

On the Effect of Information Asymmetry in Human-AI Teams

Patrick Hemmer*

Karlsruhe Institute of
Technology
Karlsruhe, Germany
patrick.hemmer@kit.edu

Max Schemmer*

Karlsruhe Institute of
Technology
Karlsruhe, Germany
max.schemmer@kit.edu

Niklas Kühl

Karlsruhe Institute of
Technology
Karlsruhe, Germany
niklas.kuehl@kit.edu

* denotes equal contribution

Michael Vössing

Karlsruhe Institute of
Technology
Karlsruhe, Germany
michael.voessing@kit.edu

Gerhard Satzger

Karlsruhe Institute of
Technology
Karlsruhe, Germany
gerhard.satzger@kit.edu

Abstract

Over the last years, the rising capabilities of artificial intelligence (AI) have improved human decision-making in many application areas. Teaming between AI and humans may even lead to complementary team performance (CTP), i.e., a level of performance beyond the ones that can be reached by AI or humans individually. Many researchers have proposed using explainable AI (XAI) to enable humans to rely on AI advice appropriately and thereby reach CTP. However, CTP is rarely demonstrated in previous work as often the focus is on the design of explainability, while a fundamental prerequisite—the presence of complementarity potential between humans and AI—is often neglected. Therefore, we focus on the existence of this potential for effective human-AI decision-making. Specifically, we identify information asymmetry as an essential source of complementarity potential, as in many real-world situations, humans have access to different contextual information. By conducting an online experiment, we demonstrate that humans can use such contextual information to adjust the AI's decision, finally resulting in CTP.

Author Keywords

Human-AI Teams; Complementary Team Performance;
Human-AI Complementarity; Information Asymmetry

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

Copyright held by the owner/author(s).

CHI Conference on Human Factors in Computing Systems (CHI '22), Workshop on Human-Centered Explainable AI (HCXAI), May 12–13, 2022, New Orleans, LA, USA

Introduction

The rising capabilities of artificial intelligence (AI) have paved the way for supporting human decision-making in a growing number of domains [26, 28, 30]. To offer humans meaningful support, particularly in high-stake settings, AI models are not only expected to provide accurate predictions but also a notion about how a decision was derived or how confidently it was made to foster humans' understanding. This idea fueled the development of techniques from the field of explainable AI (XAI) [1]. Its intention is to enable domain experts to assess when to rely on AI advice to improve decision-making performance [2, 3, 31].

Ideally, this form of XAI-assisted decision-making achieves complementary team performance (CTP)—a task performance that surpasses both human and AI performance when conducting the task alone. However, current research reveals that achieving CTP is challenging [3, 11]. Most studies show that XAI-assisted decision-making yields higher team performance than humans conducting the task alone. Still, this performance is often inferior to the one of the AI alone [3, 11], leaving the question unanswered why CTP could not have been accomplished.

A possible explanation for this observation may be that in order for human-AI decision-making to result in CTP, a more fundamental prerequisite is the presence of sufficient complementarity potential (CP) between humans and AI. In this context, we hypothesize that a source of CP emerges from unique human contextual information (UHCI). In practice, domain experts often have access to further information not available to the AI during training as not all data might be digitally available due to technical or economic reasons. Thus, we investigate whether humans' decision-making benefits from the presence of UHCI when receiving AI assistance.

We conduct an online experiment within the domain of real estate appraisal. We employ an AI model that predicts real estate prices and provides an uncertainty estimate solely based on tabular data. Humans have additional access to a corresponding picture of the house and, thus, are equipped with UHCI. Our results demonstrate that the presence of UHCI can enable humans to adjust AI predictions resulting in a task performance that surpasses the one of humans and AI alone, i.e., CTP. From this finding we can derive several implications for future XAI research.

In general, sufficient CP might constitute a requirement for effective XAI-assisted decision-making. Therefore, researchers need to investigate the mutual effects between CP and XAI. On the one hand, CP may positively influence XAI-assisted decision-making. For example, UHCI may activate analytical instead of intuitive thinking and thereby could indirectly trigger conscious engagement with explanations which improves team performance. On the other hand, XAI can also amplify the effect of CP. For example, feature importance can be used to detect whether the individual perceived UHCI is really unique or also taken into account by the AI. In future work, we aim to formalize the notion of CP, evaluate the impact of different XAI types within our specific study setup, and further assess how humans can learn to rely on AI advice appropriately.

