

Supplementary Information

A Additional Experimental Results

We make two observations matching the theoretical assumptions and results in Corvelo Benz and Gomez Rodriguez [8] in the study.

Figure [3] shows the probabilities of guessing red given the initial and final guess of participants. We observe that participants’ initial guess is monotonic in their initial confidence, and that the final guess is monotonic in their initial confidence and the AI confidence.

Given the record of initial confidence $h_{i,j}$, the initial $d_{i,j}$ and final $d'_{i,j}$ color guess made by the participant j in the game instance i , we compute the best monotone π_{mono} and best joint π_{joint} guess for all games played. The best joint guess maximizes the expected accuracy of the guess given only confidence h_i and a_i —the guess is red (black) if the mean probability of a red card for game instances with confidence h_i and a_i is higher (lower) than 50%. The best monotone guess maximizes the expected accuracy among guesses made by rational decision makers—guesses that are monotone in confidence h_i and a_i .

We measure the expected accuracy of each guess averaged across participants and games,

$$\sum_j \frac{1}{\#\text{participants}} \cdot \sum_i \frac{1}{24} \cdot \mathbb{E}_{C_i \sim Be(r_i)} [\mathbb{1}[\pi(i, j) = C_i]] \quad (4)$$

where $\pi(i, j) \in \{d_{i,j}, d'_{i,j}, \pi_{\text{joint}}(i, j), \pi_{\text{mono}}(i, j)\}$ denotes the guess evaluated. For each guess, Supplementary Fig. [S1] shows a heatmap of the decision probabilities of guessing red stratified by participants’ initial confidence and AI confidence shown, and averaged over participants. Supplementary Fig. [S2] shows the utility in terms of expected accuracy (Eq. [4]) of the best joint guess and the best monotone guess that could be made by the participants based on their own and the AI confidence values. Although the expected accuracy of the best joint guess is mostly equivalent across groups and the expected accuracy of the AI alone is the same by design, the expected accuracy of the best monotone guess is lower in groups with lower degree of alignment. However, while the difference in the best monotone guess is barely noticeable, the difference in the expected accuracy of the final guess is more prominent.

A.1 Expected Calibration Error

Table S1: Expected Calibration Error (ECE) measured empirically for each group. As expected, all values are relatively low. The AI in group $\bullet \rightarrow \leftarrow \bullet$, while empirically the least aligned, also obtains lower ECE than in groups $\leftarrow \bullet \bullet \rightarrow$ and $\bullet \bullet$. The post-processing algorithm does not significantly reduce ECE of the AI in group $\boxed{\bullet \rightarrow \leftarrow \bullet}_R$.

Group	ECE
$\leftarrow \bullet \bullet \rightarrow$	0.090
$\bullet \bullet$	0.055
$\bullet \rightarrow \leftarrow \bullet$	0.036
$\boxed{\bullet \rightarrow \leftarrow \bullet}_R$	0.033

