APPENDIX: COMPLEMENTARITY IN HUMAN-AI COLLABORATION: CONCEPT, SOURCES, AND EVIDENCE

PREPRINT

Patrick Hemmer
Karlsruhe Institute of Technology
patrick.hemmer@kit.edu

Max Schemmer
Karlsruhe Institute of Technology
max.schemmer@kit.edu

Niklas Kühl
University of Bayreuth
kuehl@uni-bayreuth.de

Michael Vössing
Karlsruhe Institute of Technology
michael.voessing@kit.edu

Gerhard Satzger
Karlsruhe Institute of Technology
gerhard.satzger@kit.edu

A Additional Details on the Complementarity Potential and Complementarity Effect

In the following, we elaborate more comprehensively that Equation 6 is an alternative representation of Equation 3:

Equation 6:
$$CP = CP^{inh} + CP^{coll}$$

Case $L_{AI} \leq L_H$:

$$CP = \frac{1}{N}\sum_{i=1}^{N}\left(\max\left(0, l_{AI}^{(i)} - l_H^{(i)}\right) + \min\left(l_H^{(i)}, l_{AI}^{(i)}\right)\right)$$

$$CP = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} 0 + l_{AI}^{(i)}, & \text{if } l_{AI}^{(i)} \leq l_H^{(i)} \\ l_{AI}^{(i)} - l_H^{(i)} + l_H^{(i)}, & \text{if } l_{AI}^{(i)} > l_H^{(i)} \end{cases}$$

$$CP = \frac{1}{N}\sum_{i=1}^{N} l_{AI}^{(i)} = L_{AI}$$

Case $L_{AI} > L_H$:

$$CP = \frac{1}{N}\sum_{i=1}^{N}\left(\max\left(0, l_H^{(i)} - l_{AI}^{(i)}\right) + \min\left(l_H^{(i)}, l_{AI}^{(i)}\right)\right)$$

$$CP = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} l_H^{(i)} - l_{AI}^{(i)} + l_{AI}^{(i)}, & \text{if } l_{AI}^{(i)} \leq l_H^{(i)} \\ 0 + l_H^{(i)}, & \text{if } l_{AI}^{(i)} > l_H^{(i)} \end{cases}$$

$$CP = \frac{1}{N}\sum_{i=1}^{N} l_H^{(i)} = L_H$$

Combining both cases shows that CP equals the minimum of the overall individual losses of the human and the AI:

Equation 3:
$$CP = L_{T*} = \min\left(L_H, L_{AI}\right)$$

Next, we show that Equation 10 is an alternative representation of Equation 7:

Equation 10: $$CE = CE^{inh} + CE^{coll}$$

Case $L_{AI} \leq L_H$:

$$CE = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} l_{AI}^{(i)} - l_I^{(i)} + 0 & , l_{AI}^{(i)} > l_I^{(i)} \geq l_H^{(i)} \\ l_{AI}^{(i)} - l_H^{(i)} + l_H^{(i)} - l_I^{(i)} & , l_{AI}^{(i)} > l_H^{(i)} > l_I^{(i)} \\ 0 + l_{AI}^{(i)} - l_I^{(i)} & , l_H^{(i)} \geq l_{AI}^{(i)} > l_I^{(i)} \\ 0 + l_{AI}^{(i)} - l_I^{(i)} & , l_I^{(i)} > l_{AI}^{(i)} \\ 0 & , \text{otherwise} \end{cases}$$

$$CE = \frac{1}{N}\sum_{i=1}^{N}\left(l_{AI}^{(i)} - l_I^{(i)}\right)$$

$$CE = \frac{1}{N}\sum_{i=1}^{N} l_{AI}^{(i)} - \frac{1}{N}\sum_{i=1}^{N} l_I^{(i)}$$

$$CE = L_{AI} - L_I$$

Case $L_{AI} > L_H$:

$$CE = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} l_H^{(i)} - l_I^{(i)} + 0 & , l_H^{(i)} > l_I^{(i)} \geq l_{AI}^{(i)} \\ l_H^{(i)} - l_{AI}^{(i)} + l_{AI}^{(i)} - l_I^{(i)} & , l_H^{(i)} > l_{AI}^{(i)} > l_I^{(i)} \\ 0 + l_H^{(i)} - l_I^{(i)} & , l_{AI}^{(i)} \geq l_H^{(i)} > l_I^{(i)} \\ 0 + l_H^{(i)} - l_I^{(i)} & , l_I^{(i)} > l_H^{(i)} \\ 0 & , \text{otherwise} \end{cases}$$

$$CE = \frac{1}{N}\sum_{i=1}^{N}\left(l_H^{(i)} - l_I^{(i)}\right)$$

$$CE = \frac{1}{N}\sum_{i=1}^{N} l_H^{(i)} - \frac{1}{N}\sum_{i=1}^{N} l_{AI}^{(i)}$$

$$CE = L_H - L_I$$

Combining both cases shows that CE equals the difference between the minimum of the overall individual losses and the team loss:

Equation 7: $$CE = min(L_H, L_{AI}) - L_I$$

B Experiment 1: The Effect of Information Asymmetry

B.1 In-person Pilot Study

For our in-person pilot study, we conducted two workshops. Each one lasted 90 minutes and aimed to elaborate on the usefulness of the house images for the participants and generate insights for the experimental design. The participants were 13 interviewees and two researchers. The first workshop was conducted with a smaller group to facilitate a more extensive exchange. The second workshop focused on collecting broad ideas. We discussed three task instances in each session. For each house, we collected feedback in a structured way by asking the participants to make a prediction, once alone and once together with the AI. In this context, they were also asked to make notes about how the AI suggestions and the additional information supported their decision-making. In general, both workshops confirmed the usefulness of unique contextual information. We also received helpful comments for further refinement of our study. In the first workshop, a participant mentioned that the "picture was the first indication for me", and another one stated, "I was already 70% sure what I was going to type when I saw the picture." Both comments indicate the importance that participants see in the unique contextual information. In the second workshop, participants highlighted the need for more information about the underlying data by stating, "show the user the summary statistics." They also mentioned the importance of a training section, for example, by saying, "I need examples of wrong (and right) predictions."

B.2 Experiment Details

Figure 1 displays the procedure of the experiment. Moreover, Figure 2 highlights the participant user interface for the task tutorial and Figure 3 depicts the AI tutorial in the experiment. Figures 4 and 5 display the training examples. Lastly, Figures 6 to 20 display the 15 task instances of the experiment.

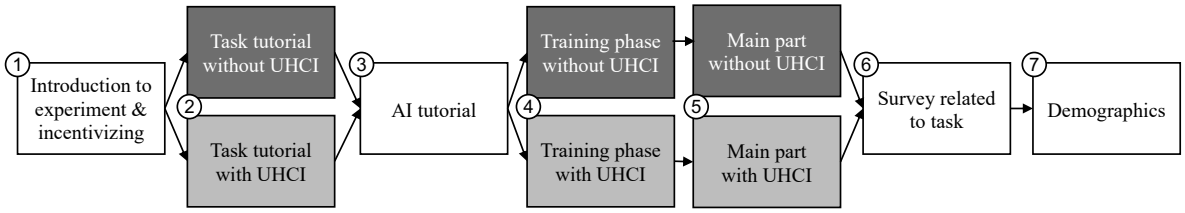


Figure 1: Sequence of the individual steps in the experiment (UHCI = unique human contextual information).

Introduction to house price prediction

In the following you will be asked to predict house prices.

Therefore, we provide you with data and images of 15 houses.

The houses are all located in Southern California and the data was collected in 2019.

The mean house price is 703,120 $.

The least expensive house in the data set costs 195,000 $.

The most expensive house in the data set costs 2,000,000 $.

For our participants who are more familiar with the metric system, 10 square feets are approximately one square meter.

An example for house data that we will provide you is the following:

Street:	Wheatville St
City:	Chula Vista, California (USA)
Number of Bedrooms:	4
Number of Bathrooms:	2
Square Feet:	1873

Your task will be to predict the house price based on the data in the table and the image.

Figure 2: Task tutorial. In the condition without UHCI the image is not displayed.

