ENHANCING AI USE: HOW COMPLEMENTARY SYSTEM INFORMATION DRIVES DELEGATION FREQUENCY AND EFFECTIVENESS

Appendix A: Updates to Pre-registration

During the process of analyzing the data and writing the paper, we agreed that changes to the wording of the hypotheses and the type of analyses conducted would be necessary to make our results and the paper clearer. Note that none of these changes alter the relationships we test or the results we produce; they only make it easier to communicate them.

The wording of our previously pre-registered hypotheses is as follows:

H1a: Displaying AI's certainty ex-ante a delegation decision has a negative effect on users' trust in AI if the user does not see the AI's actual performance ex-post.

H1b: Displaying AI's certainty ex-ante a delegation decision has a negative effect on the effectiveness of delegation to AI if the user does not see the AI's actual performance ex-post.

H2a: Observing AI perform ex-post a delegation decision on average has a negative effect on users' trust in AI if the user has not seen the AI's certainty ex-ante.

H2b: Observing AI perform ex-post a delegation decision on average has a negative effect on the effectiveness of delegation to AI if the user has not seen the AI's certainty ex-ante.

H3a: Providing users with AI's certainty ex-ante and AI's performance ex-post a delegation decision improves the calibration of trust in AI as compared to having less information on the AI's error boundaries.

H3b: Providing users with AI's certainty ex-ante and AI's performance ex-post a delegation decision increases the effectiveness of delegation as compared to having less information on the AI's error boundaries.

For Hypotheses 1a, 2a, and 3a we pre-registered delegation frequency as our measure for trust. In the manuscript we now directly use the wording of the measure, namely delegation frequency. For Hypothesis 3a, we changed the wording such that we no longer include the calibration of trust in AI but simply trust in our manuscript. However, we also include the analysis of the interaction between AI certainty before and AI performance after a delegation decision on trust calibration in Table 7 of the Appendix. We find support also for our previously pre-registered Hypothesis 3a regarding the calibration of trust. In order to have uniformly and consistently formulated effects for our hypotheses, we have decided to change the wording of our Hypothesis 3a in the manuscript.

Appendix B: Additional Analyses

B.1. Robustness Checks

We have conducted further mixed-effect logistic regressions in which we include binary treatment variables for treatments 2, 3, and 4. We find that when comparing treatment 4 to the baseline condition, the same positive effects found through testing for an interaction also hold. The results are presented in Table 1.

	Delegation Frequency	Delegation Effectiveness
Intercept	-2.133***	-4.557***
	(0.181)	(0.280)
T2	0.453**	0.240
	(0.196)	(0.174)
T3	-0.397**	-0.423**
	(0.198)	(0.178)
T4	0.645***	0.381**
	(0.199)	(0.177)
Log Likelihood	-21,501.5	-10,389.7
AIC	43,015.0	20,791.4
Number of observations	56,000	56,000

Significance levels: *** p<0.01, ** p<0.05, * p<0.1; Standard errors are in parentheses.

Table 1 Mixed-Effects Logistic Regression with Random Effects for Subjects and Images

Additionally, we present results based on the same mixed-effects logistic regression models as shown in the main paper (Table 2), but report standard errors obtained from a nonparametric cluster bootstrap that resamples participants (Subject ID) with replacement. We present the results in Table 2. This accounts for potential within-subject correlation due to repeated measures, providing inference that is robust to arbitrary within-subject dependence. Notably, the estimated coefficients remain unchanged, and the cluster-bootstrap standard errors are of similar magnitude to the model-based standard errors, indicating that the inclusion of

random intercepts for individuals already captures most of the within-subject dependence. This robustness check supports the reliability of the fixed effects reported in the main analysis.

Table 2 Mixed-Effects Logistic Regressions with Clustered Standard Errors

	Delegation Frequency	Delegation Effectiveness
Fixed Effects:		
AI Outcome	-0.397*	-0.423**
	(0.203)	(0.190)
AI Certainty	0.453***	0.240
	(0.176)	(0.162)
AI Certainty × AI Outcome	0.589*	0.565**
	(0.303)	(0.270)
Constant	-2.133***	-4.557***
	(0.121)	(0.118)
Log Likelihood	-21,501.5	-10,389.7
AIC	43,015.0	20,791.4
BIC	43,068.6	20,845.0
Number of Observations	56,000	56,000

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$. Standard errors in parentheses are cluster-bootstrap standard errors, resampling participants (Subject ID) with replacement ($B = 300$). Random intercepts for subjects and images included.

