AI と人間の性格相性が信頼行動に与える影響：経済実験による検証*

平野 泰斗[a]　中津畑 航[b]　中山 遼人[c]　中島 愛衣[d]　小林 奈津美[e]

要約

本研究は，人間と AI の性格が一致したときに，AI への信頼行動がどのように変化するかを検証した．実験では，被験者の HEXACO 特性を測定し，AI の性格を類似・反対・統制の３条件で提示し，AI を想定した信頼ゲームを実施した．その結果，類似条件の被験者は反対条件よりも AI に多く投資し，性格の類似が信頼を高めることが示された．一方で，外向性の一致による効果は確認されなかった. また正直さ-謙虚さの高い人ほど AI に投資しやすく，また AI に肯定的な人ほど信頼行動を示す傾向が示された．これにより，人間が AI を「自分に近い存在」と感じると信頼が高まること，そして特定の性格特性や AI への態度が信頼の形成に寄与することが分かった. 今後は実際に HEXACO 特性を反映したペルソナ AI を構築し, 外的妥当性の高い検証を進める必要がある. 本研究は AI を単なる道具ではなく，信頼できるパートナーとして社会に浸透させるために，性格モデルを応用した設計の有効性を示す.

JEL 分類番号： D91, O33

キーワード：信頼ゲーム, AI への態度, HEXACO モデル，社会的アイデンティティ理論

*なお，本論文に関して，開示すべき利益相反関連事項はない.

[a] 同志社大学 cgfh0115@mail3.doshisha.ac.jp
[b] 同志社大学 cgfh0449@mail3.doshisha.ac.jp
[c] 同志社大学 cgfh0451@mail3.doshisha.ac.jp
[d] 同志社大学 cgfh0445@mail3.doshisha.ac.jp
[e] 同志社大学 cgfh0258@mail3.doshisha.ac.jp

1. イントロダクション

現代社会において，AI 技術は私たちの生活のあらゆる側面に深く浸透し，その重要性は日々増している．しかし，このような技術の社会への浸透には，ユーザーが AI に対して信頼を抱くことが不可欠である．この文脈において，AI スティグマと呼ばれる，AI の助言がユーザーの不信感によって無視される問題が指摘されており，AI 技術の普及を妨げる重大な障壁となっている(Jakob, 2024).

この課題を克服するため，心理学におけるパーソナリティ・モデルを AI に適用し，人間のような性格を与えることで，ユーザーとの間に信頼関係を構築する研究が増加している（Jiang et al., 2020）．人同士の性格の類似は，対人関係の満足度や信頼関係の構築に影響することが知られており（Leikas and Salmela-Aro, 2014），特に外向性の類似は，関係の安定に寄与することが示唆されている（Selfhout et al., 2010）．先行研究では，AI と人間の性格特性の類似が与える影響についても研究が進められている．例えば，Kuhail et al.（2024）は学術アドバイスの文脈で，外向的なユーザーが自身の性格に合ったチャットボットに対してより高い信頼とエンゲージメントを示すことを明らかにした．しかしこの研究には，サンプルサイズの小ささや状況の限定性といった課題が残されている．

そこで本研究では，先行研究の課題を克服するため，ビッグファイブに道徳的・倫理的行動を評価する上で特に重要とされる「正直さ-謙虚さ」の次元を含めた，HEXACO モデルを使用する．また，HEXACO モデルが複数の言語で一貫して 6 つの因子を再現することを実性を持つ強力な根拠となる． さらに，Wakabayashi（2014）が HEXACO の「外向性」と「誠実性」がビッグファイブとほぼ同じ概念を測定していると結論付けているため，ビッグファイブを用いた先行研究の知見を本研究にも同様に適用できる．

また，AI との HEXACO 性格特性の類似による影響を，ゲーム理論実験において，参加者に事前に相手の性格特性を明確に認識させた上で行う．これにより，AI がユーザーにとって単なるツールではなく，より信頼できるパートナーとして受け入れられるための有効なアプローチを提示し，AI 技術のさらなる社会浸透に貢献する．

2. 先行研究と仮説

本稿では，社会的アイデンティティ理論 （Tajfel and Turner,Tuner 1979）を拡張し，AI と人間の関係に適用する．この理論によれば人間は自らを特定の集団（内集団）に分類し，その帰属意識を高めることで自尊心を維持する．その結果，内集団のメンバーは外集団よりも好意的に評価される傾向がある．また，態度や興味が類似しているコンピュータ制御のアバターに対する投資額は大きくなり，類似性が信頼を高めることも示唆されている（Clerke andHeerey，2024）．我々はこれらの研究に基づき，AI と人間の性格が類似すると，人は AI を内集団として認識し，親近感や一体感を抱くと考える．本稿ではこの現象を「内集団親近

