予測精度の高いAIを利用しても、人間の意思決定は正確にならない。
AI with high prediction performance does not necessarily make human decisions more accurate.

香川 璃奈[1]，本田 秀仁[2]，野里 博和[3]
Rina Kagawa, Hidehito Honda, Hirokazu Nosato
[1]筑波大学，[2]追手門学院大学，[3]産業技術総合研究所
University of Tsukuba, Otemon University, AIST
kagawa-r@md.tsukuba.ac.jp

概要
人間が他者からの助言を参照して意思決定を行う際に、助言をそのまま採用するわけではない現象は自己中心的助言割引として知られる。昨今のAIの急速な発展により、AIを助言として人間が意思決定を行う場面が今後増加すると予想される。しかし、AIの精度と、それを利用した人間の意思決定の関係性は明らかでない。本研究ではシミュレーションと行動実験を通じて、AIの助言の予測誤差が小さくなるほど、それを利用した人間の意思決定が正確になるとは限らないことを示した。

キーワード：自己中心的助言割引，AI，意思決定，algorithm aversion, algorithm appreciation

背景

人間は、意思決定を行う際に、他者からの助言を参照することがある。その一方で、人間が他者からの助言をそのまま採用するわけではない現象が、自己中心的助言割引として知られている(Yaniv & Kleinberger, 2000)。昨今では深層学習を始めとするいわゆるAI技術[1]の急速な発展に伴い、人間がAIによる予測や分類の結果を助言として利用して意思決定を行う場面も多くなっている。それと同時に、AIによる助言は人間による助言と比較して、利用されにくい(algorithm aversion) (Dietvorst, Simmons, & Massey, 2015)、または利用されやすい(algorithm appreciation) (Logg, Minson, & Moore, 2019)という報告が認められる。しかし、人間がAIの助言をどのように利用するのか、人間による助言の利用とどのように異なるのか、一定の知見には至っていない。特にAIによる助言の精度と、それを利用した人間の意思決定の変化に関する系統的な特徴は明らかになっていない。

本研究では数値予測を対象として、AIの予測精度の向上に伴い、それを利用した人間の意思決定がどのように変化するかを、理論的および実験的検証により議論する。実社会におけるAIの利用場面を想定して、以下の2点を前提とする。

1. 人間は、AIの予測結果を知る前に、自分なりの予測をする。その後、AIの予測結果を踏まえて人間が最終的な判断を下す。AIの予測結果が自動的に人間の意思決定に代わるのではない。
2. AIの予測結果は予測誤差を伴って示される。予測誤差が狭いほど精度の高いAIとみなす。例：「AIは、写真の男性の体重を72.5-75.5kgだと予測した」という予測結果より「AIは、写真の男性の体重を73.8-74.2kgだと予測した」という予測結果の方がAIの精度は高いと判断する。

目的

数値予測タスクにおいて、予測誤差を伴う助言を利用する認知過程のモデルを提案し、以下の2点についてシミュレーションおよび行動実験により検討する。

RQ1：助言の予測誤差が変化すると、人間の最終判断の正確性はどのように変化するか？

RQ2：助言の予測誤差の変化に伴う人間の最終判断の正確性の変化は、助言者によって変化するか？

助言を利用するシナリオの整理と問題定義

まず、判断者は助言なしに判断J_1を行う(図1における「80」)。次に、“x_l–x_h” ($x_l < x_h$) という予測誤差(Err_{advice})を持つ助言が示される。x_lとx_hのうちJ_1に近い値をADV_{near} (図1における「75.5」)とし、もう片方をADV_{far} (図1における「72.5」)とする。本研究では、正答($target \geq 0$)はx_lとx_hの平均値とする(図1の例では74.0)。最後に、助言を確認した人間が最終判断J_2 (図1における「75」)をする[2]。

[1] 本稿では、何かしらの数理的技術に基づく識別または予測手法を指すこととする。

[2] この手続きは、助言に基づく意思決定を評価するための実験プロトコルとして知られるjudge-advisor

本稿では、人間の最終判断の正確性として、Err_{advice}の変化に伴う$Err_{J_2} = |J_2 - target|$の変化を評価する。

$$\text{Advice}: \begin{cases} "ADV_{far} - ADV_{near}" & \text{if } J_1 - target \geq 0, \\ "ADV_{near} - ADV_{far}" & \text{if } J_1 - target < 0. \end{cases}$$

$$ADV_{far} = \begin{cases} target - Err_{advice} & \text{if } J_1 - target \geq 0, \\ target + Err_{advice} & \text{if } J_1 - target < 0. \end{cases}$$

$$ADV_{near} = \begin{cases} target + Err_{advice} & \text{if } J_1 - target \geq 0, \\ target - Err_{advice} & \text{if } J_1 - target < 0. \end{cases}$$