Related Work

In line with the continuous development of XAI algorithms [1], a growing body of research has started to investigate their effect on task performance in AI-assisted decision-making scenarios in online experiments. A popular idea is to enable humans to question the AI's decisions through insights about the uncertainty of the prediction [3, 9, 31] or through explanations that aim to shed light on the AI's decision-making [13, 17]. In this context, studies analyze



Figure 1: An overview of the interface that contains the information provided to the participants in the online experiment. The *UHCI treatment* provides all available information to the human (a). In the *no UHCI treatment*, the image is withheld (b).

the effects of different XAI techniques, ranging from feature-based [5, 18, 21] over example-based [19, 29] to rule-based [2, 22] explanations, and also consider an entire spectrum between full human agency and full automation [14]. Even though these studies reveal that human decision-making generally benefits from this algorithmic support, the combined human-AI performance usually remains inferior to the AI conducting the task alone, as humans struggle to anticipate when the AI provides correct and incorrect advice [3, 24]. As the development of these assistance methods has so far been predominantly driven by an algorithmic perspective [8], necessary prerequisites from a human-centered perspective that contribute to enabling CTP have been underexplored. Consequently, researchers have recently started to argue for placing the human at the center of technology design [7, 16]. Thus, with our work, we aim to contribute to identifying essential elements in the interplay between humans and AI that have to be considered to enable effective decision-making.

Methods

We select a house prediction data set consisting of tabular data and house images [27] for the online experiment. The data set consists of 15,474 instances, of which we allocate 80% to the training and 20% to the test set. Additionally, we draw a hold-out set of 15 properties from the test split as the samples for our experiment. We train an AI model—a random forest regression—only on the tabular features street, city, number of bedrooms, number of bathrooms, and square footage of the house. The image of the property is withheld from the AI model. It achieves a performance measured as the mean absolute error (MAE) of \$163,080 on the hold-out set, which is comparable to that on the entire test set. Additionally, based on the individual trees of the random forest, we generate a predictive distribution and

display the 5% and 95% quantile as indicators for AI uncertainty.

The experiment consists of two treatments. In the first, participants are provided with information about each property's street, city, number of bedrooms, number of bathrooms, and square footage (*no UHCI treatment*). In the second treatment, they are additionally provided with an image of the property (*UHCI treatment*). With this image, they receive additional contextual information compared to the participants of the first treatment. Figure 1 displays the information provided in the respective treatments. We recruited participants via prolific.co and randomly assigned them to one of the two treatments.

Before the actual set of tasks, participants in both treatments had to undergo an in-depth introduction to the data set and the task [13, 20], including summary statistics about the properties' prices, followed by a question to verify their understanding. Additionally, we stressed that the AI did not have access to the image during training. After informing participants about the start of the actual decision-making task, the study procedure was as follows for each of the 15 instances: first, they were asked to provide a prediction on their own to prevent them from entering a state of low cognitive activation [10]. Consequently, they received the AI's prediction together with its confidence estimate. Then, participants were asked to adjust the prediction of the AI in the best possible way. Finally, participants were asked to fill out a questionnaire to collect demographics after completing all instances. In general, participants received a base payment of 5 pounds with the incentive that the best 10% would receive an additional pound. The whole task lasts approximately 30 minutes.

In total, we recruited 120 participants. To ensure the quality of the collected data, we removed participants enter-

ing house prices higher than the communicated maximum property price in the data set of \$2,000,000. Additionally, we identified outliers for removal using the median absolute deviation [15, 23]. After applying these criteria, we collected the data from 101 participants over both conditions, of which 53 were in the *no UHCI treatment* and 48 in the *UHCI treatment*.

Results And Discussion

In Figure 2, we display the human and the AI-assisted performance for both conditions. We evaluate the significance of the results using the Student’s T-tests with Bonferroni correction. Its prerequisites have been verified ex-ante. Participants conducting the task alone in the *no UHCI treatment* achieve a MAE of \$251,282, while the test persons in the *UHCI treatment* yield a MAE of \$200,510. We observe a significant difference between both conditions without and with UHCI of \$50,772 ($t = 4.6118, p < 0.001$).

Looking at the team performance after adjusting the AI’s prediction, we find that the human-AI team in the *no UHCI treatment* achieves a MAE of \$160,095. In contrast, the human-AI team in the *UHCI treatment* yields a performance in terms of MAE of \$148,009—a reduction of \$12,086. This performance improvement turns out to be significant on the 0.05 level ($t = 2.9571, p = 0.0155$). In both treatments, the human-AI teams outperform the AI alone (MAE: \$163,080). While the difference between the performance of the human-AI team in the *UHCI treatment* is significant ($t = -4.6798, p < 0.001$), the difference in the *no UHCI treatment* does not result in a significant improvement ($t = -1.1596, p = 0.99$). To summarize, we find that in the presence of UHCI, humans become capable of positively adjusting the AI predictions resulting in CTP.

