人-XAIインタラクションにおけるユーザ意思決定モデルの検討

An Attempt to Model User Decision-Making in Human-XAI Interaction

福地庸介 [1*]　山田誠二 [1,2]
Yosuke Fukuchi[1]　Seiji Yamada[1,2]

[1] 国立情報学研究所
[1] National Institute of Informatics
[2] 総合研究大学院大学
[2] SOKENDAI

Abstract: 説明可能 AI(XAI) 手法や大規模言語モデルの発展により、AI モデルの出力に関して多種多様な説明を生成できるようになっている。本研究の目標は、XAI から生成される説明のそれぞれがユーザの判断に与える影響を予測した上で適切な説明を選択することで、ユーザをより良い判断に導くことである。本稿ではこの目標に向けて、説明がユーザの判断に与える影響を予測するユーザモデルを構築し、モデルをもとにした説明選択手法を提案する。評価実験の結果、AI の性能が高い場面で提案する説明選択手法がユーザをより良い判断に導くことができる可能性が示唆された。

1　はじめに

人工知能 (AI) 手法をベースとする知的意思決定支援システム (IDSS) [1, 2] は、支援に関する説明性の導入によってユーザをより良い意思決定に導くことができる。これまで多くの説明可能 AI（XAI）手法が開発されており [3, 4, 5, 6]、また、大規模言語モデル（LLM）の発展によって AI の予測を正当化する様々な説明を事後的に生成することも可能になっている。従来研究ではこうした手法を IDSS に組み込み、AI の予測とともに説明を提示することの有効性が示されている [7, 8, 9]。

IDSS が様々な説明を利用できるようになる中、「多くの説明の候補のうち IDSS はどれをユーザに提示すべきか」という説明の選択が次の課題になると考える。説明の理解不能性 [10, 11, 12]、情報過多 [13, 14]、文脈との不適合 [15] など様々な原因がユーザの意思決定に悪影響を与える可能性が指摘されている。また、言語的な説明のニュアンスの微妙な違いも、文脈やタスクの状況、ユーザの認知的・心理的状態次第で、時にユーザの意思決定をミスリードする可能性がある。逆に、IDSS が状況を考慮しながら戦略的にユーザをより良い判断に導く説明を選択できれば、IDSS による意思決定支援をさらに発展させることができると期待できる。

本稿では、IDSS が提示する説明を動的に選択する手法を提案する。提案手法の特長は、説明がユーザの意思決定にどのような影響を与えるかを各試行ごとに予測し、予測結果をもとにユーザを AI が推奨する判断に導こうとする点である。提案手法を評価するため、本稿では XAI ベースの IDSS を用いたユーザスタディについても報告する。ここでは、ALL（可能なすべての説明を提示する）、ARGMAX（AI が最も可能性の高い予測に対する説明のみを提示する）という 2 条件の結果を、提案手法によって選択された説明の結果と比較した。結果からは、(i) AI の精度が高い場合は ARGMAX 戦略が、精度が低い場合は ALL がユーザをより良い判断に導くこと、(ii) AI の精度が高い場合に提案手法は ARGMAX 戦略を上回るが、AI の精度が低いときは ALL と同程度であることが示唆された。

2　人-XAIインタラクション

本研究が究極的に目指すのは、ユーザが AI を適切に利用するためのインタラクションをデザインすることである。AI に対する人間の過信や不信を避けることは、人と AI のインタラクションにおける基本的な問題である。ここで過信とは、人が AI の能力を過大評価し、その判断に盲目的に従ってしまっている状態であり、逆に不信とは、AI の能力を過小に評価することで人が AI を活用できていない状態を指す。