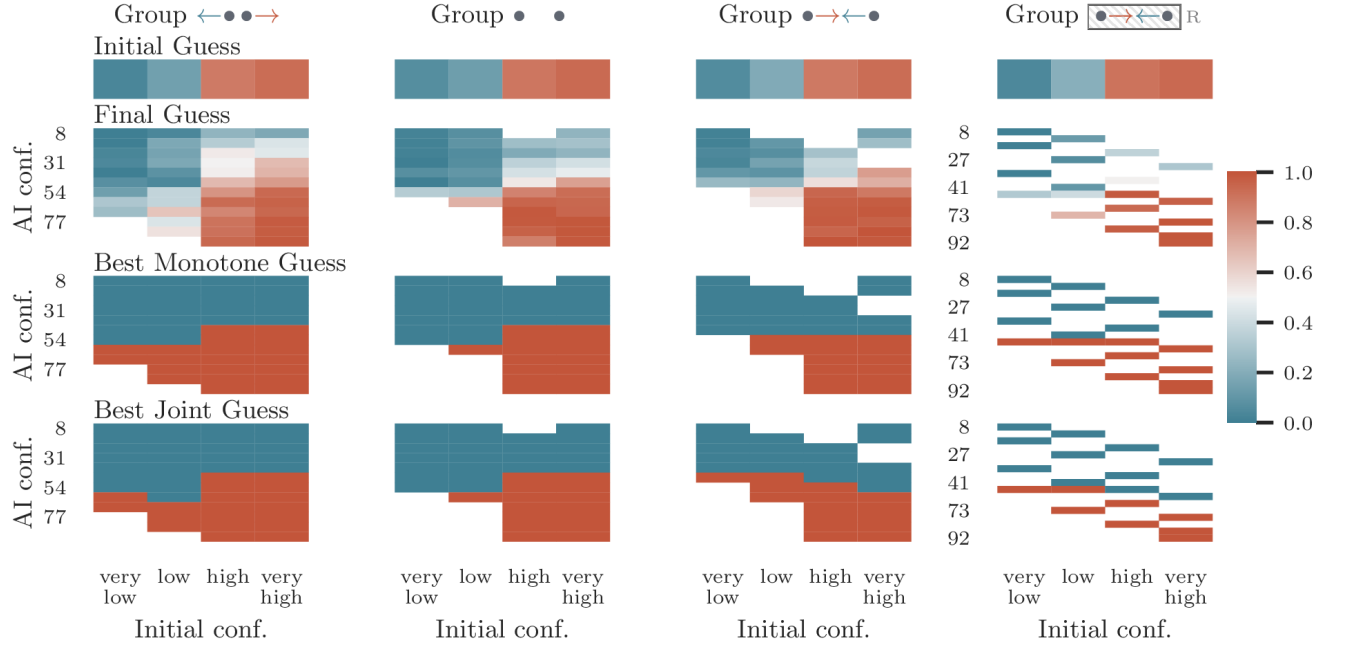
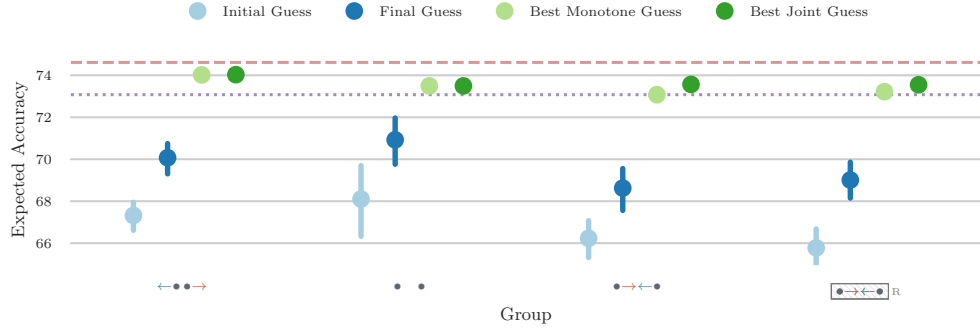
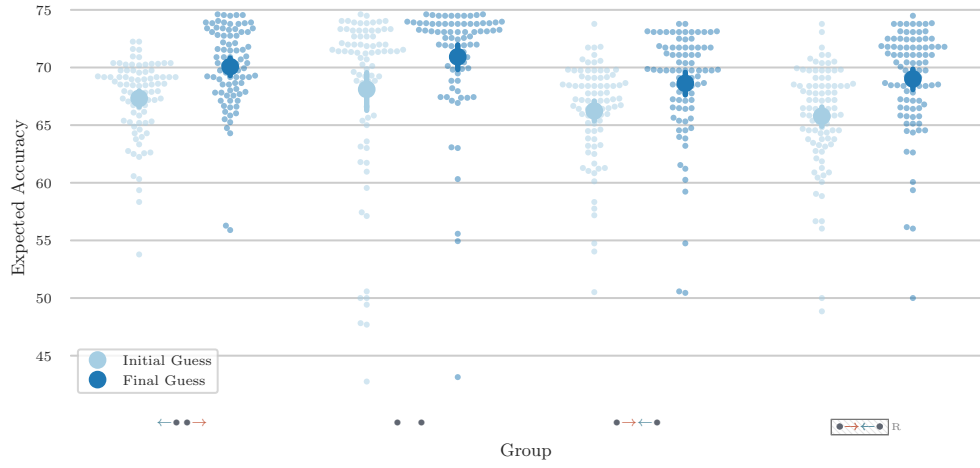