As an additional robustness check, we estimated ordinary least squares (OLS) regressions using the same predictors as in the logistic models. Although the dependent variables (Delegation Frequency and Delegation Effectiveness) are binary, the OLS specification allows us to assess whether the direction and relative magnitude of the effects are consistent across model types. The results, shown in Table 3, confirm the same pattern of effects and reinforce the main findings.

As a further robustness check, we re-estimated the mixed-effects models including additional individual-level control variables: Trust in AI, Interpersonal Trust, and Need for Control. These controls were selected based

Table 3 OLS Regression Results

	Delegation Frequency	Delegation Effectiveness
	(1)	(2)
AI Outcome	-0.020^{***}	-0.009^{***}
	(0.005)	(0.003)
AI Certainty	0.055^{***}	0.015^{***}
	(0.005)	(0.003)
AI Certainty × *AI Outcome*	0.062^{***}	0.021^{***}
	(0.007)	(0.004)
Constant	0.195^{***}	0.069^{***}
	(0.004)	(0.002)
Observations	56,000	56,000
R^2	0.012	0.003
Adjusted R^2	0.012	0.003
Residual Std. Error (df = 55996)	0.417	0.266
F Statistic (df = 3; 55996)	228.414^{***}	49.221^{***}

Note: *p<0.1; **p<0.05; ***p<0.01. OLS regression with binary dependent variables; included here as a robustness check for direction and relative magnitude of effects.

on their theoretical relevance for human–AI collaboration. The results are shown in Table 4 and confirm that the main effects and interactions of AI Outcome and AI Certainty remain robust and statistically significant even when controlling for individual dispositions.

B.2. Additional Results

In order to account for the panel data structure of our study, we include an analysis of potential learning effects within the experiment. To do this, we divided the dataset into two halves based on the order of image presentation: the first 50 images and the second 50 images for each participant. We then analyzed our main

Table 4 Mixed-Effects Model Results with Additional Controls (Robustness Check)

	Delegation Frequency	Delegation Effectiveness
Fixed Effects		
AI Outcome	-0.402**	-0.443***
	(0.188)	(0.169)
AI Certainty	0.381**	0.159
	(0.187)	(0.166)
AI Certainty × AI Outcome	0.496*	0.530**
	(0.267)	(0.239)
Trust in AI	0.380***	0.330***
	(0.064)	(0.058)
Interpersonal Trust	-0.215**	-0.195**
	(0.086)	(0.077)
Need for Control	-0.295***	-0.287***
	(0.083)	(0.074)
Constant	-1.994	-4.332***
	(1.249)	(1.168)
Controls	yes	yes
AIC	42988.700	20764.000
BIC	43283.500	21058.800
Log-Likelihood	-21461.300	-10349.000
Number of Observations	56,000	56,000

Note: *p<0.1; **p<0.05; ***p<0.01. Standard errors are in parentheses. AI Certainty and AI Outcome are binary variables indicating whether the information was shown (1) or not (0). Included controls: Trust in AI, Interpersonal Trust, Need for Control and demographic controls (gender, age, education, and income; coefficients not shown).

treatment effects separately for each sub-dataset, as reported in Tables 5 and 6. Our findings show that for the first 50 images, the main effects of seeing AI certainty and AI outcomes are largely stable, while the interaction between the two, although positive, is not statistically significant for delegation frequency or effectiveness. In contrast, for the second 50 images, the interaction between AI certainty and outcome becomes statistically significant and positive across both outcomes. This suggests that the combination of both transparency signals becomes more effective over time, potentially as users develop a better understanding of the AI's behavior.

Table 5 First 50 Images: Mixed-Effects Logistic Regression with Random Effects for Subjects and Images

	Delegation Frequency	Delegation Effectiveness
Fixed effects:		
AI Certainty	0.534***	0.342*
	(0.194)	(0.181)
AI Outcome	-0.362*	-0.382**
	(0.196)	(0.187)
AI Certainty × *AI Outcome*	0.430	0.412
	(0.276)	(0.259)
Constant	-2.244***	-4.537***
	(0.182)	(0.255)
AIC	21793.9	10210.5
BIC	21843.3	10259.9
Log Likelihood	-10891.0	-5099.2
Number of Observations	28002	28002

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$; Standard errors are in parentheses.

