人と AI の協働におけるバイアスの相互作用：効果と信頼性の検討
Bias interactions in human-AI collaboration: Examinations of effects and trustworthy

白砂 大[†]，本田 秀仁[‡]，香川 璃奈[††]
Masaru Shirasuna, Hidehito Honda, Rina Kagawa
[†]静岡大学, [‡]追手門学院大学, [††]産業技術総合研究所
Shizuoka University, Otemon Gakuin University, National Institute of Advanced Industrial Science and Technology
m.shirasuna1392@gmail.com

概要

昨今、医療診断をはじめとする実社会の判断場面において、AI(Artificial Intelligence)による意思決定支援の導入が進んでいる。本研究では、人と AI の両者の「バイアスの相互作用」(e.g., 過大評価か過小評価か) に着目し、人のバイアスの程度によって AI が持つべき特性が異なる可能性を検証した。計算機シミュレーションと行動実験を通して、人のバイアスと逆方向のバイアスをもつ AI は、人の判断精度を高めやすい一方で、人が AI に抱く信頼感は高くないことが示された。本研究は、正確で信頼できる AI が常に良いとは限らないという実用的示唆を提供する。

キーワード: 意思決定支援(decision support); 人と AI の協働(human-AI collaboration); バイアス(bias); AI の信頼性(AI trustworthy)

1. はじめに

昨今、AI(Artificial Intelligence)による意思決定支援の導入が進んでいる。例えば医療診断場面において、ある医療画像を見て AI が異常の有無を判定し、それに基づいて医者が判断を行うような診断支援システムが開発されている(e.g., Reverberi et al., 2022)。

このような AI を設計・開発するうえでは一般に、正確で信頼できる AI がよいと考えられる。しかし、最終的に AI を使うのは人である。そのため、AI 単独の判断の精度よりも、人と AI の判断をかけ合わせた際の精度を考えることが重要となる。例えば、人が特定の方向のバイアス(e.g., 過大評価)を持っていれば、AI がそれと逆のバイアス(e.g., 過小評価)を持っているようにふるまうことで、両者のバイアスが打ち消し合わされ、より正確な判断ができるかもしれない。実際に、認知科学における集合知の知見では、集団内における個々の判断が多様であることが、集団意思決定をよくするうえで重要だと主張されている(e.g., Surowiecki, 2004)。

一方、AI に対する信頼性を考えると、人は自分と異なる回答を下す AI を信頼できないと感じるかもしれない。一般に、人は自分と異なる意見を受け入れにくく、似た意見・特性の他者に協力的になる(e.g., Balliet et al., 2014)。そのため、自分と逆のバイアスを持つようにふるまう AI を受け入れることが効果的だったとしても、人はそれを信頼できないと感じる可能性がある。

本研究では、AI による意思決定支援の中でも、正答・誤答を明確に定義でき、瞬時の判断が求められる二者択一の知覚判断場面に着目した。この課題構造は、先述のような病気の異常の有無を診断するシステムを抽象化したものである(課題の詳細は後述)。そのうえで、人のバイアスの程度によって AI が持つべき特性が異なる可能性を検証した。特に、AI がバイアスを持つようにふるまうことの効果、および AI に対する信頼性という 2 つの側面について、理論(計算機シミュレーション)と実験(行動実験)の双方から検証した。なお本研究は、追手門学院大学研究倫理委員会の承認を得て実施された(承認番号 2024-23)。

2. 実験 1: 計算機シミュレーション

正確な AI が必ずしも適切とは限らず、人が持つ事前信念によって AI が持つべき特性が変わり得ることを、理論的に検証した。

2.1. 方法

題材: Vicente & Matute (2023)で扱われた仮想的な医療判断課題を題材とした。具体的には, 患者の細胞の拡大画像を模した格子画像(後述の図 2 も参照)が呈示され、濃い部分の面積が全体の 50%より多いか(50%より多ければ「陽性」とする)を、人が AI のアシストを見ながら識別する課題とした。このシミュレーションでは、「陽性」の刺激が呈示された場面を仮定し、その正答率(人が正しく『陽性』と判断できる率)を算出した。