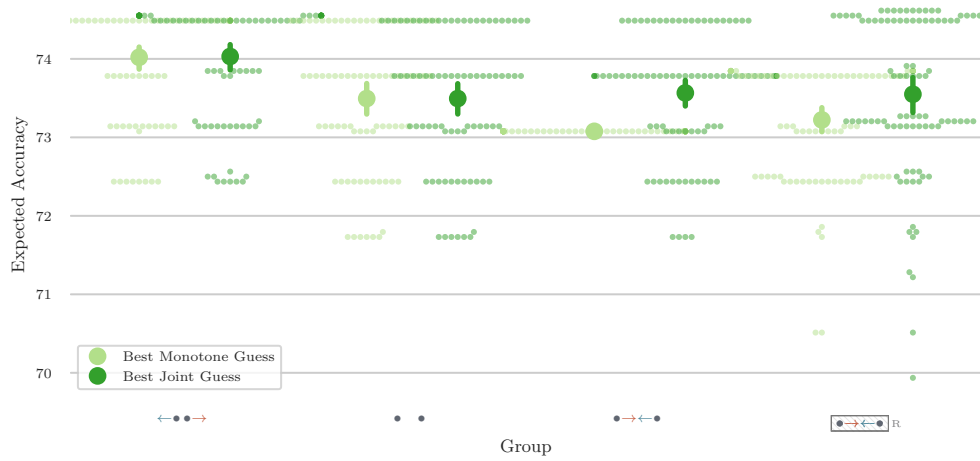
Figure S1: Heatmap of decision probabilities of guessing red stratified by participants' initial confidence and AI confidence shown, and averaged over participants. The initial confidence recorded is discretized into four bins—very low, low, high, very high—denoting the confidence of the participants that the color of the picked card will be red. Bins with 10 or less data points are not displayed.



(a) Expected accuracy averaged across participants and games for each group and type of guess. Red dashed line indicates the expected accuracy of the optimal guess π^* and purple dotted line indicates the expected accuracy of the AI playing alone. Both are by design of the study equal across groups and, thus, just drawn as lines.

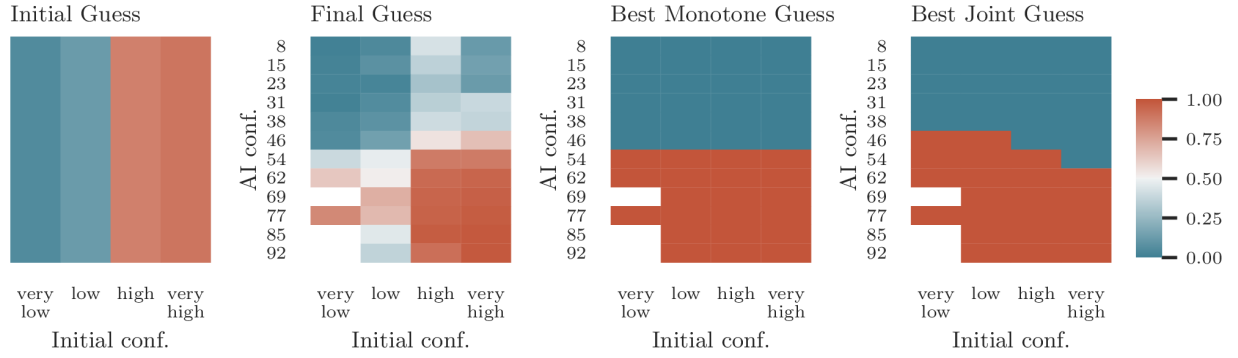


(b) Expected accuracy of each individual participant averaged across games for each group for the initial and final guess.