一般に、XAI が生成する説明を提示することで、人は AI の予測をより正しく理解できるようになり、適

*連絡先：国立情報学研究所
所属機関住所: 東京都千代田区一ツ橋２丁目１－２
E-mail: fukuchi@nii.ac.jp

切な AI 利用につながると考えられる。しかし、先行研究では、内容や状況によっては XAI による説明が時にネガティブな影響をもたらすことが示されている。Maehigashi et al. は、AI のサリエンシーマップの提示が、タスクの難易度やサリエンシーマップの理解可能性によって信頼に異なる影響を与えることを示した [10]。Herm は、XAI の説明のタイプがユーザの認知負荷、タスクパフォーマンス、タスク時間に強く影響することを明らかにした [14]。Panigutti et al. は、ユーザが説明の質を低いと感じていても、ユーザが説明に影響されることを発見した [16]。これらの結果は、説明が AI に対する不信を誘発したり、逆に説明から導かれる結論が誤りだとしても、ユーザが AI の予測に盲目的に従ってしまうリスクを示唆している。

本研究では、ユーザが IDSS からの説明に影響されながらする判断を計算機的にモデル化し説明がユーザに与える影響を予測することで、より良い説明を選択できるようにすることを試みる。Wiegreffe et al. [17] は、LLM によって生成された説明に対するユーザの容認度を予測することによって説明を評価する手法を提案している。このアプローチは、ユーザによる説明の知覚をモデル化し、モデルによる予測を説明の選択に利用するという点で本研究と同様のコンセプトを共有している。一方、本研究の関心は説明に対するユーザの知覚でなく、結果として説明がユーザの判断をどのように変化させ、それがタスクのパフォーマンスにどう影響するか、という点にある。Pred-RC [18, 19, 20] は、AI にタスクを割り当てるかどうかのユーザの判断に AI 性能に関する説明が与える影響を予測することで、ユーザの過信や不信を防止するような説明を選択する手法である。本研究では、タスクの割り当てという 2 値判断からさらに踏み込んで、説明がユーザの具体的な判断に与える影響を予測し、説明を通じてユーザに積極的に働きかけることでタスクのパフォーマンスを向上させることを目指す。

3 提案手法

本稿では、状況に応じて IDSS が提示する説明を選択する手法を提案する。手法の基本的なアイデアは、可能な説明の組み合わせのそれぞれでユーザの判断を予測し、ユーザの判断と AI が推奨する判断の距離が最小となるように説明の組を選択するというものである。

U は、説明に影響されながら判断を行うユーザのモデルである。

$$U(\boldsymbol{c}, \boldsymbol{x}, d_u) = P(d_u|\boldsymbol{c}, \boldsymbol{x}). \tag{1}$$

U は、$\boldsymbol{c}$ と $\boldsymbol{x}$ の条件下でのユーザの判断 d_u を確率分布として表現する。ここで、$\boldsymbol{x}$ は IDSS がユーザに選択的に提示する情報 (説明) の組み合わせ、$\boldsymbol{c}$ は、AI の予測、タスクの状態、ユーザの状態など、それ以外の全ての文脈情報を表す。

加えて、AI の推奨する判断 d_{AI} を出力する方策 π を考える。π は、$\boldsymbol{c}$ に基づいて d_{AI} を決定する。この推論はユーザの意思決定と並行して行われる：

$$\pi(\boldsymbol{c}, d_{\mathrm{AI}}) = P(d_{\mathrm{AI}}|\boldsymbol{c}). \tag{2}$$

U によって予測される各 $\boldsymbol{x}$ が d_u に与える影響と π が決定する d_{AI} を比較することで、d_u と d_{AI} の距離を最小化する $\hat{\boldsymbol{x}}$ を予測することができる。

$$\hat{\boldsymbol{x}} = \mathrm{distance}(U(\boldsymbol{c}, \boldsymbol{x}, d), \pi(\boldsymbol{c}, d)). \tag{3}$$