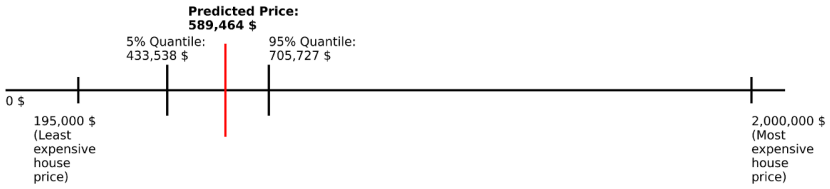
Introduction to Artificial Intelligence

After conducting the prediction on your own, you will be asked to adjust the AI's prediction of the same task.

The AI produces its prediction based on the tabular data and has not access to the image.

To better understand how recommendations of artificial intelligence look like, we want to show you an example. In the following, we show you an AI prediction.

Street:	Wheatville St
City:	Chula Vista, California (USA)
Number of Bedrooms:	4
Number of Bathrooms:	2
Square Feet:	1873



From the plot you can see that the average predicted price of the AI for the particular house is 589,464 $. Additionally, we provide you with information of the certainty of the AI in the form of quantiles. The 5% quantile tells you that with a probability of 95% the prediction of the AI is greater than the respective quantile value (433,538 $). Similarly, the 95% quantile tells you that the AI's prediction is with a probability of 95% smaller than the quantile's value (705,727 $).

In our example, with a probability of 95% the AI's prediction is greater then 433,538 $ and also smaller than 705,727 $.

The AI can be seen as a strong and reasonable base estimator. Therefore, your task will be to adjust the AI's base estimate in a best possible way.

Figure 3: AI tutorial. In the condition without UHCI the image is not displayed.

Figure 4: Training example 1. First, humans estimate the house price alone, then together with the AI. Afterwards, the listing price of the house ($749,000) is revealed. In the condition without UHCI the image is not displayed.

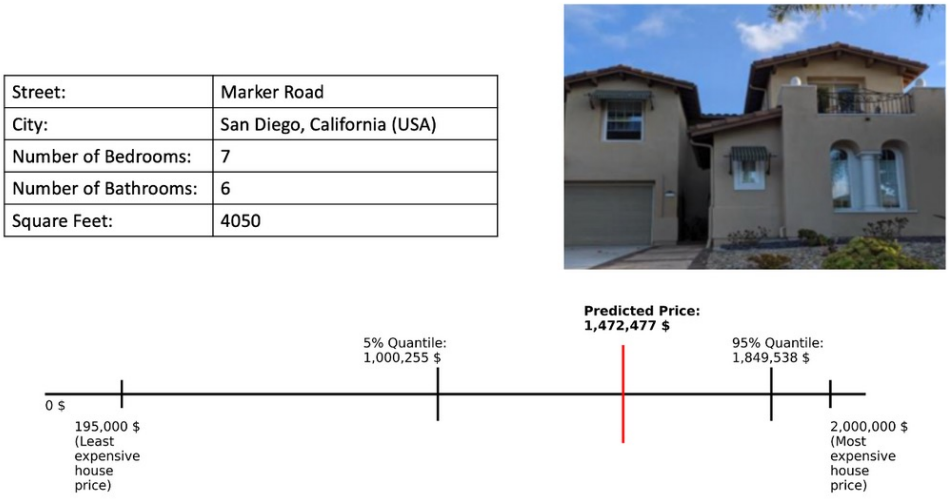


Figure 5: Training example 2. First, humans estimate the house price alone, then together with the AI. Afterwards, the listing price of the house ($1,495,000) is revealed. In the condition without UHCI the image is not displayed.

Street:	Milano Court
City:	Yorba Linda, California (USA)
Number of Bedrooms:	5
Number of Bathrooms:	4
Square Feet:	4211

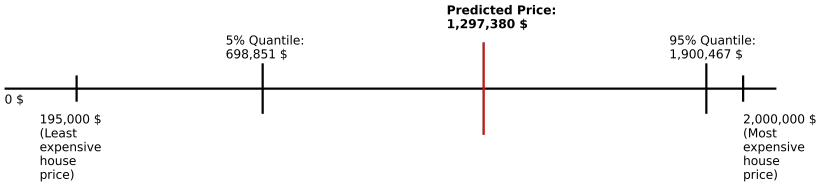


Figure 6: Task instance number 1. In the condition without UHCI the image is not displayed.

Street:	Evelyn Avenue
City:	California City, California (USA)
Number of Bedrooms:	3
Number of Bathrooms:	2
Square Feet:	1410

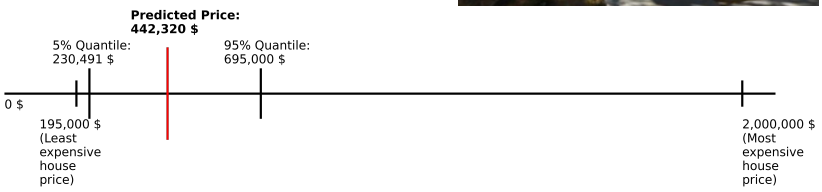


Figure 7: Task instance number 2. In the condition without UHCI the image is not displayed.

Street:	Blue Stake Road
City:	Phelan, California (USA)
Number of Bedrooms:	3
Number of Bathrooms:	2
Square Feet:	2500

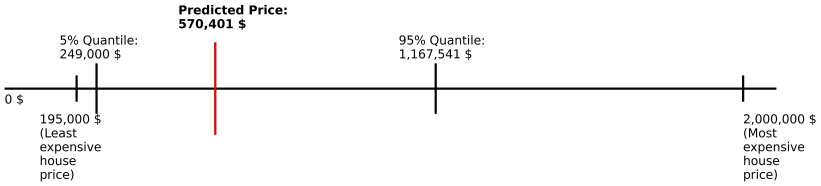


Figure 8: Task instance number 3. In the condition without UHCI the image is not displayed.

Street:	Schick Lane
City:	Lakewood, California (USA)
Number of Bedrooms:	3
Number of Bathrooms:	2
Square Feet:	1922

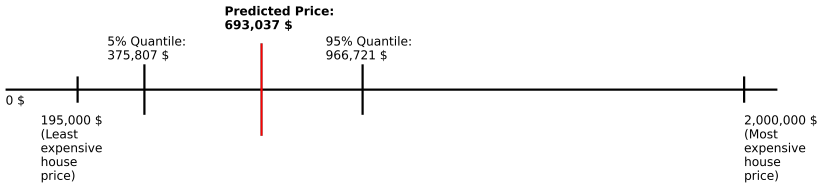


Figure 9: Task instance number 4. In the condition without UHCI the image is not displayed.

Street:	176th Street East
City:	Lancaster, California (USA)
Number of Bedrooms:	3
Number of Bathrooms:	2
Square Feet:	1296

Figure 10: Task instance number 5. In the condition without UHCI the image is not displayed.

Street:	Grand Avenue
City:	Wildomar, California (USA)
Number of Bedrooms:	4
Number of Bathrooms:	2
Square Feet:	2429

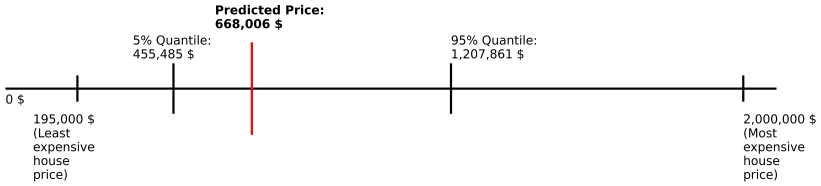


Figure 11: Task instance number 6. In the condition without UHCI the image is not displayed.

Street:	East Avenue Q
City:	Palmdale, California (USA)
Number of Bedrooms:	3
Number of Bathrooms:	2
Square Feet:	1519

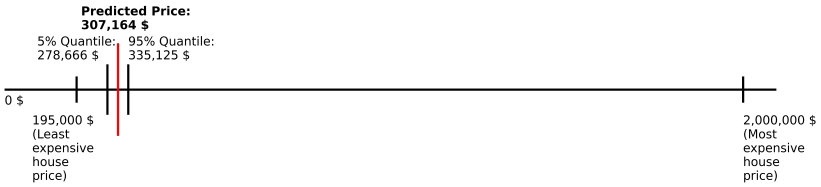


Figure 12: Task instance number 7. In the condition without UHCI the image is not displayed.