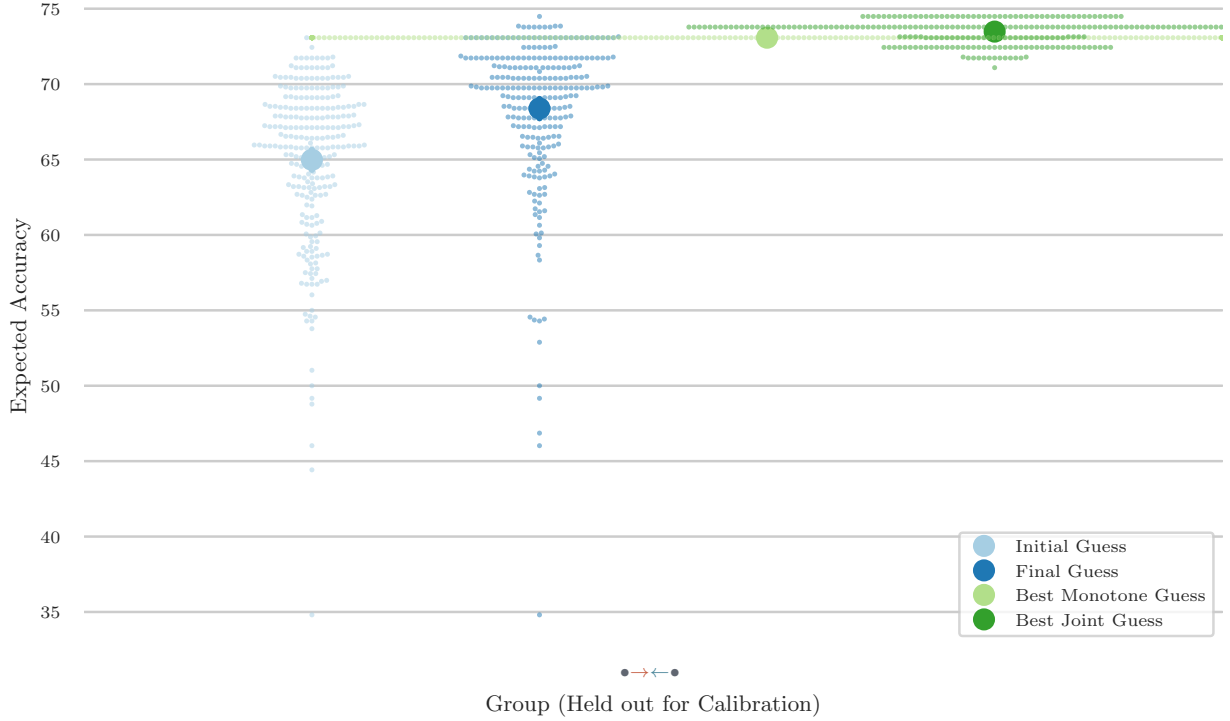


(c) Expected accuracy of each individual participant averaged across games for each group for the best joint and best monotone guess. Some points were not drawn due to space constraints.

Figure S2: Accuracy across groups and guesses in expectation. Plot (a) compares means and 95% confidence intervals, while plot (b) and (c) show individual expected accuracies of each participant as well.



(a) Heatmap of decision probabilities of guessing red stratified by participants' initial confidence and AI confidence shown, and averaged over participants. The initial confidence recorded is discretized into four bins—very low, low, high, very high—denoting the confidence of the participants that the color of the picked card will be red. Bins with 10 or less data points are not displayed.



(b) Expected accuracy of each individual participant averaged across games for each guess type. Means over participants are plotted as large circle.

Figure S3: Summarization plots for held-out calibration data from group ●→←●. Plot (a) shows a heatmap of the probability of guessing red. Plot (b) shows a swarm plot of the expected accuracy of each individual participant averaged across games for each guess type.

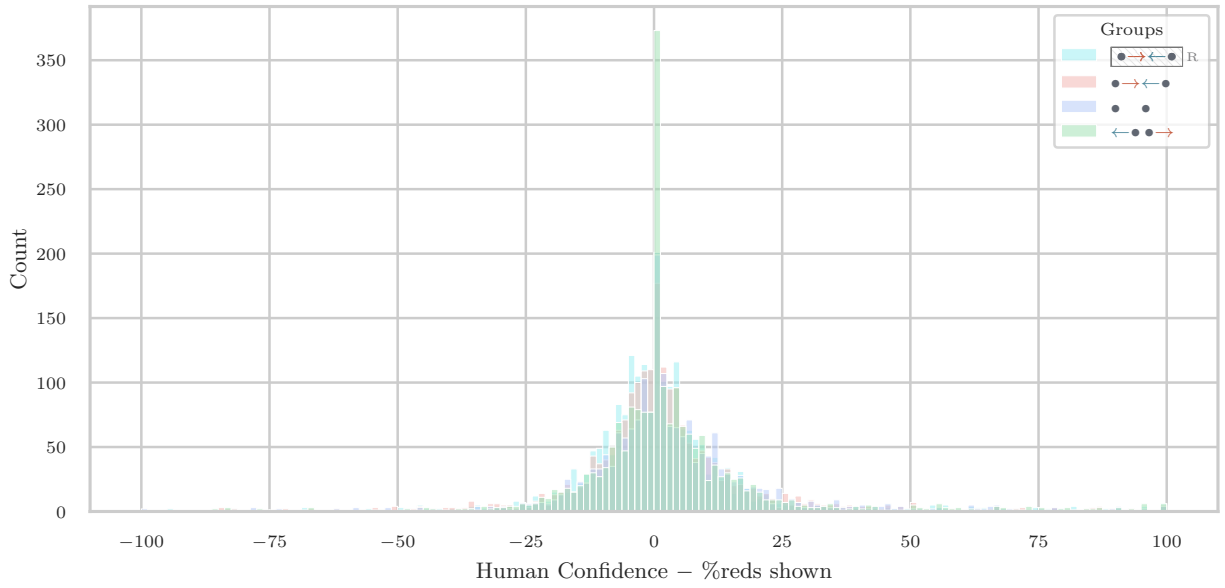


Figure S4: Histogram of the difference between reported initial confidence and percentage of reds shown to participants for each group. We observe no difference in distribution across groups.