提案手法は試行毎に $\hat{\boldsymbol{x}}$ を計算しユーザに提示することで、判断をより良いものに導くことを目指す。

4 実験

4.1 タスクと実装

IDSS の支援付株式取引シミュレータに提案手法を実装し評価を行った。図 1 にシミュレータのスクリーンショットを示す。シミュレーションはウェブサイト上で行われた。参加者は仮想的に 300 万円を与えられ、株価チャート、AI による将来の株価予測、予測の説明を見ながら 60 日間の株取引を行う。

シミュレーションでは、参加者は各日の始値と株価チャートを確認し、手持ちの資金で株を買うか、持っている株を売るか、ポジションを持ち続けるかを決定した。説明の影響を明確にするため、参加者は 1 日に 2 回判断を示すよう求めた。まずユーザは、IDSS のサポートなしにチャート情報のみで初期の注文 d' を決定した。次に、IDSS は株価予測モデルの出力を表現する棒グラフとその説明を示した。ユーザの意思決定の自律性を高めるため、d_{AI} は明示的に示さなかった[1]。最後に、IDSS からの支援を踏まえた最終的な判断 d を入力する。この後、シミュレータは直ちに翌日に移行する。最終日に持ち越したポジションは、翌日からの 5 日間の平均株価をもとに現金化し、参加者のトータル・パフォーマンスを計算した。

ユーザの支援を行う機械学習モデルとして株価予測モデルを開発した。このモデルは、株価のチャートを入力として今後 5 営業日の平均株価を予想し、上昇 (+2% 以上)、中立 (-2%〜+2%)、下落 (-2%以下) の 3 クラスに分類する。さらに、予測された株価の変動を入力として d_{AI} を出力する π を強化学習によって実装した。

[1] d_{AI} を $\boldsymbol{c}$（常に d_{AI} を表示する場合）または $\boldsymbol{x}$（選択的に d_{AI} を表示する場合）に含めることで、d_{AI} をユーザに与える設定には容易に拡張できる。

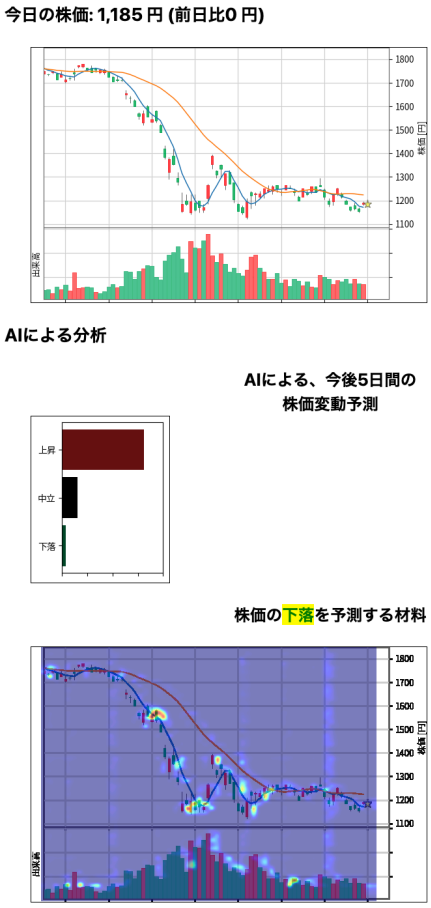

図 1: タスクのインタフェースの一部

ユーザに提示する説明の候補として、サリエンシーマップと文の 2 種類を用意した。サリエンシーマップは Score-CAM [21] によって生成した。この手法は、予測クラスごとにサリエンシーマップを生成できるため、各取引日につき 3 つのマップを得た。また、画像を入力できる Open-AI API の GPT-4V [22] によって、文による説明を作成した。GPT-4V には、各予測クラス（上昇・中立・下落）を正当化する説明文を各 2 つ生成するよう求めるプロンプトを、チャートの画像とともに入力した。したがって、各チャートに対して合計 6 つの文を得た。結果として、各取引日につき 3 枚のサリエンシーマップと 6 つの文による説明が得られ、提案手法は $2^9 = 512$ 通りの組み合わせから説明を選択した。