Appendix C: Experimental Details

C.1. Treatment Manipulations

Study 1 refers to the main experiment, Study 2 to Robustness Check 1 (Section 6.1), and Study 3 to Robustness Check 2 (Section 6.2). We show the different information on AI certainty and AI outcome which

Table 6 Second 50 Images: Mixed-Effects Logistic Regression with Random Effects for Subjects and Images

	Delegation Frequency	Delegation Effectiveness
Fixed Effects:		
Certainty	0.381*	0.199
	(0.213)	(0.190)
AI Outcome	-0.445**	-0.373*
	(0.216)	(0.194)
AI Certainty × AI Outcome	0.779**	0.641**
	(0.304)	(0.270)
Constant	-2.084***	-4.404***
	(0.193)	(0.271)
AIC	21901.9	11201.7
BIC	21951.4	11251.2
Log Likelihood	-10945.0	-5594.9
Number of Observations	27998	27998

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$; Standard errors are in parentheses.

we used for Study 1 and Study 3 in Figure 1 and Figure 2. For Study 2, the manipulation for AI certainty was displayed on a different screen as subjects had not classified the image themselves in advance. An example is displayed in Figure 3. The AI outcome is the same as in Study 1 and 3, as shown in Figure 2.

C.2. Item Scales

In the following we report the items adapted from the literature that we have used to measure potential control variables derived from the literature. The items were measured on a 7-point-Likert-scale ranging from strongly disagree to strongly agree if not stated otherwise.

Trust: Adapted from Yagoda and Gillan (2012) In the following, please tell us about your opinion of the AI used for the task. The Artificial Intelligence was:

- Reliable
- Dependable

Table 7 Mixed-Effects Logistic Regression with Random Effects for Subjects and Images (Trust Calibration)

	Trust Calibration
Fixed Effects:	
AI Outcome	-0.088
	(0.098)
AI Certainty	0.679***
	(0.098)
AI Outcome × *AI Certainty*	0.271*
	(0.139)
Constant	-1.293***
	(0.177)
Log Likelihood	-24848.5
AIC	49709.0
Number of Observations	56000

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$; Standard errors are in parentheses.

Figure 1 Exemplary Comparison of AI Certainty versus no AI Certainty Treatments (Study 1 and 3).

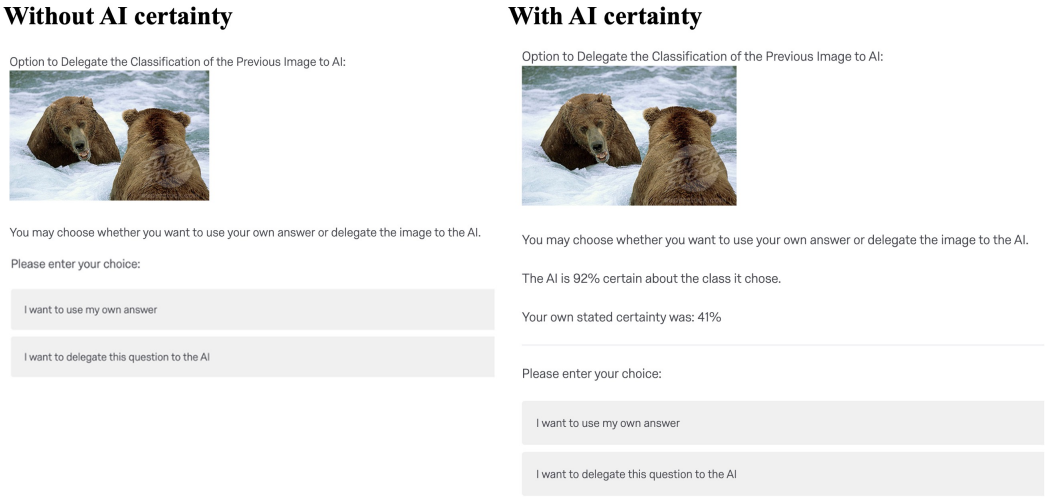


Figure 2 Example of AI Outcome (Study 1, 2, and 3).

