Outsmarting the AI: When do Explanations Help Human Users?

Bachelor/Master project "Information Systems"

Table of Contents

Prof. Dr. Mario Nadj
University of Duisburg-Essen
Faculty of Computer Science

Chair and Website
Information Systems and AI in Marketing
ai.ris.uni-due.de

Address Campus Essen
Universitätsstraße 9
D-45141 Essen

www.uni-due.de

1: Project Overview

Background: A recurring problem in studies involving human users equipped with artificial intelligence (AI) assistance is that they rarely perform better than the AI by itself. Naturally, this begs the question whether complementarity between human and AI is possible at all, and which conditions might help humans to better understand the AI and use its outputs in a more informed, performant way. Such conditions could be provided, for instance, through means of explainable AI (XAI) augmentations.

Project Goal: Develop and pretest a user study on human performance when using AI assistance and XAI augmentations. Conduct and evaluate the experiment, to investigate whether humans can reach complementary performance with AI.

Highlights: Train AI models; develop the experimental interface from scratch; conduct an interesting and highly relevant experiment; freedom to pursue individual solutions; close cooperation with the chair team

2: Project Background

When human and AI interact to make decisions, the ultimate goal is to surpass whatever performance either one achieves on its own (Bansal et al., 2021). In practice, however, humans tend to exhibit too much reliance on the AI (Bertrand et al., 2022). Hence, they mostly accept what the AI proposes, adding little critical oversight to the process.

An aspect to this problem that has seen little scientific scrutiny is, how overlap between the AI's and the human's skills shapes these results. In most studies, the AI performs better than the human by itself, but it is not clear whether AI performs well on different instances than the human, or whether they succeed mostly at similar instances. This has important implications for human-AI complementarity. If humans get many instances right by themselves, but fail at the same ones when relying on AI, there is more potential to reach complementary performance, than there is if human and AI largely fail and succeed at the same instances.

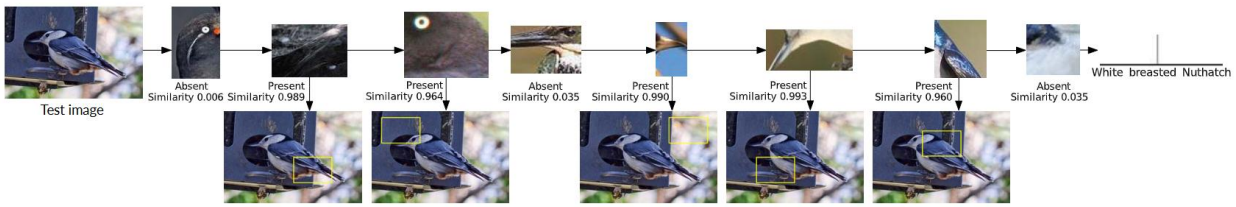

Figure 1: Example classification made on a CUB-200 bird (Nauta et al., 2021).

Furthermore, the effect of XAI on human reliance behavior remains unclear. While XAI has been shown to potentially increase overreliance, for instance due to persuading users of the AI's correctness, this has not been investigated with the different skills of human and AI in

mind. This project will look at the effect that combining contrastive examples with interactive XAI has on human-AI task performance (Miller, 2019, Bertrand et al., 2023). For this purpose, the group will use the CUB-200 bird classification dataset (Welinder et al., 2010).

3: Project Objective

The objective of this specific Bachelor/Master project is to investigate how different XAI methods affect user reliance behavior, with special consideration given to baseline differences in performance between human and AI. To achieve this, the project is split into two phases.

First, the participating students will assess how human and AI differ not only in terms of their overall performance, but also on the instance level. They will train two different AI models on the CUB-200 dataset, and in parallel conduct a pretest with fellow students to determine which instances human and AI get correct respectively. Second, based on this comparison, the group will conduct a user experiment with 20 to 40 participants to determine the effect that enabling interaction with contrastive examples has on user reliance behavior. During the experiment, eye-tracking data will be collected in addition to task data to analyze which parts of the explanatory information were actively considered during decision making. The objectives of this specific project are the following:

- **Pretesting** of instance-level performance of AI models and human participants
- **Implementing** the experimental setup based on the pretesting results
- **Conducting** and analyzing the experiment

Depending on group size, adjustments to the project scope can be made. This will be announced at the kickoff and in subsequent milestone meetings.

Each finished artifact has to be properly documented to facilitate understanding and further development by students in follow-up projects.

4: Project Registration and further Information

All Bachelor- and Master students interested in this IS-project may apply via our online application form. The link to the application will be **available here** (Link hinterlegt) as of the start of the application period on **Monday, 28.09.2026.** The deadline for the application is **Sunday, 11.10.2026, 23:59.**

For successful completion of the project, Master students receive 12 ECTS (corresponding to 360 working hours) and Bachelor students receive 6 ECTS (corresponding to 180 working hours)

- Master students will invest 18 hours per week into the project.
- Bachelor students will invest 9 hours per week into the project.

In addition to project-specific deliverables during the process, we also require some fixed deliverables from each group. On the 04.12.2026, the project team needs to submit a one-page management summary on the current state of the project. A template will be provided by the Chair. As final deliverables, the project team has to submit four items by 28.02.2027. These are (1) a 5-minute video presenting the results (video for orientation), (2) a scientific poster summarizing the results (whether the poster will be considered for presentation will be determined by our Chair, and a final centralized vote among the Information Systems Chairs), and (3) a project report along with the (4) project code. The project report will contain the final results on 20 to 50 pages (Bachelor) or 30 to 60 pages (Master). The exact number of pages depends on group size and the number of bachelor and master students and will be communicated at kickoff. A template is provided on the WIPRO page. The (4) code needs to be accessible as Open Source via GitHub with appropriate documentation. We want to put up the best posters of each semester (2) at the Chair's premises to highlight your achievements. For outstanding projects, the final video (1) and a link to your code (4) will be shown on the Chair's page.

5: Project Schedule (Dates are to be announced)

Deliverables for each milestone are marked in blue.

6: Grading

The final grade will be calculated as follows based on the sum of the milestones:

- Milestone 1: 10%
- Milestone 2: 5%
- Milestone 3: 10%
- Milestone 4: 5%
- Milestone 5: 30%
- Milestone 6: 40%

7: Sources

- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. & Weld, D. (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 (pp. 1-16).
- Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How cognitive biases affect XAI-assisted decision-making: A systematic review. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 78-91).
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, 1-38.
- Bertrand, A., Viard, T., Belloum, R., Eagan, J. R., & Maxwell, W. (2023, April). On selective, mutable and dialogic xai: A review of what users say about different types of interactive explanations. In Proceedings of the 2023 CHI conference on human factors in computing systems (pp. 1-21).
- Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., & Perona, P. (2010). Caltech-UCSD birds 200.
- Nauta, M., Van Bree, R., & Seifert, C. (2021). Neural prototype trees for interpretable fine-grained image recognition. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 14928-14938). IEEE.