株価予測モデルの性能が結果に影響すると事前に予想したため、精度が 0.75 の High-accuracy シナリオと 0.33(チャンスレベル) の Low-accuracy シナリオの 2 シナリオで実験を行った。

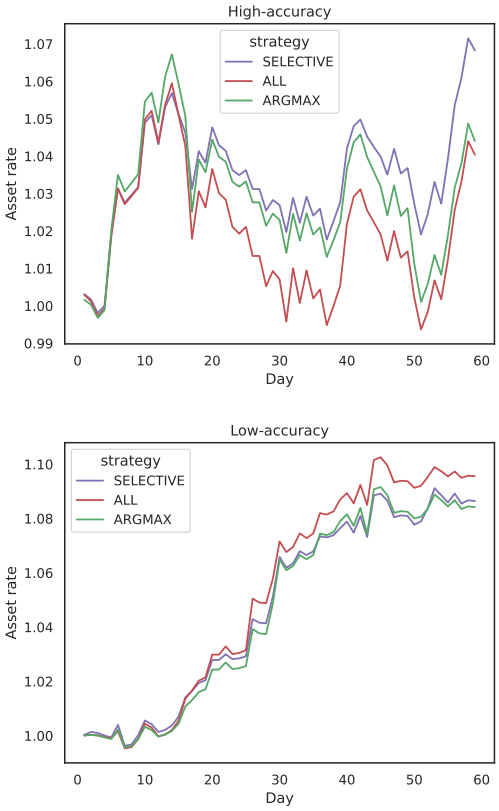

図 2: 資産合計の推移

4.2 結果

図 2 にユーザの総資産額の推移を示す。SELECTIVE は提案手法によって説明を選択した結果、ALL は可能なすべての説明を提示した結果、ARGMAX は AI が最も可能性が高いと予測するクラスに対する説明のみを提示した結果を示す。

ALL と ARGMAX は、シナリオ間で異なる結果となった。High-accuracy の場合、ARGMAX が ALL を上回った。これは、ARGMAX では精度の高い株価予測モデルの予測を説明が支持する形となり、ユーザをより良い判断に導いた一方、ALL はそれ以外の説明も提示することでユーザの判断がより中立的になり、結果として ALL よりも劣る結果になったと考えられる。一方、Low-accuracy の場合は ALL がユーザに多角的な視点を与えることがポジティブに機能し、ARGMAX を上回る結果となった。

High-accuracy シナリオでは、提案手法は概ね ALL と ARGMAX を上回る成績となった。より詳細には、提案手法は最初 ARGMAX を下回ったが、16 日目にスコアが逆転し、その差は 39 日目付近で一旦縮まったが、最終日まで再び広がった。16 日目付近の株価は急落したため、IDSS は参加者にポジションの反転に導く

必要があった。ここで、ARGMAX が株価の下落の予測を説明したのに対し、提案手法では下落だけでなく中立の説明も示しており、これがユーザを正しく株売りに誘導した可能性がある。

しかし、Low-accuracy シナリオでは提案手法のスコアは ARGMAX と同程度であり、ALL を下回った。この理由として、株価予測モデルの精度の低さが π の性能に悪影響を与えたことが考えられる。結果、ユーザの判断をガイダンスする先である d_{AI} の性能が悪化するため、ユーザを却ってミスリードしてしまった可能性がある。今後の方向性としては、ユーザを $d_{mathrmAI}$ に導くことだけを目的とするのではなく、AI の性能に基づいてユーザの判断を誘導する際の強度を調整することが考えられる。

5 結論

説明によってユーザをより良い判断に導く IDSS の実現に向けて、本稿では XAI から生成される説明群から説明を選択する手法を提案した。提案手法は、説明がユーザの意思決定にどのような影響を与えるかを各試行ごとに予測し、予測結果をもとにユーザを AI が推奨する判断に導こうとする。IDSS による支援付きの株取引シミュレーションによる評価の結果、AI の精度が高い場合には提案手法がユーザをよりよい判断に導くことができること、AI の精度が低い場合には誘導の強度を調整する等の課題があることが示唆された。

謝辞

本研究は、JST CREST (JPMJCR21D4)・AIP チャレンジプログラムの支援を受けた。

参考文献

[1] Gloria Phillips-Wren. *Intelligent Decision Support Systems*, pages 25–44. 02 2013.