Street:	Momat Avenue
City:	Wildomar, California (USA)
Number of Bedrooms:	4
Number of Bathrooms:	2
Square Feet:	1738

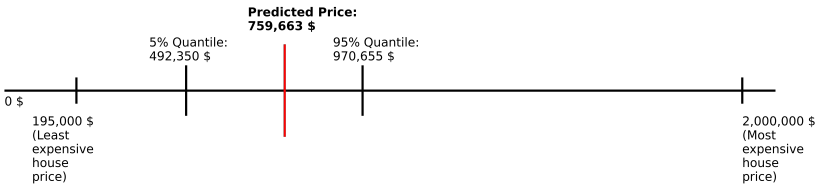


Figure 13: Task instance number 8. In the condition without UHCI the image is not displayed.

Street:	Winfield Drive
City:	Ladera Ranch, California (USA)
Number of Bedrooms:	4
Number of Bathrooms:	3
Square Feet:	3225

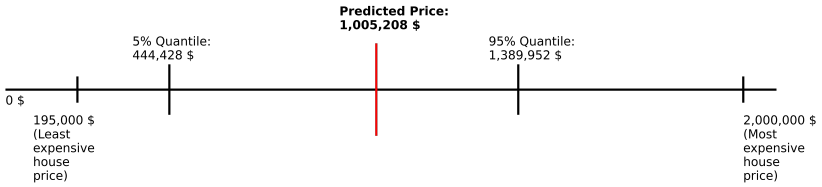


Figure 14: Task instance number 9. In the condition without UHCI the image is not displayed.

Street:	Somerset Place
City:	Rowland Heights, California (USA)
Number of Bedrooms:	5
Number of Bathrooms:	4
Square Feet:	3300

Figure 15: Task instance number 10. In the condition without UHCI the image is not displayed.

Street:	Elm Avenue
City:	Lancaster, California (USA)
Number of Bedrooms:	4
Number of Bathrooms:	2
Square Feet:	1517

Figure 16: Task instance number 11. In the condition without UHCI the image is not displayed.

Street:	East Street
City:	Lancaster, California (USA)
Number of Bedrooms:	4
Number of Bathrooms:	2
Square Feet:	1204

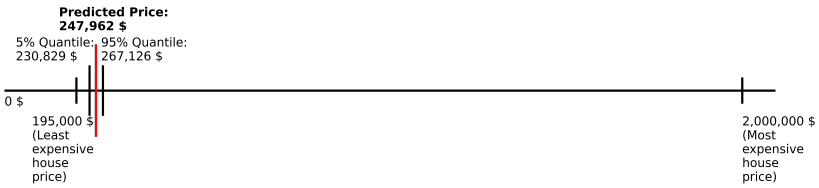


Figure 17: Task instance number 12. In the condition without UHCI the image is not displayed.

Street:	Oceo Circle S
City:	Palm Springs, California (USA)
Number of Bedrooms:	3
Number of Bathrooms:	3
Square Feet:	2722

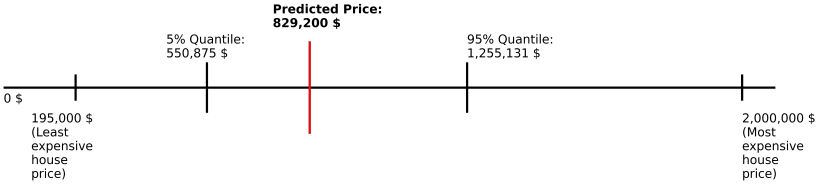


Figure 18: Task instance number 13. In the condition without UHCI the image is not displayed.

Street:	Stone Mountain Drive
City:	Chino Hills, California (USA)
Number of Bedrooms:	4
Number of Bathrooms:	2
Square Feet:	1901

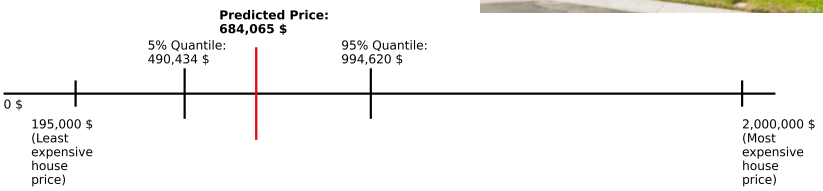


Figure 19: Task instance number 14. In the condition without UHCI the image is not displayed.

Figure 20: Task instance number 15. In the condition without UHCI the image is not displayed.

B.3 Participant Statistics

Table 1 contains descriptive information about the characteristics of the experiment's participants.

Table 1: Summary of participants' characteristics.

Number per condition	Without UHCI = 53 With UHCI = 48
Age	Mean = 24.44 Standard deviation = 6.4
Gender	Female = 56% Male = 43% Non-binary = 1%
Education	High school = 41% Bachelor's = 44% Master's = 10% Other = 5%

B.4 Additional Experimental Results

A detailed analysis of the 15 randomly drawn images shows that not just single instances contributed to the overall performance improvement (see Tables 2 and 3). Instead, the majority of instances show improvements through the availability of unique contextual information with regard to humans' individual performance as well as the resulting team performance.

With regard to the condition without UHCI, we can observe that humans tend to integrate the final team prediction between their own initial estimate and the sole AI prediction while placing more weight on the suggestions provided by the AI (see Table 3). This is also reflected in the weight of advice (WOA), which can be calculated by dividing the difference between the team prediction and the human's initial estimate by the difference between the AI prediction and the human's initial estimate. Averaging over all task instances and winsorizing values less than 0 and greater than 1 results in a mean WOA across participants of 0.78 (without UHCI).

Table 2: MAE per task instance across conditions including the AI.

	Human Alone (w/o UHCI)	**Human + AI (w/o UHCI)**	**Human Alone (with UHCI)**	**Human + AI (with UHCI)**	**AI Alone (w/o UHCI)**
1	586,151	378,355	420,379	347,126	382,620
2	216,878	232,652	160,923	219,565	244,320
3	226,972	293,968	130,918	197,839	290,401
4	148,012	98,931	136,125	111,329	132,137
5	170,550	24,016	95,120	20,994	11,307
6	172,467	70,976	320,731	116,384	70,994
7	154,081	42,395	124,110	38,411	32,164
8	117,031	275,065	100,081	241,956	334,663
9	353,905	141,030	257,381	131,112	94,792
10	424,649	426,770	335,818	422,991	511,022
11	267,734	54,404	151,944	60,250	40,832
12	218,095	63,810	168,858	59,450	48,962
13	424,402	237,999	286,364	171,198	240,800
14	191,910	49,953	193,027	56,425	7,835
15	96,390	11,104	125,870	25,100	3,358

A similar observation can be found in the condition with UHCI. Humans also tend to integrate the final team prediction between their own initial estimate and the sole AI prediction while placing a similar weight on the AI suggestions, which is reflected in a mean WOA of 0.76 (with UHCI). Even though the overall weight placed on the AI predictions remains similar, we find that unique contextual information improves humans' individual predictions and the corresponding team predictions for many task instances.

In the following, we discuss examples of task instances where the availability of UHCI, on average, worsened and improved the prediction performance. We particularly highlight images 6 and 15, noting that in these cases, the visual aspects of the images had a detrimental impact on humans' individual predictions and the joint team predictions. Focusing on image 6, we observe that providing it as unique human contextual information led to a significantly lower prediction accuracy. This image shows an attractive scene with a swimming pool, which stands out visually. Its ground truth rating was significantly lower than the average individual human prediction, suggesting that the aesthetic appearance of the house was misleading in terms of accurate listing price predictions. Similarly, image 15, which depicts a large house, also had a lower listing price. The visual appearance of the house misled participants to overvalue the house price.

In contrast to task instances 6 and 15, for the majority of the task instances, unique contextual information results in performance improvements. A notable example is image 3, which shows a more basic-looking house with a low listing price of $280,000. The AI prediction is relatively high with $570,401. The image resulted in more accurate individual human estimates and overall more accurate team predictions.

Table 3: Average participant predictions per task instance across conditions including the AI predictions and the respective listing house prices.

	Human Alone (w/o UHCI)	Human + AI (w/o UHCI)	Human Alone (With UHCI)	Human + AI (With UHCI)	AI Alone	Ground Truth
1	1,130,075	1,314,475	1,284,621	1,347,040	1,297,380	1,680,000
2	412,350	430,652	352,006	417,565	442,320	198,000
3	500,576	573,968	371,811	475,316	570,401	280,000
4	464,710	644,377	486,350	627,691	693,037	560,900
5	403,851	257,883	314,286	252,452	246,306	234,999
6	631,967	691,089	1,005,980	806,469	668,006	739,000
7	413,534	311,734	387,447	313,202	307,164	275,000
8	487,795	700,065	471,118	664,873	759,663	425,000
9	802,699	962,743	941,084	993,892	1,005,208	1,100,000
10	997,187	928,891	1,049,821	933,482	838,978	1,350,000
11	507,232	298,611	393,402	307,250	287,832	247,000
12	416,379	262,810	367,858	258,450	247,962	199,000
13	670,147	832,001	970,041	919,319	829,200	1,070,000
14	515,133	651,658	661,861	689,188	684,065	691,900
15	371,918	305,749	406,102	323,100	302,358	299,000

The overarching trend in this analysis suggests that despite the varied results with individual images, there is a consistent direction in which images, in most cases, facilitate better individual and team predictions.

