	要因	F	p	η_p^2
信頼	説明	0.0036	0.9523 *ns*	0.0000
	言語	7.109	0.0080 **	0.0176
	前後	40.69	0.0000 ***	0.0932
	説明 × 言語	2.763	0.0973 *ns*	0.0069
	説明 × 前後	4.497	0.0346 *	0.0112
	言語 × 前後	18.02	0.0000 ***	0.0435
	説明 × 言語 × 前後	0.9216	0.3376 *ns*	0.0023
共感	説明	0.0420	0.8377 *ns*	0.0001
	言語	9.362	0.0024 **	0.0231
	前後	1.3079	0.2535 *ns*	0.0033
	説明 × 言語	0.0180	0.8933 *ns*	0.0000
	説明 × 前後	1.760	0.1854 *ns*	0.0044
	言語 × 前後	27.20	0.0000 ***	0.0643
	説明 × 言語 × 前後	1.883	0.1707 *ns*	0.0047

p: $^{**}p<0.01$ $^{***}p<0.001$

表 4: 単純主効果の分析結果

	要因	F	p	η_p^2
信頼 説明 × 前後	タスク前の説明	1.426	0.2332 *ns*	0.0036
	タスク後の説明	1.091	0.2969 *ns*	0.0027
	説明なしの前後比較	10.19	0.0016 **	0.0490
	説明ありの前後比較	32.52	0.0000 ***	0.1411
信頼 言語 × 前後	タスク前の言語	0.0057	0.9398 *ns*	0.0000
	タスク後の言語	16.46	0.0001 ***	0.0399
	ドイツ語の前後比較	68.82	0.0000 ***	0.2579
	日本語の前後比較	1.929	0.1665 *ns*	0.0096
共感 言語 × 前後	タスク前の言語	2.241	0.1352 *ns*	0.0056
	タスク後の言語	17.39	0.0000 ***	0.0421
	ドイツ語の前後比較	8.728	0.0035 **	0.0422
	日本語の前後比較	19.25	0.0000 ***	0.0886

p: $^{**}p<0.01$ $^{***}p<0.001$

題内容を理解できない状況において，エージェントへの信頼がより顕著に形成されることを示している。信頼が理解や正確性評価に基づくものだとする直観的な想定とは異なり，H1 は「課題言語が信頼形成に影響する」という点で支持された。

次に，**H2（課題言語と共感）**についても，課題言語と時間の交互作用が有意であった。共感は両条件で課題後に増加したものの，その推移は課題言語によって異なっており，意味理解の制約が共感形成の過程に影響することが示された。この結果は，共感が意味理解に依存する側面を持つという理論的整理と整合的であり，H2 は支持された。

一方，**H3（説明と信頼）**については，説明の主効果は認められなかった。説明は信頼の増加過程を調整する役割を果たしたものの，一般的な信頼向上要因としては機能しなかった。このため，説明が信頼を高めるとした H3 は支持されなかった。

同様に，**H4（説明と共感）**も支持されなかった。説明は共感の変化に有意な影響を与えず，短く抽象的な説明が共感を直接喚起する手がかりとしては十分でないことが示された。

5.2 言語理解が制約された状況における信頼と共感

仮説検証の結果を踏まえると，本研究は，信頼と共感が必ずしも意味的理解を前提として形成されるわけではないことを示している。特に信頼については，課題内容を理解できない状況においても，インタラクションを通じて増加することが確認された。これは，信頼が正確性評価ではなく，エージェントが意思決定に関与するという役割認識や，インタラクションの継続によって形成される可能性を示唆する。

共感についても，意味理解が制約される中で形成されるものの，その過程は課題言語に依存して変化することが示された。この結果は，共感が単一の心理過程ではなく，理解に基づく側面と，関係的・文脈的側面の双方から構成されることを示唆している。

5.3 設計的含意と MARCH の視点

本研究の結果は，説明を中心とした AI 設計観に対して重要な示唆を与える。説明は理解を保証する手段としてだけでなく，社会的評価を形成する文脈的手がかりの一つとして機能するに過ぎない。本研究では，説明が信頼や共感を直接的に高める要因とはならなかったことから，説明中心の設計が常に有効であるとは限らないことが示唆される。