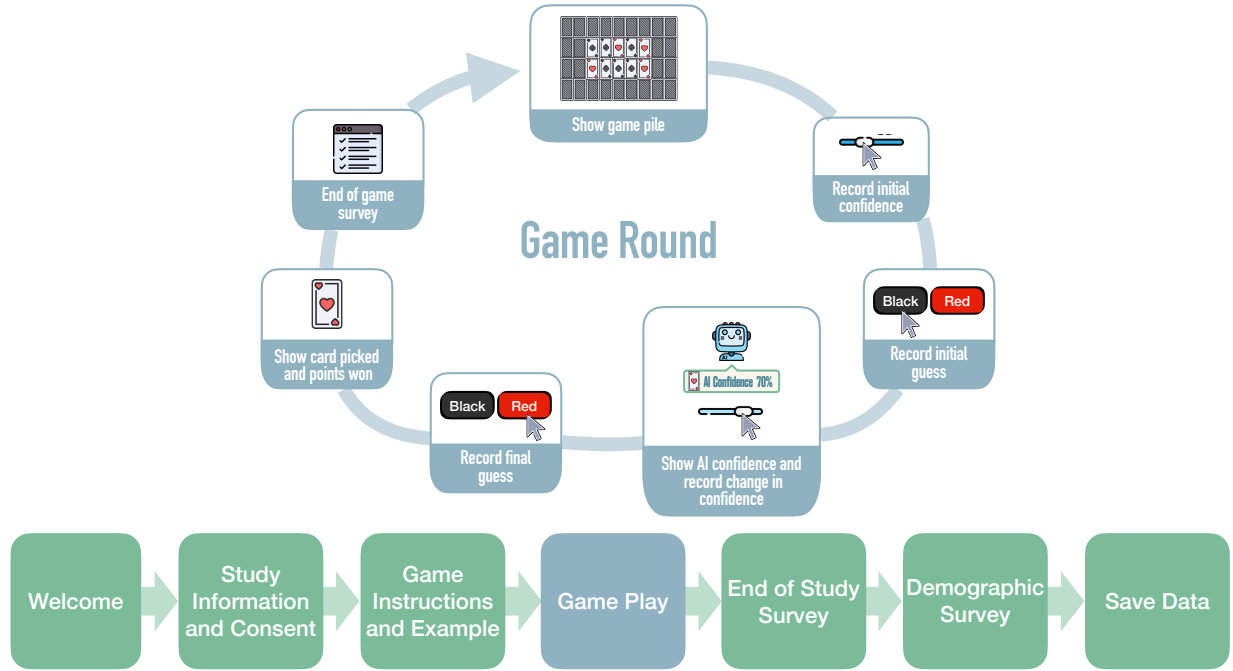


Figure S5: Flow chart showing an overview of the study timeline. Screenshots of an example game round can be found in Supplementary Fig. S6 to S9. A screenshot of the end of game survey and end of study survey can be seen in Supplementary Fig. S10. The full study setup can be found at https://hac-experiment.mpi-sws.org/?PROLIFIC_PID=test&STUDY_ID=test&SESSION_ID=test&LEVEL=B&GAME_BATCH=0.