C Experiment 2: The Effect of Capability Asymmetry

C.1 Implementation Details of the AI Model

For the intervention, we implement the approach proposed by Hemmer et al. (2022). It pursues the idea to consider the capabilities of human team members in the training process instead of training the AI model in isolation. Thus, it optimizes the AI model for team performance. During the training process, the AI model learns to make a decision for a particular task instance or delegates it to a human team member who takes over the decision. For a classification task, the AI model consists of two ML components—one for the classification and one for the delegation task. Considering human capabilities requires the availability of human decisions in addition to ground truth labels for each instance in the training data set. For each image in the training data set, we randomly sample one of the available human decisions. Images that tend to be more challenging to humans exhibit a higher disagreement among multiple human decisions. Thus, it is more likely that the selected human decision deviates from the ground truth label. By jointly training both components, the AI model learns to focus its capacity on the cases that tend to be more challenging for humans while requiring human expertise for other task instances. For the behavioral experiment, we utilize the decisions made by the classification component of the AI model.

C.2 Experiment Details

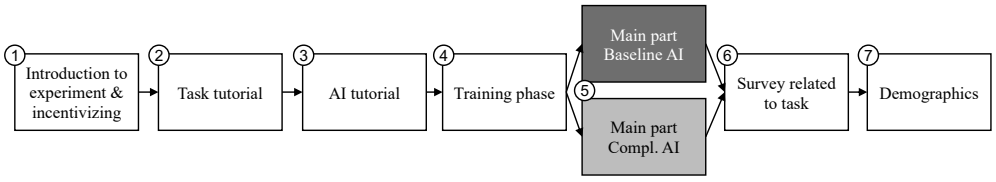


Figure 21: Sequence of the individual steps in the experiment.

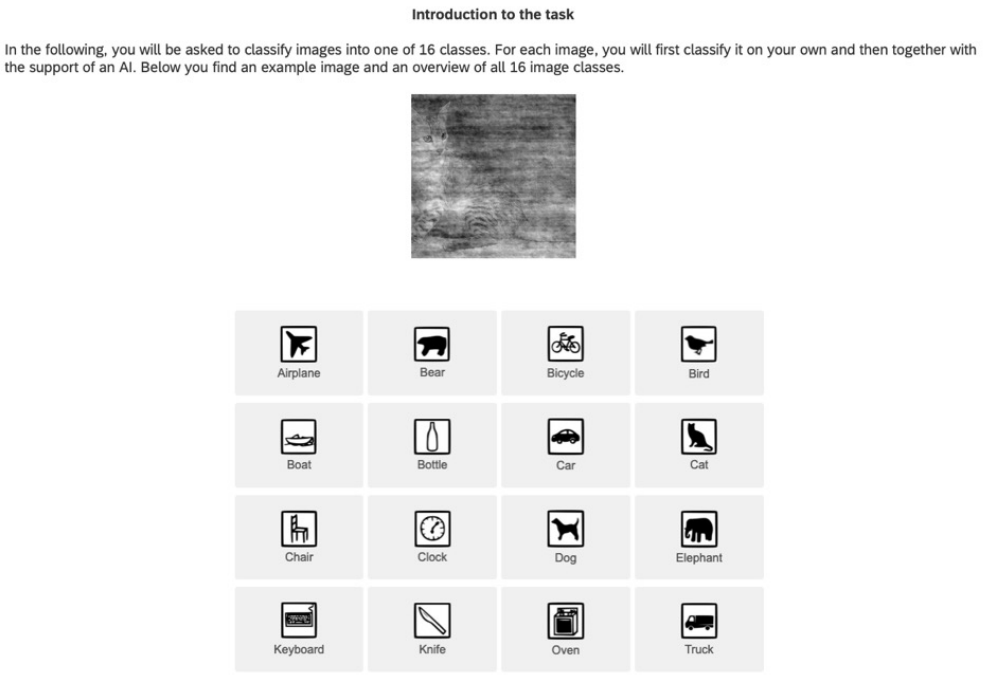


Figure 22: Task tutorial.

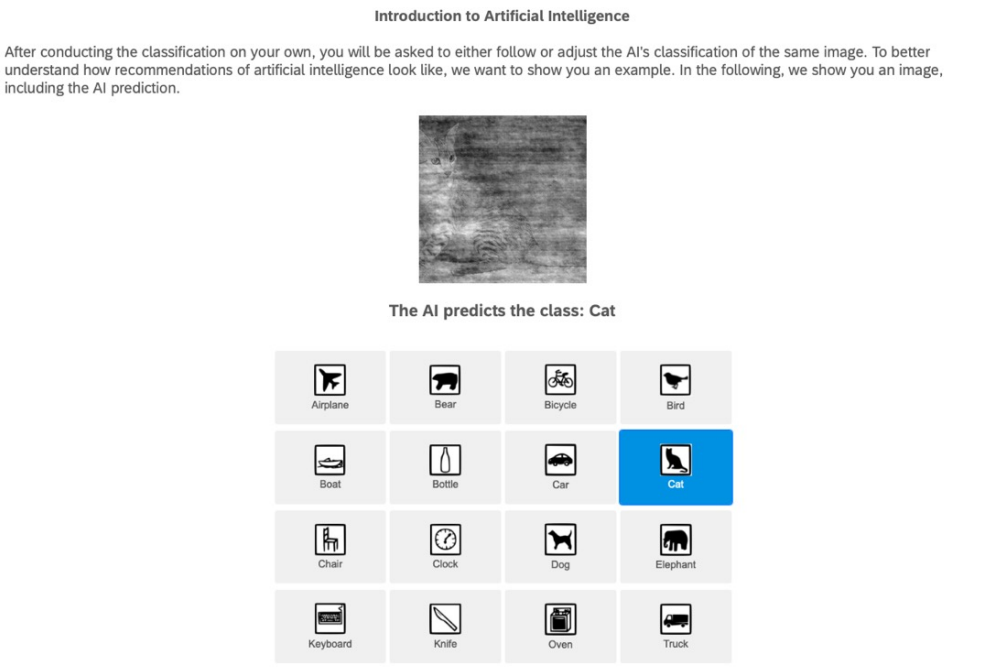


Figure 23: AI tutorial.

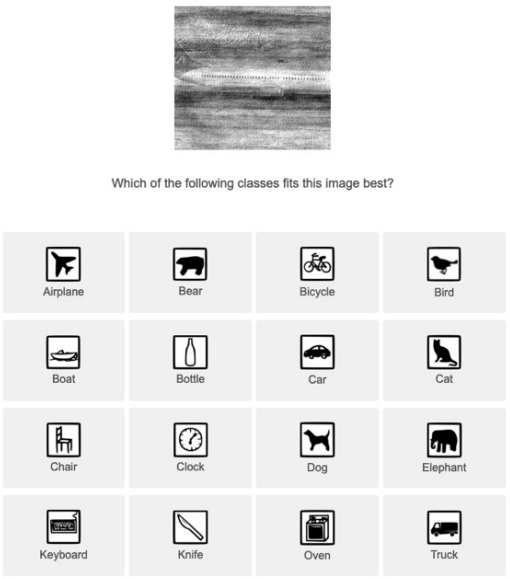


Figure 24: Training example 1. Humans have to classify the image alone to familiarize with the task. The ground truth label is not revealed.