このような理解が制約された状況を捉える概念的枠組みとして，本研究は MARCH（Machine Agent for Responsibility and Choice beyond Human explanation）を提示する。MARCH の視点からは，信頼や共感は説明の有無に依存するのではなく，理解不能な状況下における選択や役割分担の中で形成される社会的評価として捉えられる。

5.4 限界と今後の課題

本研究にはいくつかの限界がある。課題言語の理解可能性は二値的に操作されており，部分的理解の影響は検討されていない。また，低リスクな課題設定であったため，高リスク領域への一般化には慎重な検討が必要である。今後は，理解の程度が連続的に変化する状況や，より実践的な意思決定場面において，信頼と共感がどのように形成されるかを検討する必要がある。

6 まとめ

本研究は，課題言語の理解可能性および説明の有無が，エージェントに対する信頼と共感の形成にどのように影響するかを検討した。その結果，利用者が課題内容を意味的に理解できない状況においても，エージェントに対する信頼と共感が形成されうることが示された。特に信頼は，理解不能な条件においてインタラクションを通じて顕著に増加し，説明は正確性を高めることなく信頼の成長を補強した。一方，共感は言語条件に依存した変化を示し，意味理解の制約が共感形成の過程に影響することが示唆された。これらの結果は，エージェントに対する社会的評価が必ずしも意味的理解や正確性評価に基づくものではなく，インタラクション構造や文脈的手がかりによって形成されうることを示しており，理解が制約された現実的な HAI を捉える上で重要な示唆を与える。

参考文献

[1] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. PMID: 15151155.

[2] Kevin Hoff and Masooda Bashir. A theoretical model for trust in automated systems. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, page 115–120, New York, NY, USA, 2013. Association for Computing Machinery.

[3] Alexandra D. Kaplan, Theresa T. Kessler, J. Christopher Brill, and P. A. Hancock. Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, 65(2):337–359, 2023. PMID: 34048287.

[4] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[5] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. Bringing transparency design into

practice. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, IUI '18, page 211–223, New York, NY, USA, 2018. Association for Computing Machinery.

[6] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[7] Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: Informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA, 2020. Association for Computing Machinery.

[8] Byron Reeves and Clifford Nass. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press, USA, 1996.

[9] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[10] B. L. Omdahl. *Cognitive appraisal, emotion, and empathy*. Psychology Press, New York, 1 edition, 1995.

[11] Stephanie D. Preston and Frans B. M. de Waal. Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, 25(1):1–20, 2002.

[12] Ana Paiva. Empathy in social agents. *International Journal of Virtual Reality*, 10(1):1–4, 2011.

[13] Takahiro Tsumura and Seiji Yamada. Influence of agent ' s self-disclosure on human empathy. *PLOS ONE*, 18(5):1–24, 05 2023.

[14] Takahiro Tsumura and Seiji Yamada. Influence of anthropomorphic agent on human empathy through games. *IEEE Access*, 11:40412–40429, 2023.

[15] Takahiro Tsumura and Seiji Yamada. Making a human's trust repair for an agent in a series of tasks through the agent's empathic behavior. *Frontiers in Computer Science*, 6:1–17, 2024.

[16] Takahiro Tsumura and Seiji Yamada. Ikea effect and empathy for robots: Can assembly strengthen human-agent relationships? *PLOS ONE*, 20(7):1–25, 07 2025.

[17] Takahiro Tsumura and Seiji Yamada. The role of individual recognition in shaping empathy and trust toward an agent. *PLOS ONE*, 20(7):1–19, 07 2025.

[18] Takahiro Tsumura and Seiji Yamada. Shaping empathy and trust toward agents: The role of agent behavior modification and attitude. *IEEE Access*, 13:116908–116923, 2025.

[19] Deborah Johanson, Ho Seok Ahn, Rishab Goswami, Kazuki Saegusa, and Elizabeth Broadbent. The effects of healthcare robot empathy statements and head nodding on trust and satisfaction: A video study. *J. Hum.-Robot Interact.*, 12(1), feb 2023.

[20] Christopher Birmingham, Ashley Perez, and Maja Matarić. Perceptions of cognitive and affective empathetic statements by socially assistive robots. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '22, page 323–331. IEEE Press, 2022.