手続き: 人の当初の推定値が、AI によって更新され、最終的な推定が行われると仮定した。具体的には、Honda et al. (2024)に基づいて、「人の事前信念」と「AI アシストの事前信念」が正規分布から生成されると仮定し、両者の合成分布を「人の事後信念」と見なした。

$$\mu_{posterior} = \frac{\frac{\mu_{prior}}{SD^2{}_{prior}} + \frac{AI_{prior}}{SD^2{}_{AI}}}{\frac{1}{SD^2{}_{prior}} + \frac{1}{SD^2{}_{AI}}}$$

$$SD_{posterior} = \left(\frac{1}{SD^2{}_{prior}} + \frac{1}{SD^2{}_{AI}}\right)^{-1/2}$$

なお、μ_{prior}は人の事前信念(i.e., 当初、濃い部分の面積が何%だと思ったか)の平均、μ_{AI}は AI の事前信念の平均、SD_{prior}は人の事前信念の標準偏差、SD_{AI}は AI の事前信念の標準偏差、$\mu_{posterior}$は人の事後信念(i.e., AI を見た後で、濃い部分の面積が何%だと思うか)の平均、$SD_{posterior}$は人の事後信念の標準偏差とした。そして、$\mu_{posterior}$が 50 から大きくなるほど「陽性」と判断されやすい、すなわち正答率が高くなると見なし、事後分布の.50 を超える部分の面積を算出した。

パラメータ設定: μ_{prior}およびμ_{AI}は、40 から 60 まで 0.2 刻みで、全 101 通りに操作した。SD_{prior}は 2、5、10 の 3 通りに操作し(それぞれ、事前信念が強、中、弱)、SD_{AI}は 5 で固定した。よって、パラメータは全部で 101 * 101 * 3 * 1 = 30,603 通りに設定された。

2.2. 結果と考察

図 1 の、横軸はμ_{prior}、縦軸はμ_{AI}であり、ヒートマップの白色はチャンスレベル(正答率 50%)を表す。陽性刺激の出現が仮定されていたため、一見するとμ_{AI}が 50 より大きい方が、一貫して有利だと考えられる。しかし実際には、正答率が高くなる(青色)のは人と AI との事前信念の関係に依存していた。例えば、人が大きい値に強い事前信念を持っていた場合は(e.g., μ_{prior}> 50 かつ SD_{prior} = 2)、μ_{AI}が 50 未満であっても正答率は高くなった。このことから、AI が持つべき特性は人が持つバイアスによって異なると考えられる。

3. 実験 2: 行動実験

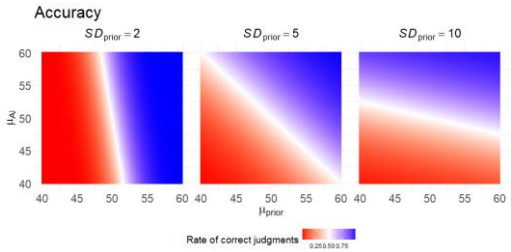
続いて、実験 1 で得られた理論的知見が実際の行動レベルでも見られるかについて検証した。

3.1. 方法

実験参加者: 日本人 512 名が実験に参加した(M_{age} = 42.7, SD_{age} = 10.4)。

題材・刺激: 実験 1 と同様の、仮想的な医療判断課題(Vicente & Matute, 2023)を実施した。具体的には、薄い黄色と濃いピンク色からなる格子画像(患者の細胞の拡大画像を模したもの; 図 2)が呈示され、「濃い部分の

図 1 計算機シミュレーションの結果



註) 左から順に、SD_{prior} = 2, 5, 10 の結果を示す。横軸はμ_{prior}、縦軸はμ_{AI}を示す。$\mu_{posterior}$ が 50 に近いほど白色で示されている。