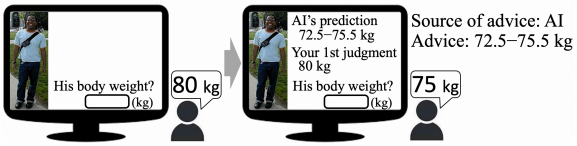

図 1: 実社会における助言利用シナリオのイメージ

予測誤差を伴う助言を利用した意思決定プロセスのモデル

助言利用の先行研究(Himmelstein, 2022; Vodrahalli, Daneshjou, Gerstenberg, & Zou, 2022)を参照し、予測誤差を伴う助言を利用した意思決定プロセスについて、以下の 3 段階のモデルを提案する。

○Pre-step: 当初の判断(J_1)は$N(target, \sigma_1)$にしたがって決定される。σ_1は判断者の知識レベルやタスクの難しさを反映する。

○Activation Step: 助言を得た判断者が、最終判断を当初の判断(J_1)から変更するか決定する段階を示す。判断者が最終判断をJ_1から変更する確率(P_{change})は当初の判断への自信$(Weight_{J_1})$と助言への信頼$(Weight_{advice})$のバランスで決定され、さらにそのバランスはJ_1とADV_{near}の距離に従うと考える。a_iとd_iは$Weight_i$の上限と下限を意味する。b_iはdis_1の変化に伴う$Weight_i$の変化の傾きを意味する。c_iは$Weight_i$の変曲点である。

$$dis_1 = \begin{cases} J_1 - ADV_{near} & \text{if } J_1 - target \geq 0, \\ ADV_{near} - J_1 & \text{if } J_1 - target < 0. \end{cases}$$

$$Weight_i(dis_1) = \frac{a_i}{1 + e^{-(b_i(dis_1 - c_i))}} + d_i$$

$$P_{change}(dis_1) = Weight_{J_1}(dis_1) \times Weight_{advice}(dis_1)$$

○Integration Step: $J_2 \neq J_1$の場合、J_2の値は以下に従う。なお$(ADV_{far} + ADV_{near})/2 = target$である。

$$J_2 = N((target + J_1)/2, \sigma_2)$$

シミュレーション実験

$N(target, \sigma_1)$に従う 500 点をランダムに生成し、各パラメーターを変化させた際の平均値を図 2 に示した。

Err_{advice}が小さくなるほどErr_{J_2}が小さくなる傾向が多くの条件で確認されたものの、その傾向は非単調かつ非線形であった。Err_{advice}が小さくなるほどErr_{J_2}が小さくなる傾向を示さない条件($\sigma_1 = 15$ かつ