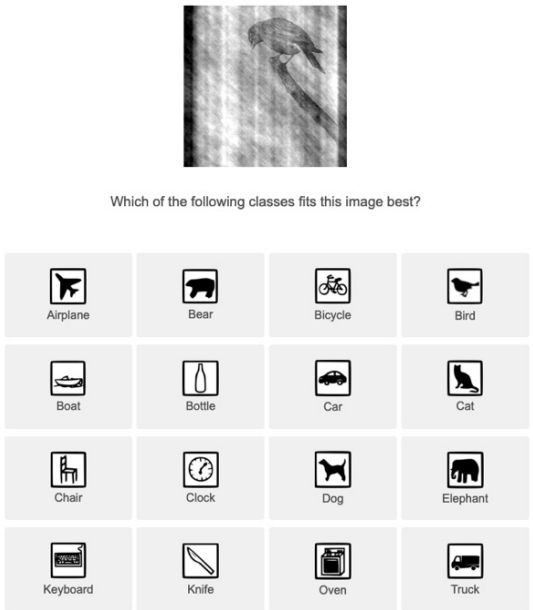


Figure 25: Training example 2. Humans have to classify the image alone to familiarize with the task. The ground truth label is not revealed.

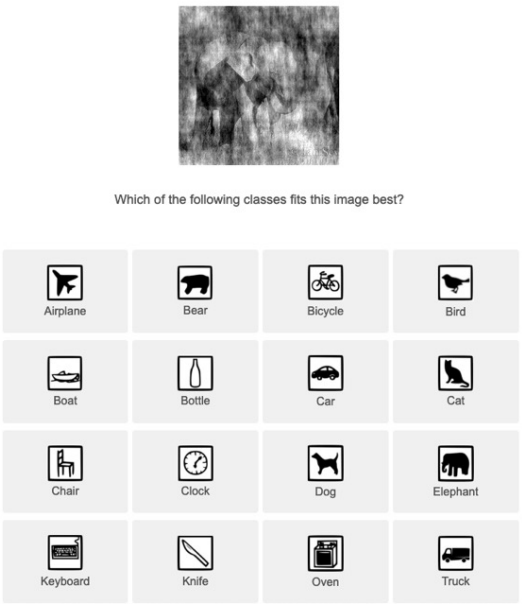


Figure 26: Training example 3. Humans have to classify the image alone to familiarize with the task. The ground truth label is not revealed.

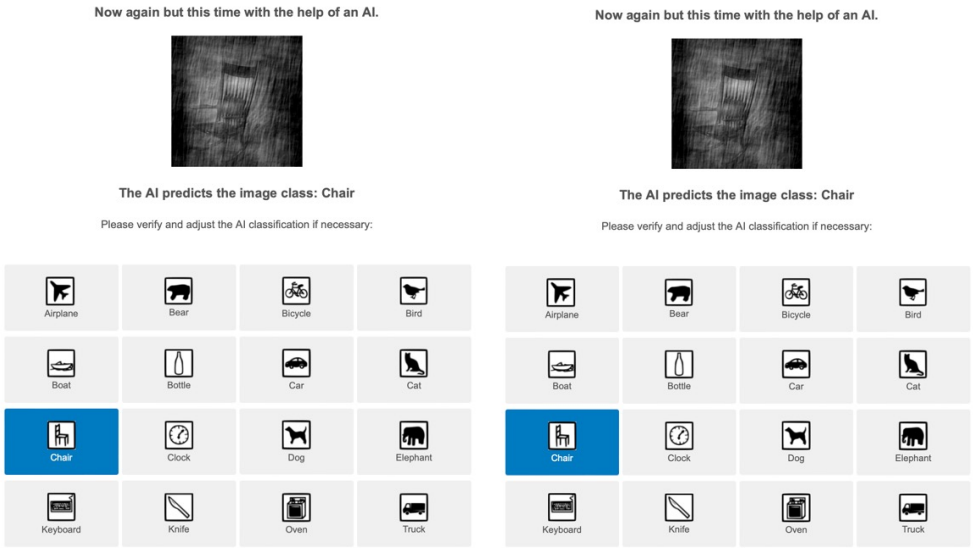


Figure 27: Task instance number 1. Left: Baseline AI. Right: Complementary AI. Ground truth label is chair.

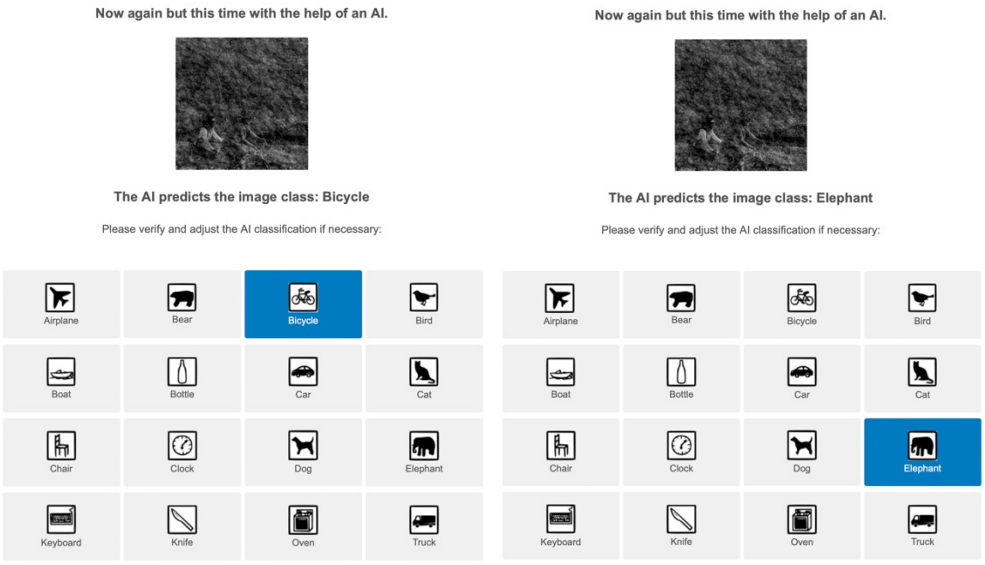


Figure 28: Task instance number 2. Left: Baseline AI. Right: Complementary AI. Ground truth label is bicycle.

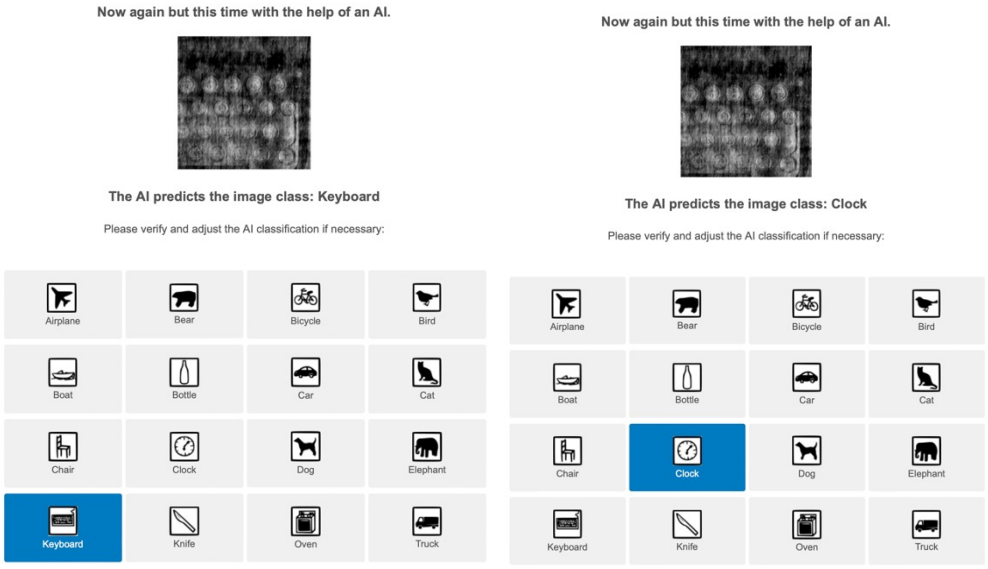


Figure 29: Task instance number 3. Left: Baseline AI. Right: Complementary AI. Ground truth label is keyboard.