面積が全体の 50%より多ければ、その患者は陽性である」という設定のもと、参加者に「陽性」か「陰性」かの判断を求める課題であった。実験刺激に関しては、濃い部分の面積が 30%から 70%まで 1%刻みで用意され、各%について異なる画像が 2 枚ずつ作成された。すなわち、全部で 41 * 2 = 82 通りの画像が作成された。

手続き: 実験はオンラインで行われた。1 試行の流れは次の通りであった。PC 画面上には、問題文と刺激画像、それに「陽性」「陰性」の 2 つのボタンが呈示された。参加者は、刺激画像を見て、2 つのボタンのいずれかをクリックするよう求められた。クリック後、「次へ」と書かれたボタンが出現し、これがクリックされると次の試行に移行した。1 セットは 82 試行からなり、全部で 3 セット行われた。1 セット内では、上述の刺激 82 枚が 1 枚ずつランダムな順番で呈示された。

なお第 2 セット終了後のみ、AI に対する信頼度も回答することが求められた。具体的には、Hoffman et al. (2023)に基づき、「この AI はよく機能していると思う」など 7 項目について、visual analog scale を用いて「0 (全くそう思わない) — 100 (非常にそう思う)」で回答させるアンケートを実施した。

図 2 行動実験のスクリーンショット

註) 左は第 1・3 セット、右は第 2 セットである。

条件: 全 3 セットのうち、第 1・第 3 セットは上記の手続きの通りに行われた(図 2 左)。第 2 セットでは、各試

行においてAIの判定も呈示された。具体的には、刺激画像の下に、AIのイラストと、AIの判定(「陽性」または「陰性」)が呈示された(図2右)。AIは、次の各基準で「陽性」と判定するよう操作されており、参加者はこの5条件のいずれかにランダムで割り当てられた。

・Positive+条件(n = 100): 濃い部分の面積が40%以上
・Positive条件(n = 104): 濃い部分の面積が45%以上
・Accurate条件(n = 103): 濃い部分の面積が50%以上
・Positive条件(n = 103): 濃い部分の面積が55%以上
・Positive条件(n = 102): 濃い部分の面積が60%以上

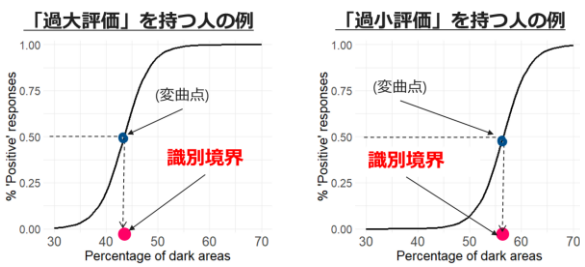
すなわちAIは、Positive+条件ほど過大評価バイアス(i.e., 濃い部分が少なくても「陽性」と判断する)、Negative+条件ほど過小評価バイアス(i.e., 濃い部分が多くないと「陽性」と判断しない)を持つようにふるまうことが想定された。

バイアスの指標「識別境界」: 本研究では、参加者が何%を基準に陽性・陰性の判断を切り替えたかを行動データから推定し、これを「識別境界」と定義した。具体的には、各参加者の回答データをシグモイド関数に当てはめ(i.e., 独立変数を「濃い部分のパーセンテージ」、従属変数を「陽性と回答したか否か」とするロジスティック回帰分析を行った)、そのときの変曲点を識別境界とした。例えば、ある参加者の識別境界が「42」であれば、その参加者は濃い部分が42%以上で「陽性」と回答する。一方、識別境界が「57」なら、濃い部分が57%以上にならないと「陽性」と回答しない(図3)。すなわち、識別境界が50より小さければ過大評価、大きければ過小評価のバイアスを持つ人だと解釈できる。

3.2. 結果と考察

まず、濃い部分の各パーセンテージ(30~70%)について、全参加者の正答率を算出したところ、平均 .887、標準偏差 0.150、第一四分位数 .825 と概して天井効果が見られた。ただし、濃い部分が「43~53%」の刺激については、いずれも正答率が第一四分位数の.825 を下回っており、比較的難しい問題だったと考えられた(平均 .671、標準偏差0.144)。よって以下では、濃い部分が「43~53%」の刺激が呈示された試行を分析対象とした。