感」と定義する．加えて，本研究は信頼ゲーム（Berg et al.，1995）を用いて，被験者の信頼行動を測定する．AI の介入によって，投資行動が単なる合理的な意思決定ではなく，内集団親近感という心理的メカニズムによって説明できるのではないかと考える．

さらに，チャットボットとの会話実験では，ユーザーと AI の外向性の類似が信頼に影響を与えることも示されている（Kuhail et al.，2024）ものの，当該研究では協調性や誠実性といった他の性格，特性との関連性は示されていない．これらを踏まえると，信頼ゲームにおける投資行動は，AI との正確類似によってどのように変化するのだろうか．この問いを検証するため，我々は以下２つの仮説を立てた．

仮説 1：被験者と AI の性格が類似していると，反対している状況と比較して，AI に対する投資額が高い

仮説 2：外向性が高い被験者は，類似している性格の AI において，より高い信頼度を示す

3. 実験デザイン

本実験では，まず被験者の性格特性（HEXACO）を測定し，結果を被験者に提示した．測定には The Japanese IPIP-HEXACO short scale（Tokiwa, 2024）を使用した．

その後，参加者は以下の 3 条件のいずれかにランダムに割り当てられた．

統制群	AI の性格については特に情報を与えず，「AI とゲームを行います」とだけ告知.
類似群	AI の性格として，被験者と類似した性格を表示し，「あなたと類似した性格傾向を持つ AI とゲームを行います」と告知.
反対群	AI の性格として，被験者から反転させた性格を表示し，「あなたと反対の性格を持つ AI とゲームを行います」と告知.

続いて本研究では後藤（2023）を参考に，信頼ゲーム前後で IOS 尺度（Aron et al., 1992）を 2 回測定し，AI に対する親近感の変化を測定した．その後，ワンショットの信頼ゲームを実施した．参加者には 100pt が与えられ，任意の額（0〜100pt）を AI に投資することが求められる．ここでは被験者は，各条件に対応する AI を相手と認識して信頼ゲームを行なったと想定される．しかし実際は，受けとった額と同額を返すように設定した．ゲーム終了後，参加者にはアンケートに回答させた．調査項目には基本的な属性情報に加え，AI に対する一般的な態度を測定する尺度である GAAIS 尺度（Schepman et al., 2020），及び AI の性格をどれほど意識したかを測る設問を含めた．一連の実験は oTree を用いて開発した（Chen et al., 2016）．

4. 結果

表１　記述統計量

指標	統制群(N=96)	類似群(N=98)	反対群(N=96)
投資額	48.14 (31.39)	58.45 (36.07)	42.30 (31.33)
IOS スコア(事前)	2.47 (1.41)	2.67 (1.68)	2.07 (1.17)
IOS スコア(事後)	2.52 (1.58)	2.47 (1.68)	2.38 (1.42)
AI の性格意識度(信頼ゲーム)	-	4.86 (1.42)	4.83 (1.57)
AI の性格意識度 (IOS)	-	4.74 (1.43)	4.61 (1.65)
GAAIS ポジ平均	3.46 (0.63)	3.44 (0.73)	3.55 (0.61)
GAAIS ネガ平均(逆転)	3.12 (0.76)	3.23 (0.82)	3.36 (0.77)
H	3.92 (0.76)	4.11 (0.73)	4.16 (0.62)
E	2.88 (0.90)	2.75 (0.93)	2.80 (0.91)
X	2.75 (0.96)	2.52 (1.13)	2.45 (1.01)
A	2.83 (0.84)	2.96 (1.02)	2.88 (0.93)
C	3.12 (0.83)	2.90 (0.92)	3.11 (0.94)
O	3.15 (0.83)	3.10 (0.96)	3.12 (0.86)
年齢	51.60 (11.17)	47.48 (10.43)	49.32 (9.90)
性別: 女性	25 (26.0%)	32 (32.7%)	17 (17.7%)
性別: 男性	71 (74.0%)	65 (66.3%)	78 (81.2%)
性別: その他	0 (0.0%)	1 (1.0%)	1 (1.0%)

注: 性別以外の数値は平均値 (標準偏差).