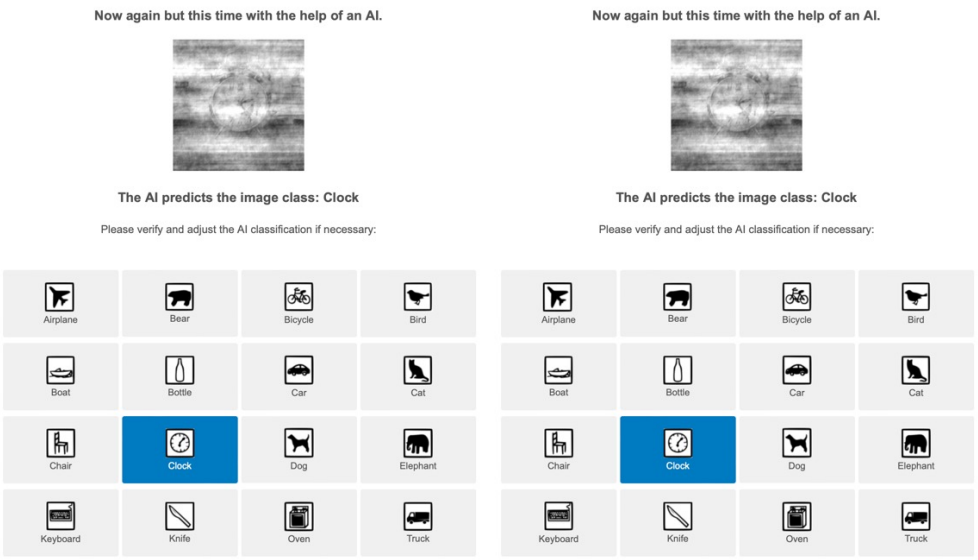


Figure 30: Task instance number 4. Left: Baseline AI. Right: Complementary AI. Ground truth label is clock.

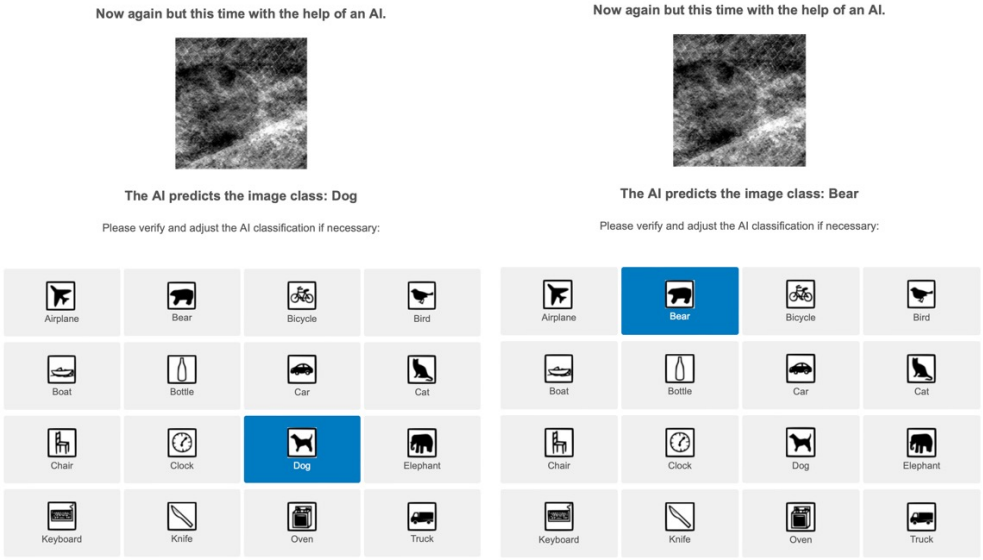


Figure 31: Task instance number 5. Left: Baseline AI. Right: Complementary AI. Ground truth label is bear.

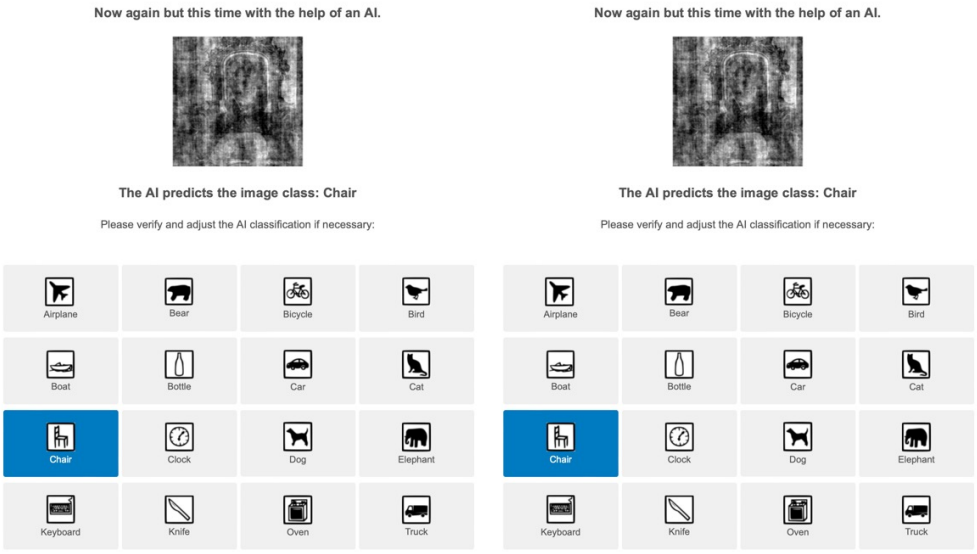


Figure 32: Task instance number 6. Left: Baseline AI. Right: Complementary AI. Ground truth label is chair.

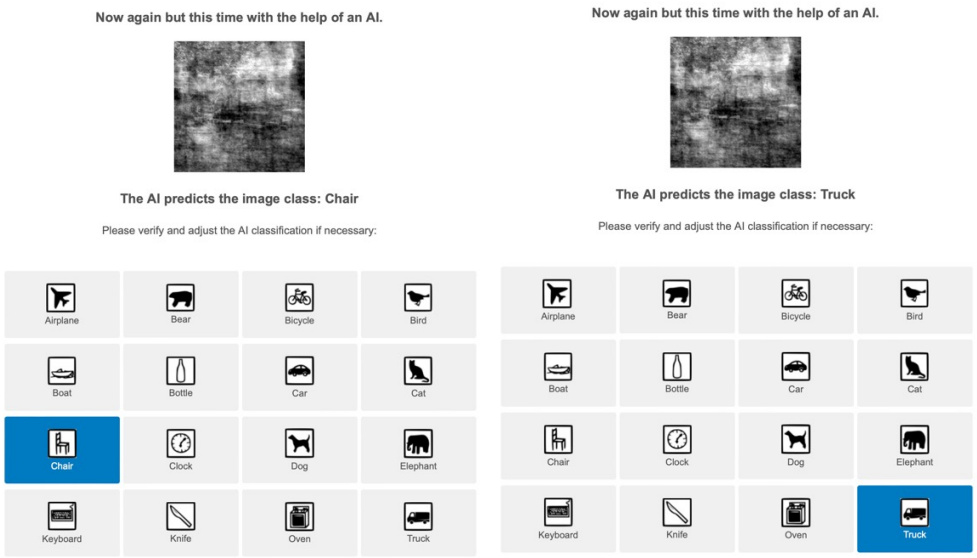


Figure 33: Task instance number 7. Left: Baseline AI. Right: Complementary AI. Ground truth label is truck.

Figure 34: Task instance number 8. Left: Baseline AI. Right: Complementary AI. Ground truth label is chair.

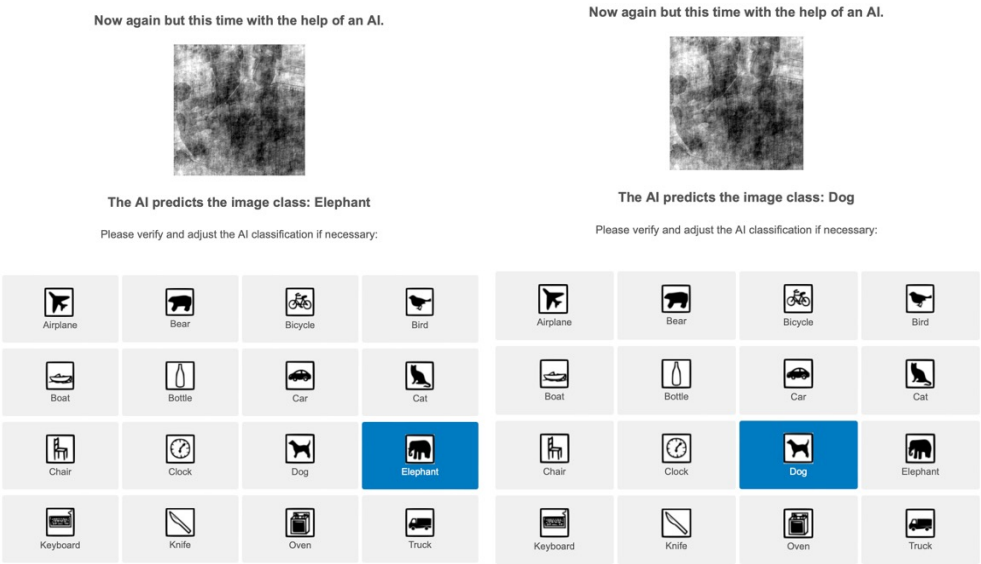


Figure 35: Task instance number 9. Left: Baseline AI. Right: Complementary AI. Ground truth label is dog.