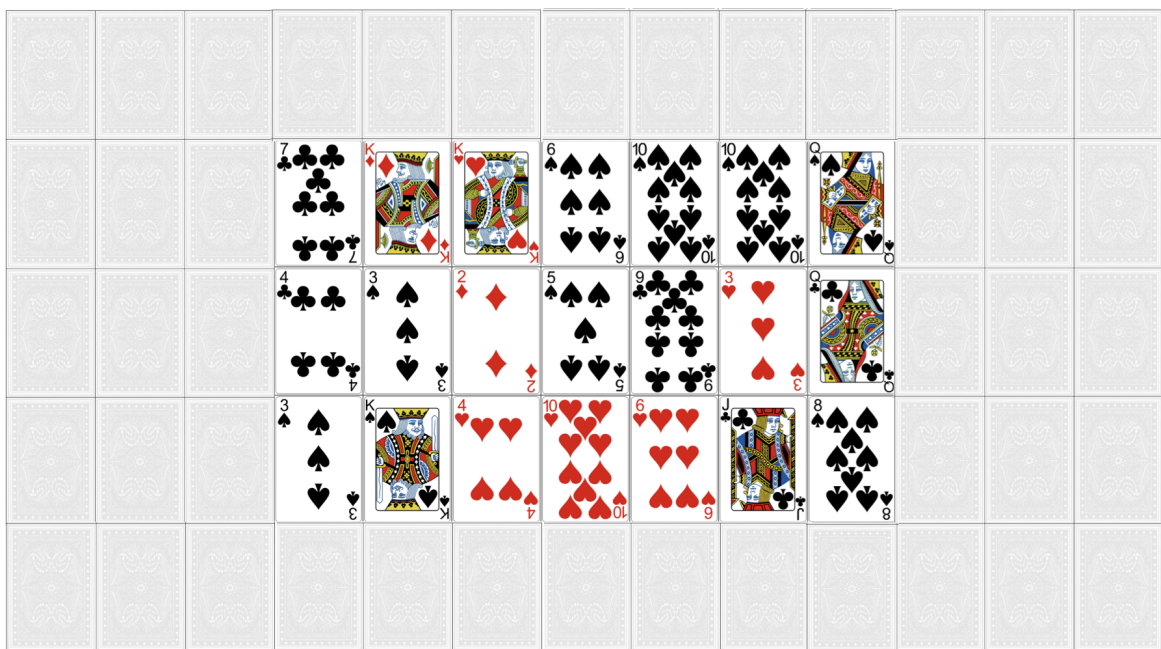
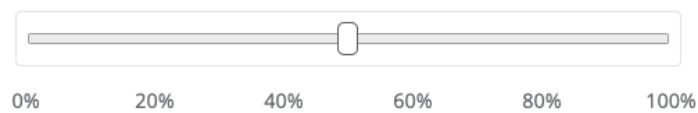


Figure S6: Example of a game pile shown to participants.

A random card will be picked from this round's game pile.
State your confidence that the picked card is red:

(Here, you need to click a value on the scale to click next.)



Next >

(a) Record initial confidence

Based on **your stated confidence of 23%** that the picked card is red,
what is your **initial guess** for the color of the picked card?

Black

Red

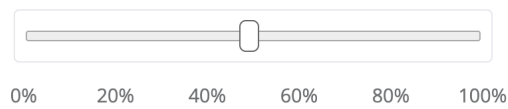
(b) Record initial guess

Figure S7: Screenshots of game round steps where participants are asked for their initial confidence and guess before being shown the AI confidence.

The AI's estimated likelihood that the picked card is red:

69%

What is your confidence that the picked card is red after seeing the AI's estimate:



Next >

(a) Show AI confidence and record change in confidence of the participant

Based on **your stated confidence of 30%** that the picked card is red,
what is your **final guess** for the color of the picked card?

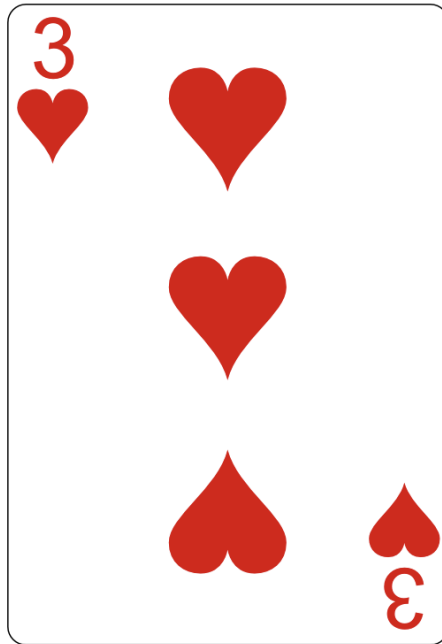
Black

Red

(b) Record final guess

Figure S8: Screenshots of game round steps where participants are shown the AI confidence and asked for their confidence and final guess.

Card Picked:



Your initial bet was **Black**.

Your final bet was **Red**.

You **WON 1** point!

You have **earned 0 point(s) from 2 round(s)** until now.

Next >

Figure S9: Screenshot of how participants are shown the card picked and points won.