図3 「識別境界」の概念図



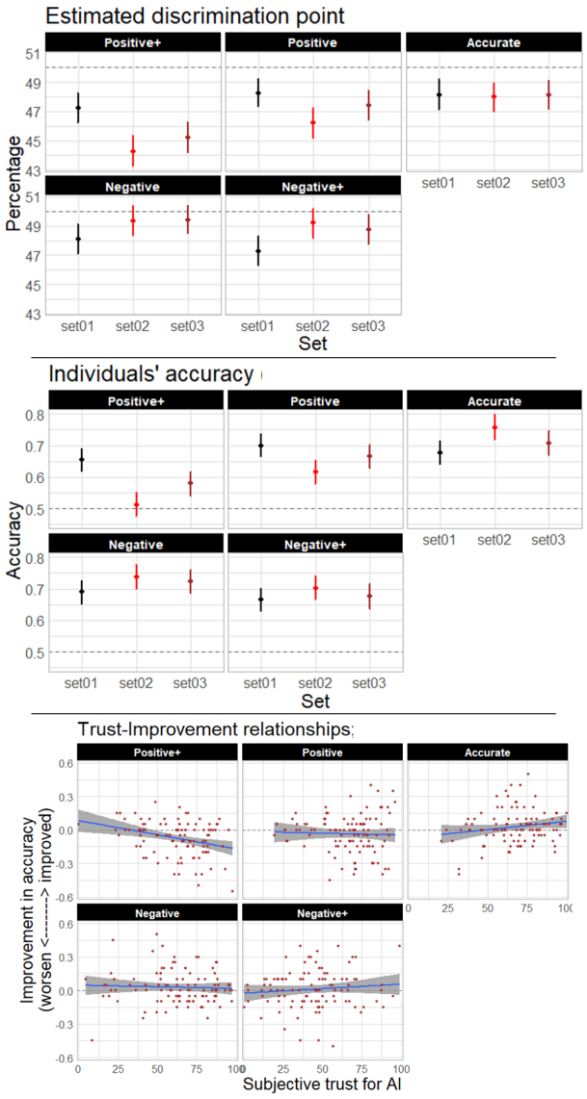
識別境界と正答率: 条件とセットを独立変数、識別境界または正答率を従属変数とする分散分析モデルを構築した。分析では、Rのbrmsパッケージ(Bürkner, 2017)を用いて、MCMC法(繰り返し数4000、バーンイン2000、チェーン数 4)によって各条件・セットにおける識別境界や正答率の95%信用区間(CI)を推定した。

図4上段(黒色)より、第1セットの識別境界は50以下の範囲に多く分布していた($M_{Positive+}$ = 47.2, $M_{Positive}$ = 48.2, $M_{Accurate}$ = 48.1, $M_{Negative}$ = 48.1, $M_{Negative+}$ = 47.3)。すなわち本実験において、多くの参加者は、AIを観察する前の時点では過大評価バイアスを見せていたといえる。

続いて、AIの効果を検証すべく、第1から第2セットの推移に着目した。識別境界について(図4上段)、どの条件においてもAIのバイアスの方向に識別境界がシフトしていた。特に、多くの参加者(過大評価)とは逆方向のバイアス(過小評価)を持つNegativeとNegative+の両条件では、第2セットの95%CIが50をまたいでおり(Negative CI_{set02} = [48.4, 50.4]; Negative+ CI_{set02} = [48.2, 50.2])、参加者の持つバイアスが修正されたといえる。実際に、正答率は(図4中段)、Negative・Negative+両条件の第2セットで向上していた(Negative M_{set01} = .690, M_{set02} = .738; Negative+ M_{set01} = .666, M_{set02} = .702)。なお第2セットで最も正答率が高かったのはAccurate条件であったが($M_{Accurate}$ = .757)、Accurate、Negativeの各条件における第3・第2セット間の正答率の差($M_{set03} - M_{set02}$)はそれぞれ –.049 (= .707 – .756)、–.015 (= .723 – .738)であった。すなわち、Accurate条件よりもNegative条件の方が、AIが取り除かれた後も正答率が維持されていた。以上より、人の判断の精度を高める際に、バイアスを持つようにふるまうAIも有効であることが示唆された。