Regarding the general potential of human-AI teaming, our

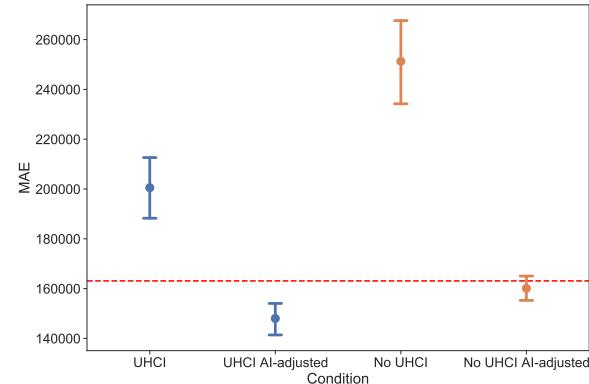


Figure 2: Performance results as MAE of the online experiment across conditions including 95% confidence intervals. The red horizontal line denotes the AI performance.

finding validates the results from [3, 6, 12] by reaching CTP in an experimental study. Moreover, it has several implications for future research on human-AI decision-making in general and XAI-assisted decision-making in specific. For example, as humans tend to become more capable of correctly adjusting AI advice, they might also be able to better question additional information beyond sole confidence estimates, e.g., different explanations.

Future work should systematically identify additional sources of CP. From a human-centered perspective, not only information asymmetry but also skill differences could play a decisive role. Moreover, we hypothesize that CTP depends not only on CP but also on how well humans can utilize it in the decision-making process. Thus, human-centered design mechanisms to effectively combine AI and human

decisions are needed to foster appropriate reliance on AI advice. Prior research on XAI has shown that a major challenge of XAI is the issue of over-trust [3, 4, 25]. Therefore, XAI needs to be designed taking a human-centered view to enable appropriate reliance and not solely increase trust. Additionally, future research needs to investigate the mutual effects between different XAI techniques and CP. In future work, we aim to formalize the notion of CP and conduct additional experiments to investigate the effect of XAI on appropriate reliance in the presence of CP.

REFERENCES

- [1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.
- [2] Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantacioglu. 2021. Does Explainable Artificial Intelligence Improve Human Decision-Making?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6618–6626.
- [3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. 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*. 1–16.
- [4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.
- [5] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.
- [6] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248* (2020).
- [7] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.
- [8] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing human-centered perspectives in explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.
- [9] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. 2021. Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI. *Management Information Systems Quarterly* 45 (2021).
- [10] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
- [11] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *Proceedings of the 25th Pacific Asia Conference on Information Systems* (2021).

- [12] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.
- [13] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [14] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.
- [15] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* 49, 4 (2013), 764–766.
- [16] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).
- [17] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.
- [18] Sina Mohseni, Fan Yang, Shiva Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji, and Eric Ragan. 2021. Machine Learning Explanations to Prevent Overtrust in Fake News Detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 421–431.
- [19] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. 2021. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems* 34 (2021).
- [20] Marcus O'Connor, William Remus, and Ken Griggs. 1993. Judgemental forecasting in times of change. *International Journal of Forecasting* 9, 2 (1993), 163–172.
- [21] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.
- [22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [23] Peter J Rousseeuw and Christophe Croux. 1993. Alternatives to the median absolute deviation. *J. Amer. Statist. Assoc.* 88, 424 (1993), 1273–1283.
- [24] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022a. Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making. *arXiv preprint arXiv:2204.06916* (2022).

- [25] Max Schemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022b. On the Influence of Explainable AI on Automation Bias. *arXiv preprint arXiv:2204.08859* (2022).
- [26] Julian Senoner, Torbjørn Netland, and Stefan Feuerriegel. 2021. Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Science* (2021).
- [27] TED8080. 2019. House prices and images - SoCal. (Date accessed: August, 01 2021).
<https://www.kaggle.com/ted8080/house-prices-and-images-socal>
- [28] Alexander Treiss, Jannis Walk, and Niklas Kühl. 2020. An uncertainty-based human-in-the-loop system for industrial tool wear analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 85–100.
- [29] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021).
- [30] N. Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanislaw Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, S. Wolfson, Ujas Parikh, Sushma Gaddam, L. Lin, Kara Ho, Joshua D. Weinstein, B. Reig, Yiming Gao, H. Toth, Kristine Pysarenko, A. Lewin, Jiyon Lee, Krystal Airola, E. Mema, Stephanie Chung, Esther Hwang, N. Samreen, S. Kim, L. Heacock, L. Moy, Kyunghyun Cho, and K. Geras. 2020. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging* 39 (2020), 1184 – 1194.
- [31] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.