State how much you agree with following statement about the AI.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I think the AI's estimate was accurate in this round.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Choose the statement that best applies to this round. *

- ☐ I won points in this round because I changed my guess based on the AI's estimate.
- ☐ I lost points in this round because I changed my guess based on the AI's estimate.
- ☐ The AI's estimate did not change my guess in this round.

Next Round >

(a) End of game survey

Please answer the following questions about your overall thoughts on the AI during the game.

State how much you agree with following statements about the AI.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Overall, I think the AI's estimates were accurate.	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Overall, the AI's estimates affected my confidence.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Overall, I trust the estimates of the AI.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

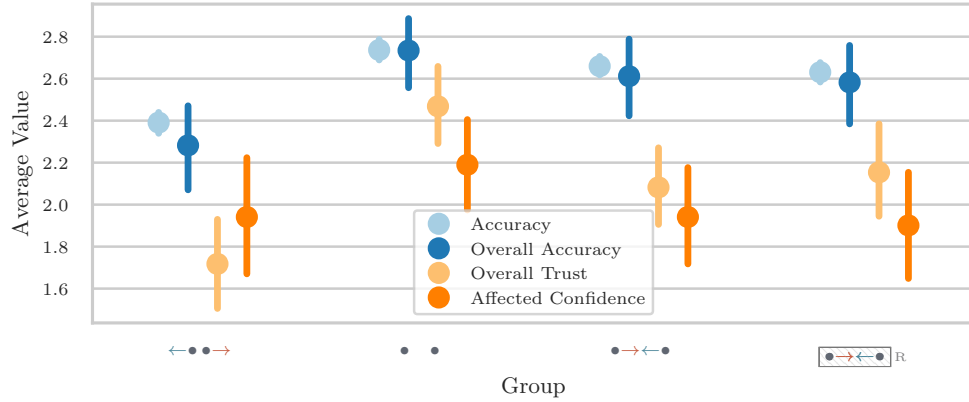
Choose the statement that best applies. *

- ☐ Overall, the AI's estimates changed my guess for the better, and I would have gained less points without the AI.
- ☐ Overall, the AI's estimates changed my guess for the worse, and I would have gained more points without the AI.
- ☐ Overall, the AI's estimates did not change my guess, and I would have gained a similar amount of points without the AI.
- ☐ Overall, the AI's estimates did not change my guess, and I believe I would have gained less points by trusting the AI.
- ☐ Overall, the AI's estimates did not change my guess, but I believe I would have gained more points by trusting the AI.