信頼性と正確性の関係: 最後に、呈示されたAIに対する信頼度、および信頼度と正答率向上との関係を分析した。まず、信頼度の指標として、第2セット直後のアンケート全7問に対する評定値の平均を、参加者ごとに算出した。また正答率向上の指標として、第3セットと第2セットの正答率の差を、参加者ごとに算出した。もしAIがより正確さを高めるのであれば、両者の間には正の相関が現れるはずである。しかし実際には、どの条件においても有意な正の相関は見られなかった(図4下段)。特にPositive+条件では、有意な負の相関が確認された($r_{Positive+}$ = -.296, p = .003; $r_{Positive}$ = -.079, $r_{Accurate}$ = .149, $r_{Negative}$ = -.095, $r_{Negative+}$ = -.178,

図 4 行動実験の結果



註) 上段と中段: 横軸は各セット、縦軸はそれぞれ識別境界と正答率を示す。エラーバーは 95%信用区間を示す。下段: 横軸は AI への信頼度、縦軸は正答率の向上度合い(第 3 セット – 第 2 セット)を示す。各点は各参加者であり、回帰直線の周囲の灰色は標準誤差を示す。全段共通: 各パネルは条件を示す(上行: 左から Positive+、Positive、Accurate。下行: 左から Negative、Negative+)。

ps > .10)。以上の結果は、信頼できる AI が常によいわけではないことを示している。言い換えれば、人と逆方向のバイアスを持つ AI は、たとえそれを受け入れることが効果的であったとしても、当人にとって信頼できないと感じられがちであるとも言える。

4. 総合考察

本研究では、AI による二者択一の単純な意思決定支援を題材に、「人が持つバイアス」と「バイアスを持つようにふるまう AI」との相互作用に着目した。そして、人のバイアスの程度によって AI が持つべき特性が異なる可能性を、計算機シミュレーションと行動実験から検証した。その結果、AI が人と逆方向のバイアスを持つようにふるまうことが時には効果的であること、一方で人は、逆方向のバイアスを持つようにふるまう AI を信頼しづらいことが、それぞれ示された。

本研究の知見は、「正確かつ信頼できる AI を設計するのがよい」という一般的な通念に疑問を投げかけ、意思決定支援 AI を設計する側、使用する側の双方に実用的示唆を与える。設計者側は、AI 単独の判断の精度を高めるだけでなく、人(AI の使用者)の認知バイアスを考慮することも重要だといえる。また使用者側も、信頼できないからといって AI のアシストを拒否するのではなく、時にはそれを受け入れることが、多様な視点を取り入れるうえで重要だろう。今後は、AI を用いた意思決定支援場面への応用可能性を検討する必要がある。例えば、単純な知覚刺激画像ではなく実際の医療画像を用いたり、医療従事者に対して同様の実験を行ったりする方法が考えられる。また、長期のタイムスパン(e.g., 1 日以上)をおいて追実験を行うなど、AI の効果の持続性もより詳細に検証すべきだろう。

謝辞

本研究の一部は JST さきがけ(JPMJPR23I3)および科研費(23K16249)の支援を受けた。

文献

Balliet, D., Wu, J., & De Dreu, C. K. W. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psych Bull*, *140*(6), 1556–1581.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *J Stat Software*, *80*(1), 1–28.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, *5*.

Honda, H., Kagawa, R., & Shirasuna, M. (2024). The nature of anchor-biased estimates and its application to the wisdom of crowds. *Cognition*, *246*, 105758.

Reverberi et al. (2022). Experimental evidence of effective human–AI collaboration in medical decision-making. *SciRep*, *12,* 14952.

Surowiecki, J. (2004). The wisdom of crowds. In *Anchor*.

Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Sci Rep*, *13*, 15737.