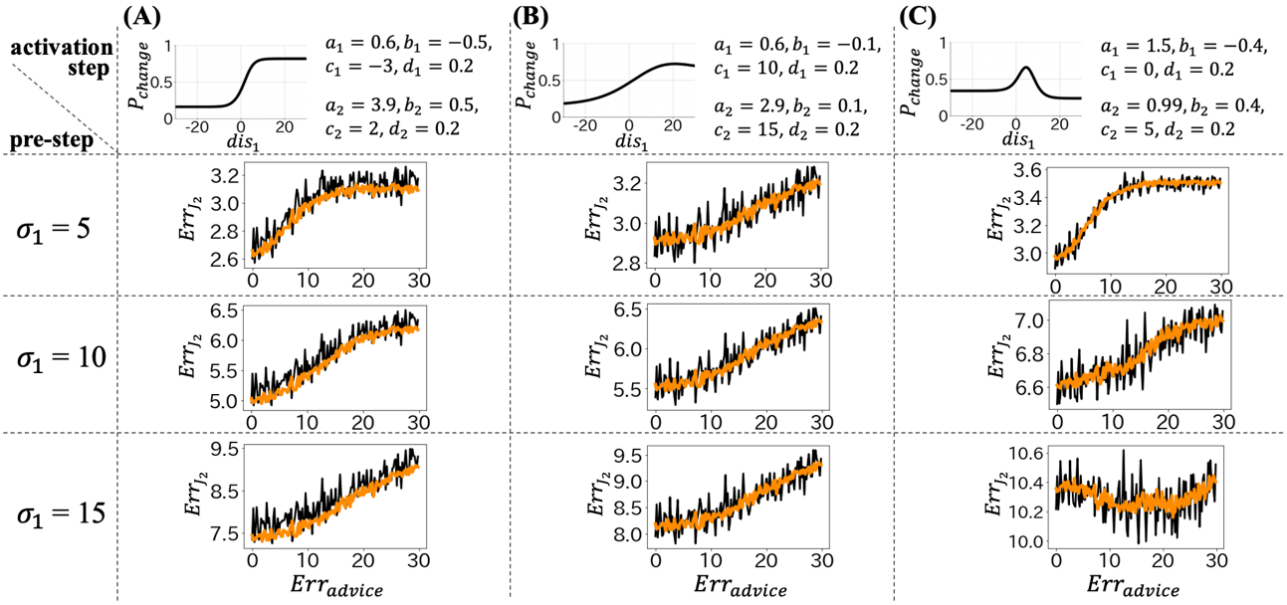

図 2: シミュレーション実験の結果

system (JAS; Sniezek & Buckley, 1995)と一致している。

activation step が(C)) も確認された。このことから、Err_{advice}が小さくなるほどErr_{J_2}が小さくなるとは限らないと示唆された。

行動実験　方法

次に、実際にシミュレーション研究と同様のシナリオを利用した行動実験を行なった。

実験参加者: 152 名(男性 104 名、女性 48 名、M_{age} = 42.6、SD_{age} = 22.7)。

刺激・手続き: 全ての実験参加者は、ランダムに提示される 60 問の数値推定タスクに回答した。

地理的な予測タスク(Himmelstein and Budescu, 2023)として、過去 8 週間のとある地下鉄駅の乗降者数の推移から翌週のその駅の乗降者数[3]を推定した。正答($target$)の平均値は 521.2 ± 198.2 (最小 : 123、最大 : 1,010)。Err_{advice}は 10 段階(0.0, 4.95, …, 44.55)を設定した。

1 問のタスクは、以下の 3 段階から構成される刺激で構成される。

まず、参加者は、過去 8 週間のある地下鉄駅の乗降者数の推移から翌週のその駅の乗降者数を推定する。

次に、助言が提示される。助言は助言者と助言内容がランダムに組み合わされて構成される。助言者は、「AI」「ある鉄道会社の職員」「以前この課題を解いた 100 人」の 3 通り[4]とした。助言内容は問題ごとに、J_1 に応じて"$ADV_{far} - ADV_{near}$"または"$ADV_{near} - ADV_{far}$"とした、ただしADV_{far}およびADV_{near}は四捨五入した整数とした。

> 助言内容の例 :「来週のこの駅の乗降者数について、AI は 118−128(万)人と推定した。」($target$ = 123, Err_{advice} =4.95 の場合)

最後に、その助言に基づき、参加者は最終的な推定値を決定する。なお、自身の当初の推定値と助言は同一の画面に示されている。

本実験は、筑波大学医学医療系医の倫理委員会の承認(承認番号 1743)を得ている。

行動実験　結果・考察

3 段階モデルの検証

Activation Step について、助言者ごとにErr_{advice}の階層性を考慮した状態空間モデルを利用して算出したdis_1とP_{change} の関係を図 3(A-1)に示した。助言者を問わず、dis_1が 0 以下のときはP_{change}が小さく、0 以上になるとP_{change}が大きくなったこと、さらに、dis_1がある一定以上になるとP_{change}が一定になったことが確認できた。

Integration Step について、$(target + J_1)/2$とJ_2の関係を図 3(A-2)に示す。$(target + J_1)/2$を固定効果、J_2を目的変数、Err_{advice}と助言者と参加者をランダム効果としたマルチレベルモデルで推測された傾きと 95% 信用区間は 1.01 [0.98, 1.05] であり、$(target + J_1)/2 = J_2$の関係が成り立つと判断できる。