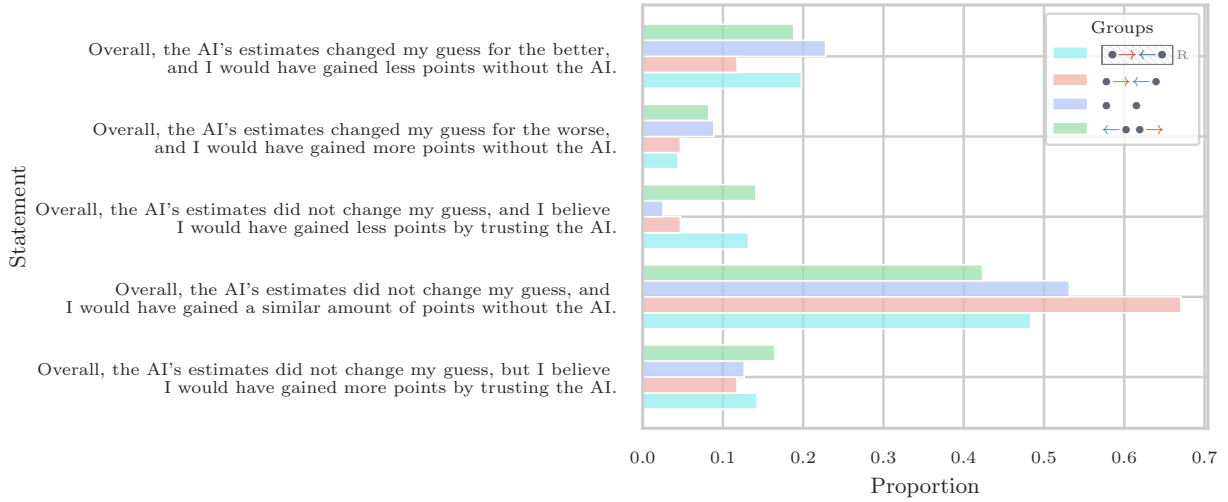
Next >

(b) End of study survey

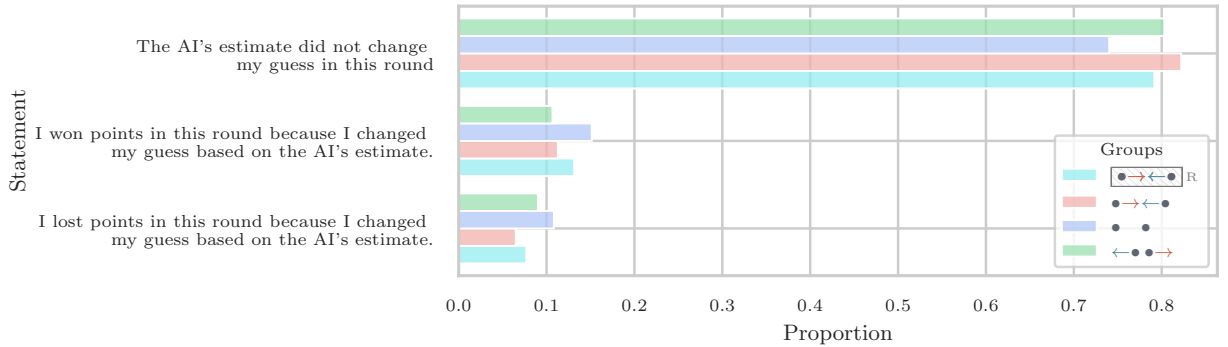
Figure S10: Screenshot of the end of game and end of study surveys filled out by participants.



(a) Averaged value of the likert scale—strongly disagree (0) to strongly agree (5)—in the end of game and end of study survey for following statements: "I think the AI's estimate was accurate in this round." (Accuracy), "Overall, I think the AI's estimates were accurate." (Overall Accuracy), "Overall, the AI's estimates affected my confidence." (Affected Confidence) and "Overall, I trust the estimates of the AI." (Overall Trust).



(b) Frequency of statement selected by participants to best apply to their whole game experience at the end of study survey.



(c) Frequency of statement selected by participants to best apply to their game experience at the end of game survey.

Figure S11: Summarized responses to end of game and end of study surveys.

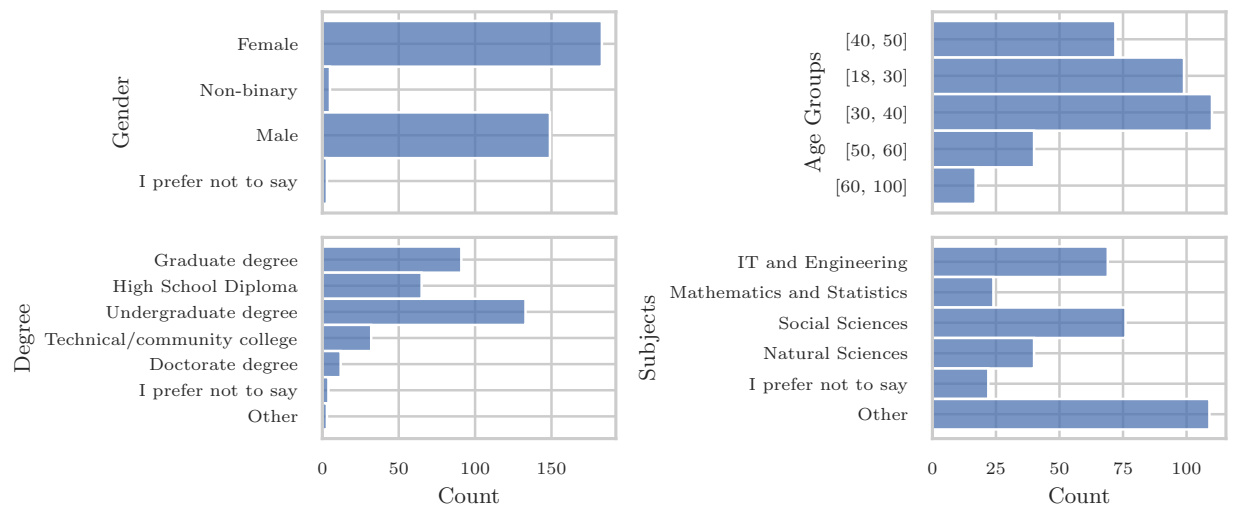


Figure S12: Demographic information provided by participants including gender, age, highest obtained degree and subject of degree.