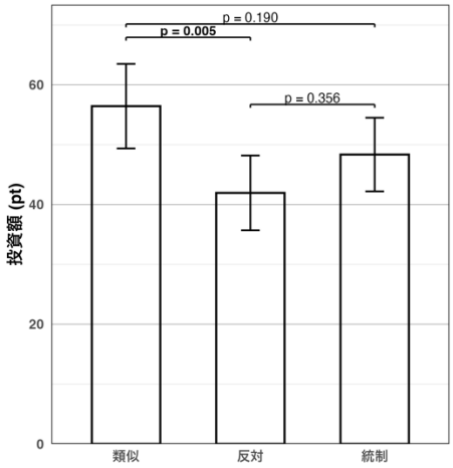

図 1　投資額の条件別平均

表 2　重回帰分析：類似群

Variable	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-57.1800	36.0401	-1.587	0.1162
H	11.2947	5.1816	2.180	**0.0319***
E	2.6101	4.4208	0.590	0.5564
X	-4.8801	3.3593	-1.453	0.1499
A	-1.4746	4.4248	-0.333	0.7397
C	-0.1653	4.3637	-0.038	0.9699
O	2.0281	4.2325	0.479	0.6330
IOS(事前)	0.3838	2.1940	-0.175	0.8615
GAAIS ポジ平均	10.7582	5.2961	2.031	**0.0452***
GAAIS ネガ平均	10.7765	4.8193	2.236	**0.0279***

注: GAAIS ネガ平均は逆転済み

仮説 1 について，分散分析の結果，条件の主効果は有意であり(*F(2, 287)=5.95, p=.003, η²=.040*)，Tukey の多重比較では，類似条件の平均投資額が反対条件より約 14.5 点高かった

($p = .005$）(図 1)．よって，仮説 1 は支持された．

仮説 2 の検証のため，投資額を目的変数とし，各群において重回帰分析を実施した(表 2)．その結果，類似群でのみ，Honesty-Humility ($\beta = 11.29, p < .05$)，GAAIS ポジ ($\beta = 10.76, p < .05$)，ネガ($\beta = 10.78, p < .05$）が有意に影響を示した．一方，仮説 2 は支持されなかった．

5. 考察

仮説 1 に関しては，類似群における投資額が反対群と比較して有意に高いことが示された．さらに，類似群の IOS スコアは事前，事後ともに反対群よりも平均値が高かった．この結果は，「内集団親近感」が信頼行動に影響を与えた可能性を示唆している．社会的アイデンティティ理論によれば，人間同士における内集団への好意的評価や優遇がみられるが，本研究では同様のメカニズムが人間と AI の関係にも適用し得ると考えられる．

また，仮説 2 においては，AI に対する信頼行動（投資額）が被験者の誠実さや謙虚さを示す Honesty-Humility（以下，H）という性格特性と正に有意な関連がみられた．Thielmann et al.（2020）の理論的枠組みにおいて，本研究で実施した信頼ゲームは相手に利用されるかもしれないというリスクを含む状況であり，*H* はその影響を最も強く受ける特性である．加えて本研究の実験デザインでは，AI と自分の性格が類似していると提示されたとき，被験者は「相手は自分と同様に誠実に行動するだろう」と推測しやすい．そのため *H* が高い被験者ほど，AI に対しても安心して投資した可能性が考えられる．反対に，社交性や会話の円滑さに関連する要素である外向性（*X*）が有意ではなかった原因は，先行研究ではチャットボットとの会話実験を扱っているが，本研究では会話を伴わない経済ゲーム実験を使用しているため，外向性の影響を引き出すのに適していなかった可能性が考えられる．

また，先行研究では AI に対する一般的態度（GAAIS）が返戻行動に影響することは示されなかった (Upadhyaya and Galizzi，2023）が，本研究では投資行動の測定の結果，GAAIS は投資額に正に有意な影響を与えることが明らかとなった．すなわち，AI に対する好意的態度は互恵性に影響しないが，信頼行動の形成には寄与する可能性がある．

6. 今後の展望

本研究では，統制の精度を高めるため，HEXACO モデルを反映したペルソナ AI の構築は行わなかった．また，仮説検証においては主に外向性に焦点を当てた．これは，外向性と対人関係や信頼行動との関連が Kuhail et al.（2024）で明確に示されていたためである．しかし，H などは，AI との信頼形成において重要な役割を果たす可能性があることがより明らかとなった．したがって，今後は HEXACO の各特性を反映したペルソナ AI を構築し，外向性に限らず多次元的に AI への信頼を検証することが望ましい．これにより，現実により近い条件での検証を可能にし，外的妥当性を高めることにつながるだろう．