[2] Marianne Cherrington, Zhongyu (Joan) Lu, Qiang Xu, David Airehrour, Samaneh Madanian, and Andrea Dyrkacz. Deep learning decision support for sustainable asset management. In Andrew Ball, Len Gelman, and B. K. N. Rao, editors, *Advances in Asset Management and Condition Monitoring*, pages 537–547, Cham, 2020. Springer International Publishing.

[3] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[4] Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.

[5] Bilal Alsallakh, Allan Hanbury, Helwig Hauser, Silvia Miksch, and Andreas Rauber. Visual methods for analyzing probabilistic classification data. *IEEE TVCG*, 20(12):1703–1712, 2014.

[6] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE TNLS*, 28(11):2660–2673, 2016.

[7] Min Hun Lee and Chong Jun Chew. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2), oct 2023.

[8] Dimitrios P. Panagoulias, Elissaios Sarmas, Vangelis Marinakis, Maria Virvou, George A. Tsihrintzis, and Haris Doukas. Intelligent decision support for energy management: A methodology for tailored explainability of artificial intelligence analytics. *Electronics*, 12(21), 2023.

[9] Devleena Das, Been Kim, and Sonia Chernova. Subgoal-based explanations for unreliable intelligent decision support systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 240–250, New York, NY, USA, 2023. Association for Computing Machinery.

[10] Akihiro Maehigashi, Yosuke Fukuchi, and Seiji Yamada. Modeling reliance on xai indicating its purpose and attention. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, pages 1929–1936, 2023.

[11] Akihiro Maehigashi, Yosuke Fukuchi, and Seiji Yamada. Empirical investigation of how robot ' s pointing gesture influences trust in and acceptance of heatmap-based xai. In *2023 32nd IEEE International Conference on Robot and Human*

Interactive Communication (RO-MAN), pages 2134–2139, 2023.

[12] Akihiro Maehigashi, Yosuke Fukuchi, and Seiji Yamada. Experimental investigation of human acceptance of ai suggestions with heatmap and pointing-based xai. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, HAI '23, page 291–298, New York, NY, USA, 2023. Association for Computing Machinery.

[13] Aidah Nakakande Ferguson, Matija Franklin, and David Lagnado. Explanations that backfire: Explainable artificial intelligence can cause information overload. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.

[14] Lukas-Valentin Herm. Impact of explainable ai on cognitive load: Insights from an empirical study. In *The 31st European Conference on Information Systems*, 2023. 269.

[15] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. Automated rationale generation: A technique for explainable ai and its effects on human perceptions. In *Proc. 24th Int. Conf. IUI*, page 263–274, 2019.

[16] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. Understanding the impact of explanations on advice-taking: A user study for ai-based clinical decision support systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[17] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States, July 2022. Association for Computational Linguistics.

[18] Yosuke Fukuchi and Seiji Yamada. Selectively providing reliance calibration cues with reliance prediction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, pages 1579–1586, 2023.

[19] Yosuke Fukuchi and Seiji Yamada. Dynamic selection of reliance calibration cues with ai reliance model. *IEEE Access*, 11:138870–138881, 2023.

[20] Yosuke Fukuchi and Seiji Yamada. Selective presentation of ai object detection results while maintaining human reliance. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3527–3532. IEEE, 2023.

[21] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[22] OpenAI. Gpt-4 technical report, 2023.