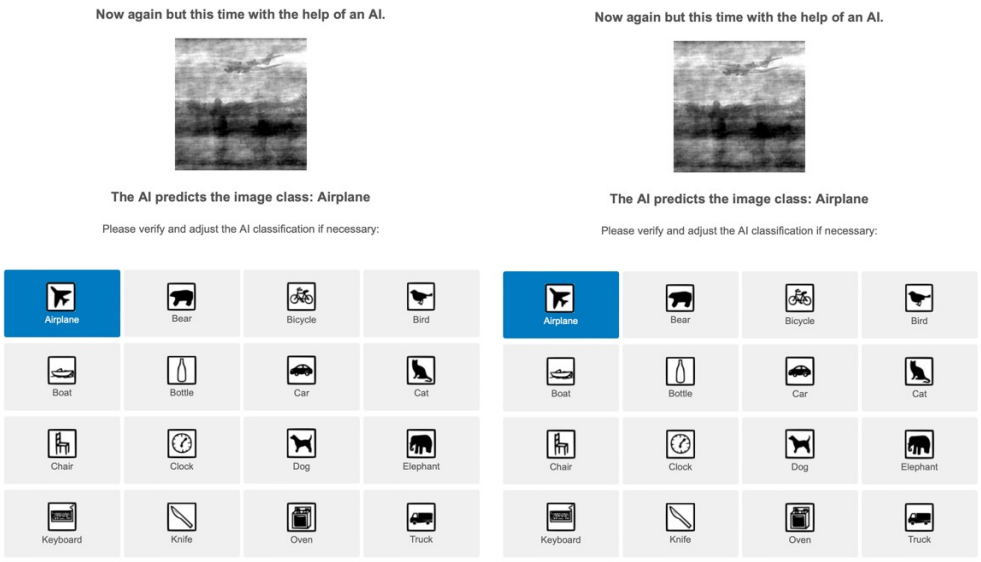


Figure 36: Task instance number 10. Left: Baseline AI. Right: Complementary AI. Ground truth label is airplane.

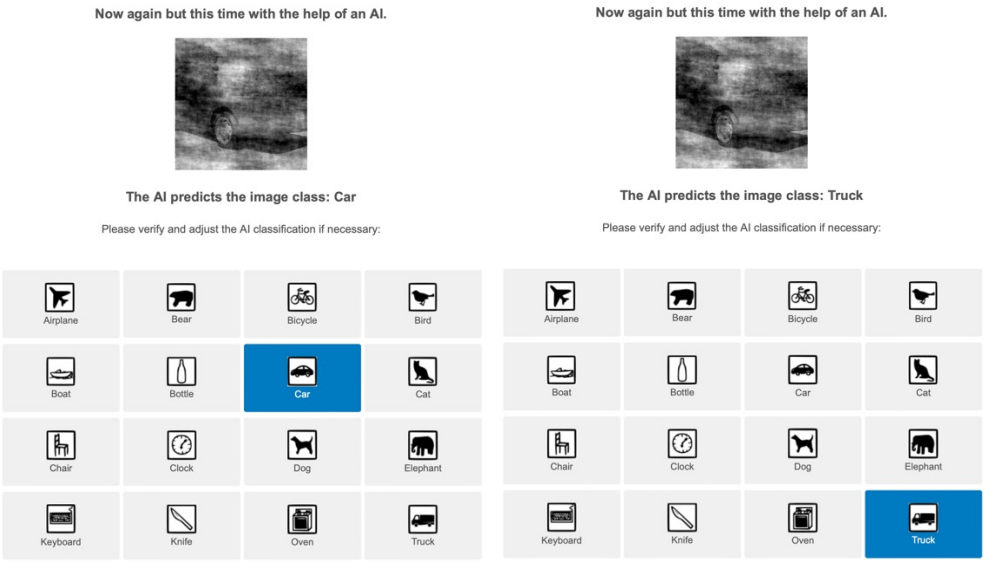


Figure 37: Task instance number 11. Left: Baseline AI. Right: Complementary AI. Ground truth label is car.

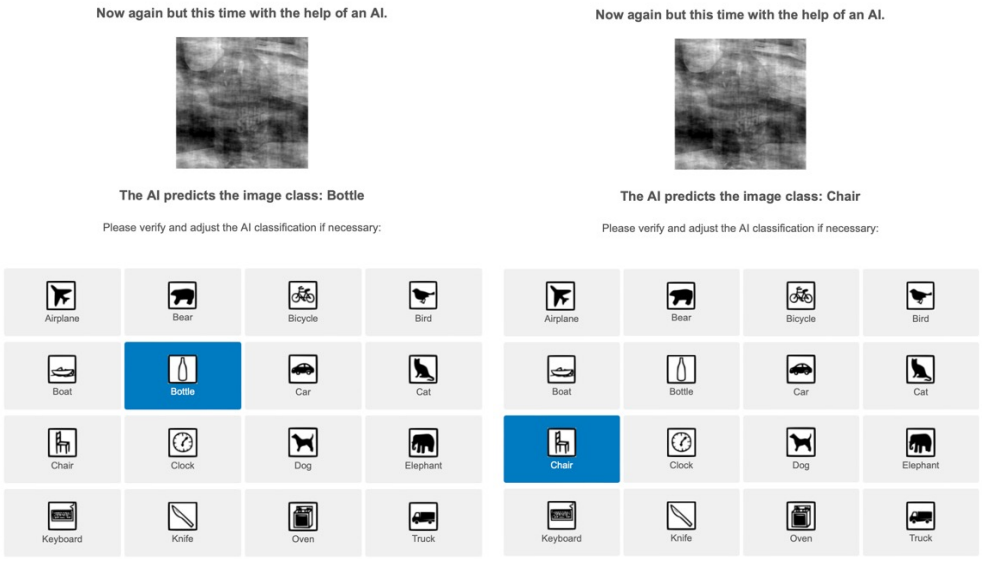


Figure 38: Task instance number 12. Left: Baseline AI. Right: Complementary AI. Ground truth label is bottle.

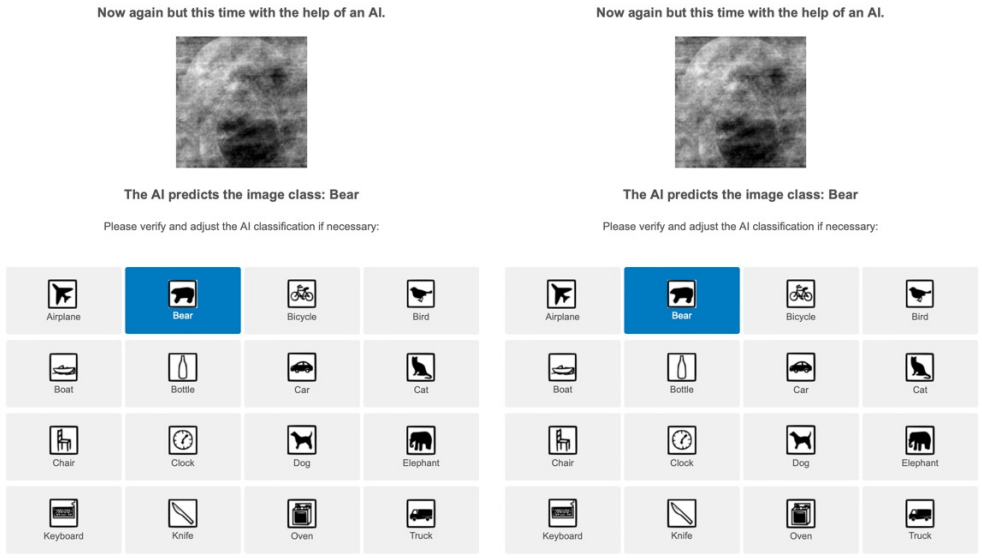


Figure 39: Task instance number 13. Left: Baseline AI. Right: Complementary AI. Ground truth label is bear.