これらの結果から、想定した 3 段階モデルは妥当であると判断した。

助言の予測誤差と最終的な推定値の正確性の関係

Err_{advice}とErr_{J_2}の関係を図 3(B)に示した。Err_{advice}とErr_{J_2}は助言者に依らず非単調かつ非線形であり、シミュレーション実験と結果に矛盾はなかった。また、助言者が「AI」または「ある鉄道会社の職員(specialists)」の場合と比較して、助言者が「以前この課題を解いた 100 人(laypeople)」の場合、判断者の最終判断の誤差がより大きくなる傾向が認められた。

助言者が「AI」の場合と「以前この課題を解いた 100

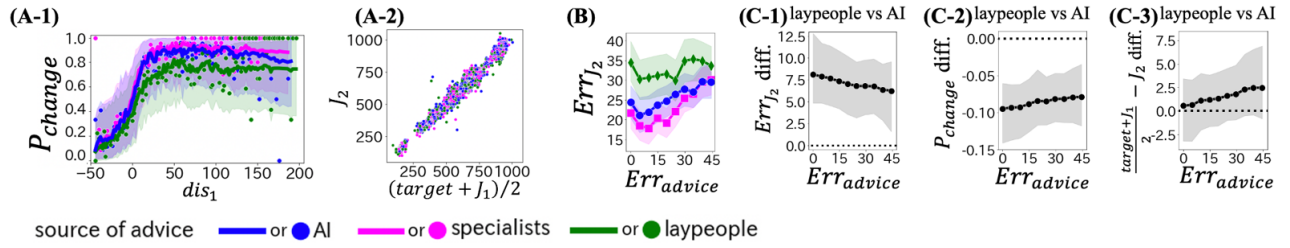

図 3: 行動実験の結果

[3]https://catalog.data.gov/dataset/mta-subway-customer-journey-focused-metrics-beginning-2015 (last accessed 2023/4/1)

[4] 実際にこれらの助言者が計算または回答したわけではなく、あくまでランダムに表示しただけである。

人」の判断者の最終判断の誤差の差を、助言の予測誤差ごとに推定した結果が図 3(C-1)である。この結果から、全てのErr_{advice}において、助言者が「以前この課題を解いた 100 人」であった場合には、「AI」であった場合と比較して有意にErr_{J_2}が大きいといえた。

助言者の違いに影響を受ける認知過程

最後に、助言者に依ってErr_{J_2}が変化する事象について、提案した 3 過程の認知モデルのうち、activation step と integration step のどちらが有意に影響を受けているのか、Err_{advice}ごとに検討した結果を図 3(C-2, C-3)に示す。利用したモデルは上記と同様である。この結果から、P_{change}は全てのErr_{advice}において、助言者が「以前この課題を解いた 100 人」であった場合に、「AI」であった場合と比較して有意に値が小さかった。この結果はErr_{J_2}と同様であった。その一方で、$(target + J_1)/2 - J_2$は全てのErr_{advice}において、助言者が「以前この課題を解いた 100 人」であった場合と「AI」であった場合との間で有意差を認めかった。この結果から、助言者が変わると、助言を確認して助言を利用するか決定する認知過程が影響を受け、最終的な人間の判断の正確性が変化する、というメカニズムが示唆された。

まとめ

本研究では、シミュレーションと行動実験を通じて、予測誤差を伴う助言を人間が利用する場合、助言の予測誤差が小さくなるほど、それを利用した人間の意思決定がより正確になるとは限らないことを明らかにした。また、助言者が変わると、助言を確認した上で助言を利用するか決定する認知過程が影響を受け、最終的な人間の判断の正確性が変化することが示唆された。今後は、他ドメインおよび数値推定に限らないタスクに実験を拡張した上で、最終的な人間の意思決定がより正確になる意思決定支援システムの開発につなげたい。

謝辞

本研究成果の一部は、国立研究開発法人新エネルギー・産業技術総合開発機構(NEDO)の委託業務(JPNP20006)、JST 未来社会創造事業の助成(JP19211284)の結果得られた。

文献

[1] Himmelstein, M. (2022). Decline, adopt or compromise? A dual hurdle model for advice utilization. *Journal of Mathematical Psychology*, 110, 102695.
[2] Himmelstein, M., & Budescu, D. V. (2023). Preference for human or algorithmic forecasting advice does not predict if and how it is used. *Journal of Behavioral Decision Making*, 36(1), e2285.
[3] Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision-making. *Organizational behavior and human decision processes*, 62(2), 159-174.
[4] Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes*, 83(2), 260-281.