近年，AI にどのような性格や振る舞いを期待するかをめぐる議論が社会的なテーマとな

っている．例えば，OpenAI のリリースした GPT-5 は応答の正確性や高速化といった点で高く評価される一方，「冷たい」と指摘され，前世代モデル GPT-4o を支持するユーザーの「keep4o 運動」が展開された．Sam Altman 氏も自身のソーシャルメディア上で「AI の知能面以外の要素を過小評価していた」と認め[1]，今後はユーザーによるカスタマイズを重視する方針を示している．この議論は，AI 研究において冷静な論理や合理性だけでなく，ユーザーに合わせた柔軟な最適化を取り込むことの重要性を示唆している．その成果は， AI を信頼できる社会的パートナーとして社会に浸透させるための基盤を築くものとなるだろう．

引用文献

Aron, A., Aron, E. N., and Smollan, D, 1992. Inclusion of Other in the Self Scale and the structure of interpersonal closeness. Journal of Personality and Social Psychology, 63(4), 596-612.

Chen, D.L., Schonger, M., and Wickens, C, 2016. oTree - An open-source platform for laboratory, online, and field experiments, Journal of Behavioral and Experimental Finance 9, Elsevier, 88-97.

Clerke, A. S., and Heerey, E. A, 2021. The Influence of Similarity and Mimicry on Decisions to Trust. Collabra: Psychology, 7(1), 23441.

後藤，晶. 2023,人間は『人工知能』と『協力』できるか: クラウドソーシングを用いた仮想的 AI エージェント実験による検討. 社会情報学, 12(1), 1-17.

Jakob, N. 2024, AI Stigma: Why Some Users Resist AI's Help?, https://www.uxtigers.com/post/ai-stigma

Jiang, H., A. Guo and J. Ma, 2020. Personality-aware Chatbot: An Emerging Area in Conversational Agents. Transactions on Emerging Topics in Computational Intelligence, 1–19. Available: https://www.researchgate.net/publication/346468952 (accessed Sep. 24, 2025).

Joyce Berg, John Dickhaut, Kevin McCabe, 1995. Trust, Reciprocity, and Social History. Games and Economic Behavior. 10, 122-142

Kuhail, M. A., Bahja, M., Al-Shamaileh, O., Thomas, J., Alkazemi, A., and Negreiros, J, 2024. Assessing the impact of chatbot-human personality congruence on user behavior: A chatbot-based advising system case. IEEE Access, 12, 71761-71782.

Leikas, S. and Salmela-Aro, K, 2014. Similarity and attraction in personality: The role of personality traits, values, and interests in young adult couples. European Journal of Personality, 28(5), 469–482

Schepman, A., and Rodway, P, 2020. Initial validation of the general attitudes toward artificial intelligence scale. Computers in Human Behavior Reports, Volume 1, January–July 2020.

Selfhout, M. H., Denissen, J. J., Branje, S. J., and Meeus, W. H, 2010. Intra-individual variability and similarity in personality development among friends. Journal of

Personality, 78(4), 1251–1282.

Tajfel, H., and Turner, J. C, 1979. An integrative theory of intergroup conflict. In W. G. Austin, and S. Worchel (Eds.), The social psychology of intergroup relations (pp. 33-37).

Thielmann, I., Spadaro, G., and Balliet, D, 2020. Personality and prosocial behavior: A theoretical framework and meta-analysis. Psychological Bulletin, 146(1), 30–90.

Tokiwa, E, 2024. The Japanese IPIP-HEXACO Short Scale: Development and Validation of Reliability and Validity. https://osf.io/preprints/psyarxiv/7hx9c_v1

Upadhyaya, N., and Galizzi, M. M, 2023. In bot we trust? Personality traits and reciprocity in human-bot trust games. Frontiers in Behavioral Economics, 2, 1164259.

Wakabayashi, A, 2014. A sixth personality domain that is independent of the Big Five domains: The psychometric properties of the HEXACO Personality Inventory in a Japanese sample. Japanese Psychological Research, 56(3), 211–223

[1] Sam Altman. [@sama], (2025, August 9), https://x.com/sama/status/1953953990372471148