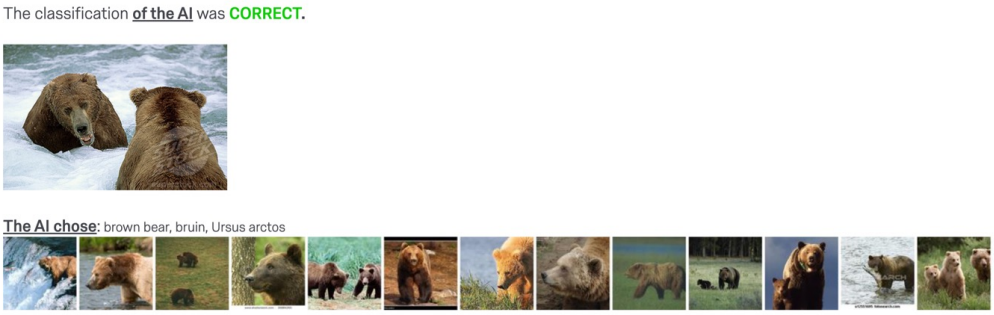
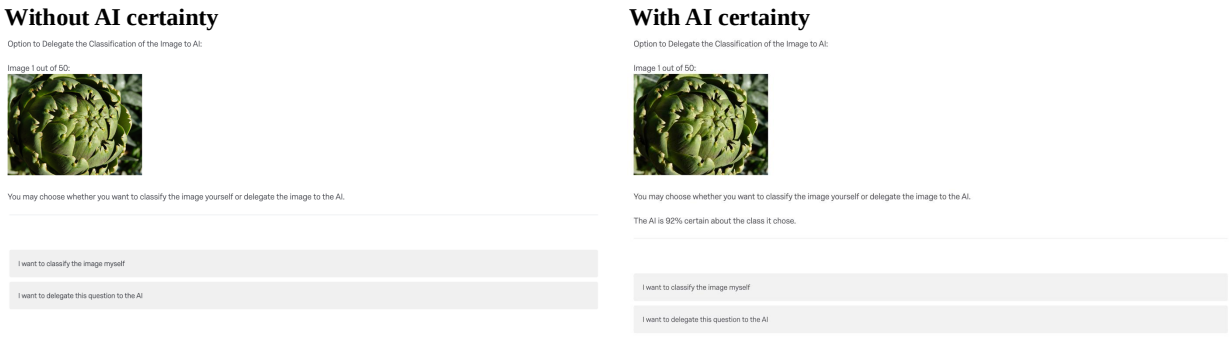


Figure 3 Example of AI Certainty (Study 2).



- Understandable
- Accessible

Interpersonal Trust: From Nießen et al. (2023) Please assess how much the following statements hold for you as a person.

- I am convinced that most people have good intentions
- You can't rely on anyone these days
- In general, people can be trusted

Need for Control: From de Bellis and Venkataramani Johar (2020)

- Most of the time I don't like giving things out of hand
- I like to be in control over the things happening around me
- Usually I prefer doing things by myself instead of delegating them
- I trust myself more than others
- Sometimes I'm afraid to lose control

Risk Aversion: Measured on a scale from zero (unwilling to take risks) to ten (fully prepared to take risks)

- How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?

Ability to Override Intuitive Responses (Cognitive Reflection Test): From (Thomson and Oppenheimer 2016)

- If you're running a race and you pass the person in second place, what place are you in?
- A farmer had 15 sheep and all but 8 died. How many are left?
- Emily's father has three daughters. The first two are named April and May. What is the third daughter's name?
- How many cubic feet of dirt are there in a hole that is 3' deep x 3' wide x 3' long?

References

de Bellis E, Venkataramani Johar G (2020) Autonomous shopping systems: Identifying and overcoming barriers to consumer adoption. *Journal of Retailing* 96(1):74–87.

Nießen D, Beierlein C, Rammstedt B, Lechner CM (2023) Interpersonal trust short scale (kusiv3). `https://zis.gesis.org/skala/Nie%C3%9Fen-Beierlein-Rammstedt-Lechner-Interpersonal-Trust-Short-Scale-(KUSIV3)`, accessed: (31.10.2023).

Thomson KS, Oppenheimer DM (2016) Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making* 11(1):1–15.

Yagoda RE, Gillan DJ (2012) You want me to trust a robot? the development of a human–robot interaction trust scale. *International Journal of Social Robotics* 4(3):235–248.