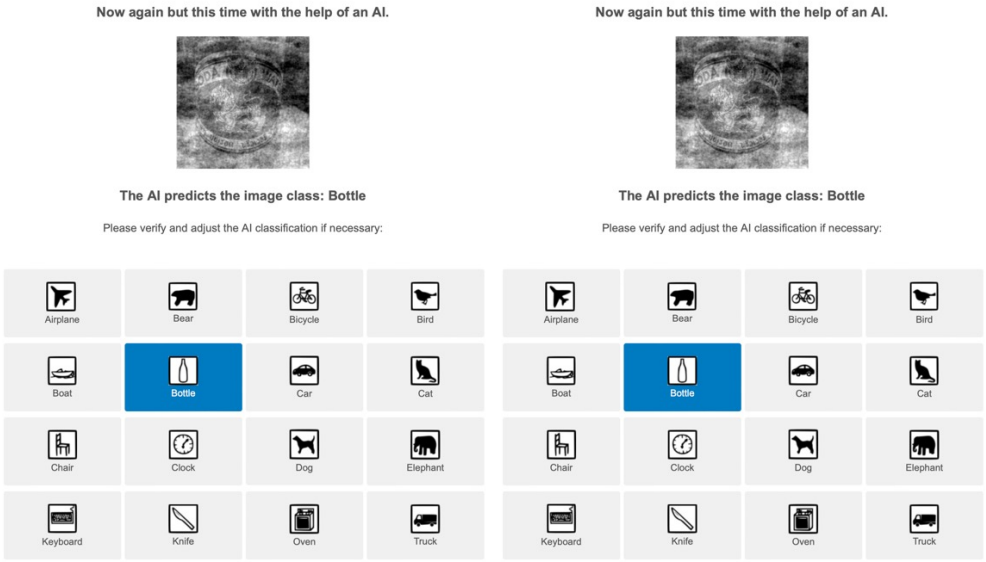


Figure 40: Task instance number 14. Left: Baseline AI. Right: Complementary AI. Ground truth label is bottle.

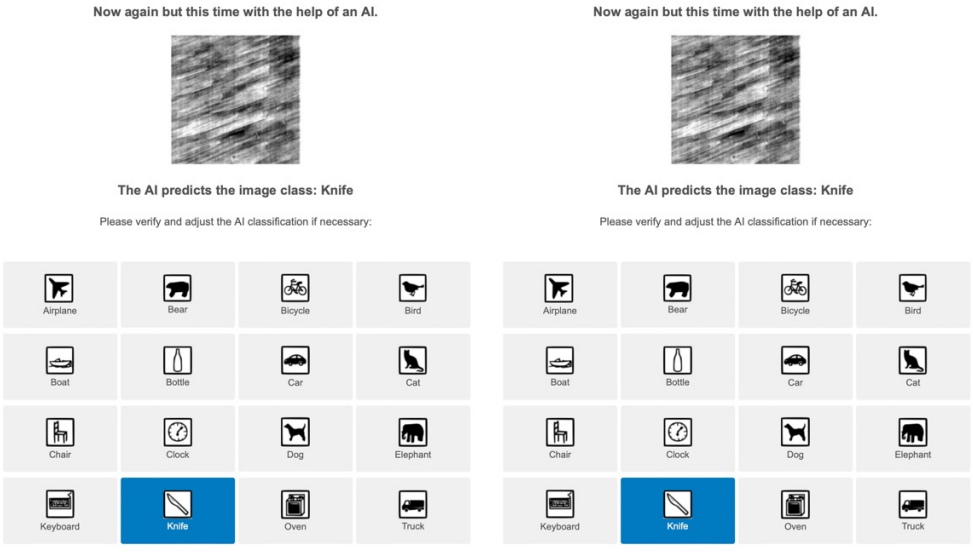


Figure 41: Task instance number 15. Left: Baseline AI. Right: Complementary AI. Ground truth label is knife.

C.3 Participant Statistics

Table 4 contains descriptive information about the characteristics of the experiment's participants.

Table 4: Summary of participants' characteristics.

Number per condition	Baseline AI = 76 Complementary AI = 68
Age	Mean = 31.07 Standard deviation = 9.0
Gender	Female = 49% Male = 50% Prefer not to say = 1%
Education	High school = 33% Bachelor's = 36% Master's = 26% Other = 5%

C.4 Additional Experimental Results

Table 5 shows a detailed comparison of the classification errors in different conditions. As expected, humans performing the task alone also show a similar performance across all individual instances. However, it is noticeable that there are instances where both AI models make correct decisions, while there are also other instances where they make different errors, essentially leading to different levels of inherent complementarity potential.

Comparing the error rates between the AI models and the human alone highlights the different degrees of complementarity potential. For example, for task instance 2, the baseline AI makes an accurate decision, while the complementary AI does not. Here, the individual human performance is also high, suggesting limited added value from AI assistance in this case. Conversely, for task instance 5, the roles are reversed in the condition with the complementary AI, highlighting a scenario where human performance can benefit significantly from AI input.

Moreover, Table 6 displays the disagreement fraction with the ground truth label for each task instance when humans initially classify the images alone, which can be interpreted as a proxy for the difficulty level for humans. Additionally, it also depicts the switch fraction after humans receive the AI suggestions. It denotes the fractions of humans for each condition that initially disagreed with the AI suggestion when classifying the image alone and selected the AI suggestion as the team decision after receiving the AI advice. For example, task instance 5 shows a high disagreement fraction in humans' initial decisions. In both conditions, a comparable fraction of humans switched to the suggestion of the AI. Whereas the advice turns out to be correct in the complementary AI condition, it is incorrect in the baseline AI condition. Interestingly, despite witnessing the incorrect AI suggestions in the complementary AI condition where humans have a low initial disagreement with the ground truth label, e.g., 2, 3, 11, and 12, they continue perceiving the AI suggestions for the more difficult task instances as useful support and tend to incorporate it in the final team decision.

To summarize, the human-AI team improves over humans' individual performance for the majority of the task instances in the complementary AI condition. This is particularly noteworthy as no potential aversion towards the AI can be observed, even though the instances for which the AI makes incorrect suggestions tend to be relatively easy for humans. This suggests that participants in the complementary AI condition were able to take advantage of the existing complementarity potential and did not start developing aversion towards the AI.

Table 5: Classification error per task instance across conditions including the Baseline AI and the Complementary AI.

	Human Alone (Baseline AI)	Human + AI (Baseline AI)	Human Alone (Compl. AI)	Human + AI (Compl. AI)	Baseline AI	Compl. AI
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.01	0.00	0.01	0.06	0.00	1.00
3	0.25	0.01	0.18	0.22	0.00	1.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.82	0.95	0.87	0.24	1.00	0.00
6	0.09	0.00	0.10	0.03	0.00	0.00
7	0.86	0.92	0.94	0.59	1.00	0.00
8	0.75	0.86	0.74	0.19	1.00	0.00
9	0.62	0.72	0.69	0.26	1.00	0.00
10	0.01	0.00	0.00	0.00	0.00	0.00
11	0.03	0.00	0.06	0.21	0.00	1.00
12	0.03	0.00	0.03	0.13	0.00	1.00
13	0.58	0.17	0.49	0.19	0.00	0.00
14	0.11	0.03	0.03	0.00	0.00	0.00
15	0.36	0.05	0.29	0.07	0.00	0.00

Table 6: Disagreement fraction per image with respect to the ground truth labels for humans conducting the task alone aggregated over both conditions. Additionally, we report the switch fraction (initial disagreement with the AI decision when classifying each image alone vs. final agreement with the AI decision in the final team decision) for the baseline AI condition and the complementary AI condition.

	Human Alone Ground Truth Disagreement Fraction	Switch Fraction (Baseline AI)	Switch Fraction (Compl. AI)
1	0.00	0.00	0.00
2	0.01	0.01	0.06
3	0.22	0.24	0.15
4	0.00	0.00	0.00
5	0.84	0.71	0.63
6	0.10	0.09	0.07
7	0.90	0.53	0.35
8	0.74	0.55	0.54
9	0.65	0.39	0.43
10	0.01	0.01	0.00
11	0.04	0.03	0.15
12	0.03	0.03	0.13
13	0.53	0.41	0.29
14	0.07	0.08	0.03
15	0.33	0.30	0.22

References

Hemmer, P., Schellhammer, S., Vössing, M., Jakubik, J., and Satzger, G. (2022). Forming effective human-AI teams: Building machine learning models that complement the capabilities of multiple experts. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2478–2484.