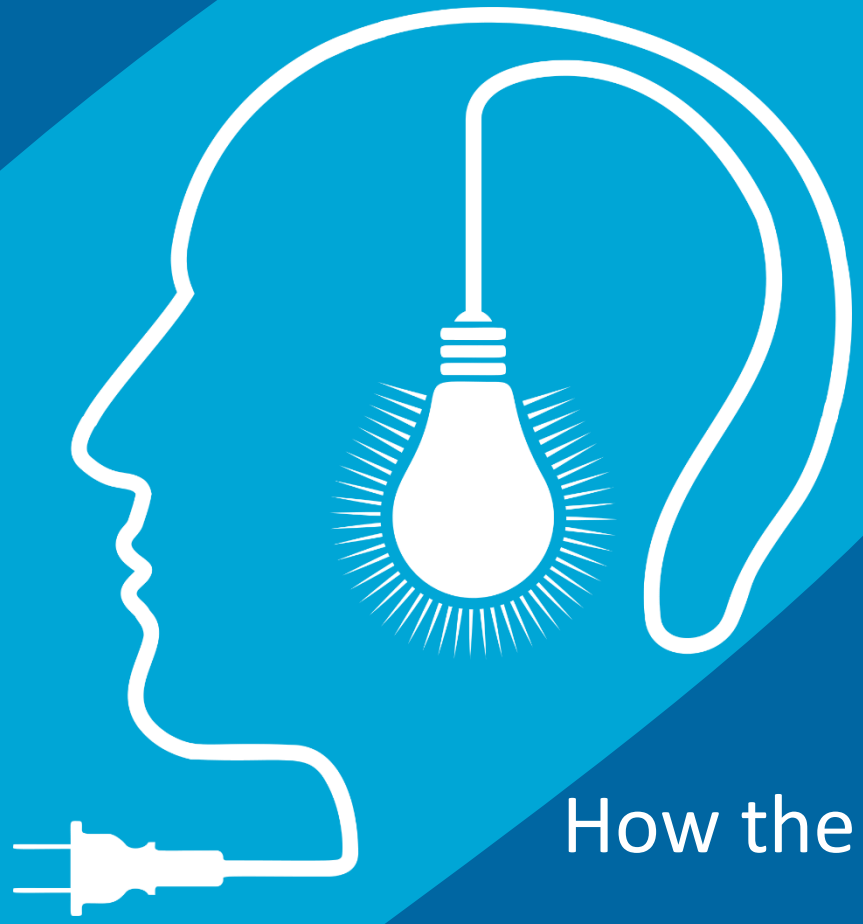


Knowing Better Than the AI

How the Dunning-Kruger Effect Shapes Reliance on Human-AI Decision Making

Lucie Kuiper 30-11-2022

TU Delft

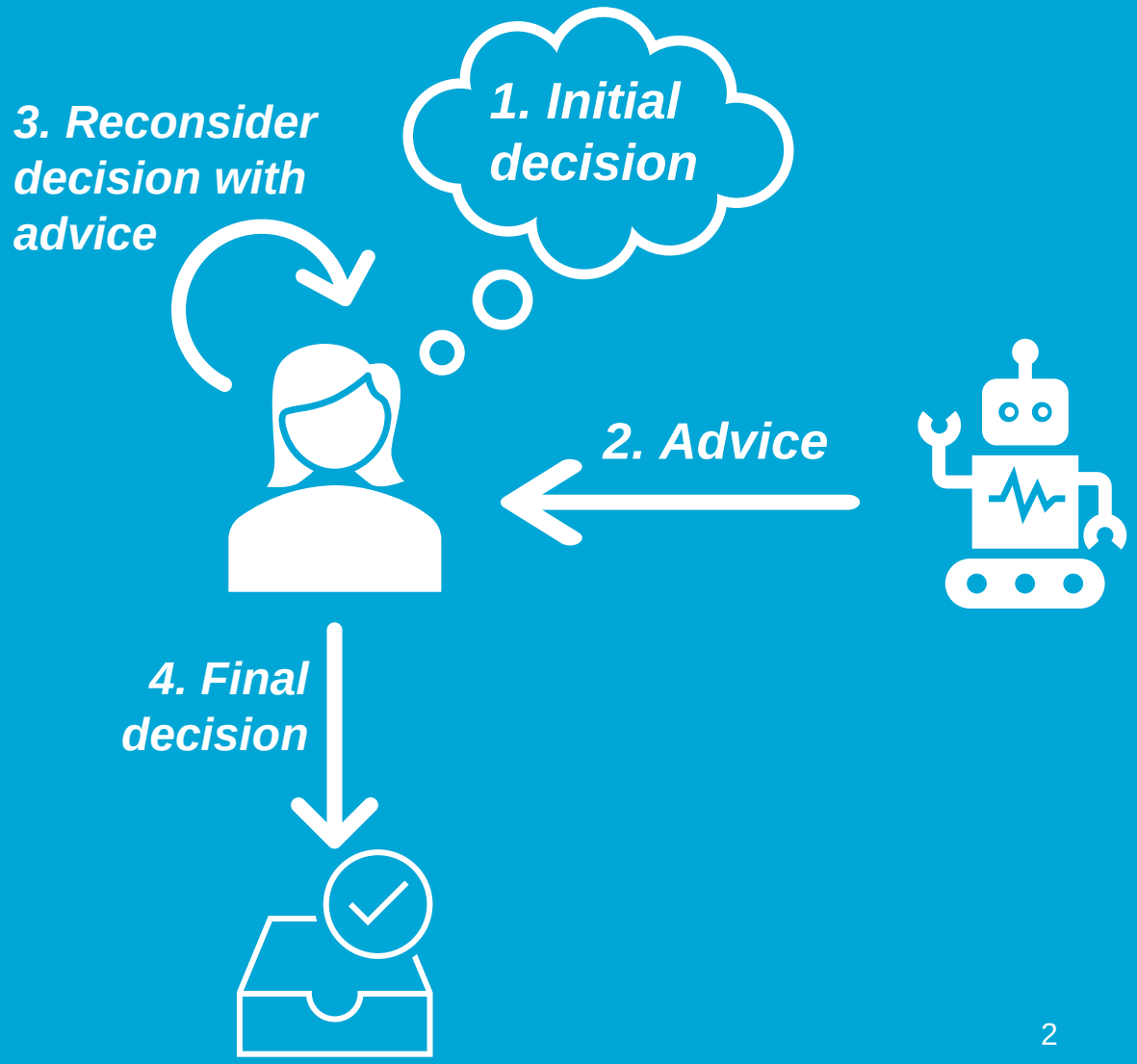
Chair:	Prof. Dr. G.J.P.M. Houben
Supervisor:	Dr. U.K. Gadiraju
Thesis committee:	Dr. M.L. Tielman
Daily supervisor:	MSc. G. He

Human-AI Decision Making

- Human and AI work together to get to shared goal
- High-stake decision making
  - Medical decisions
  - Granting out loans
- Ultimate goal: complementary performance
- Explainable AI (XAI)

TU Delft

A Decision Needs to Be Made



Reliance and Appropriate Reliance

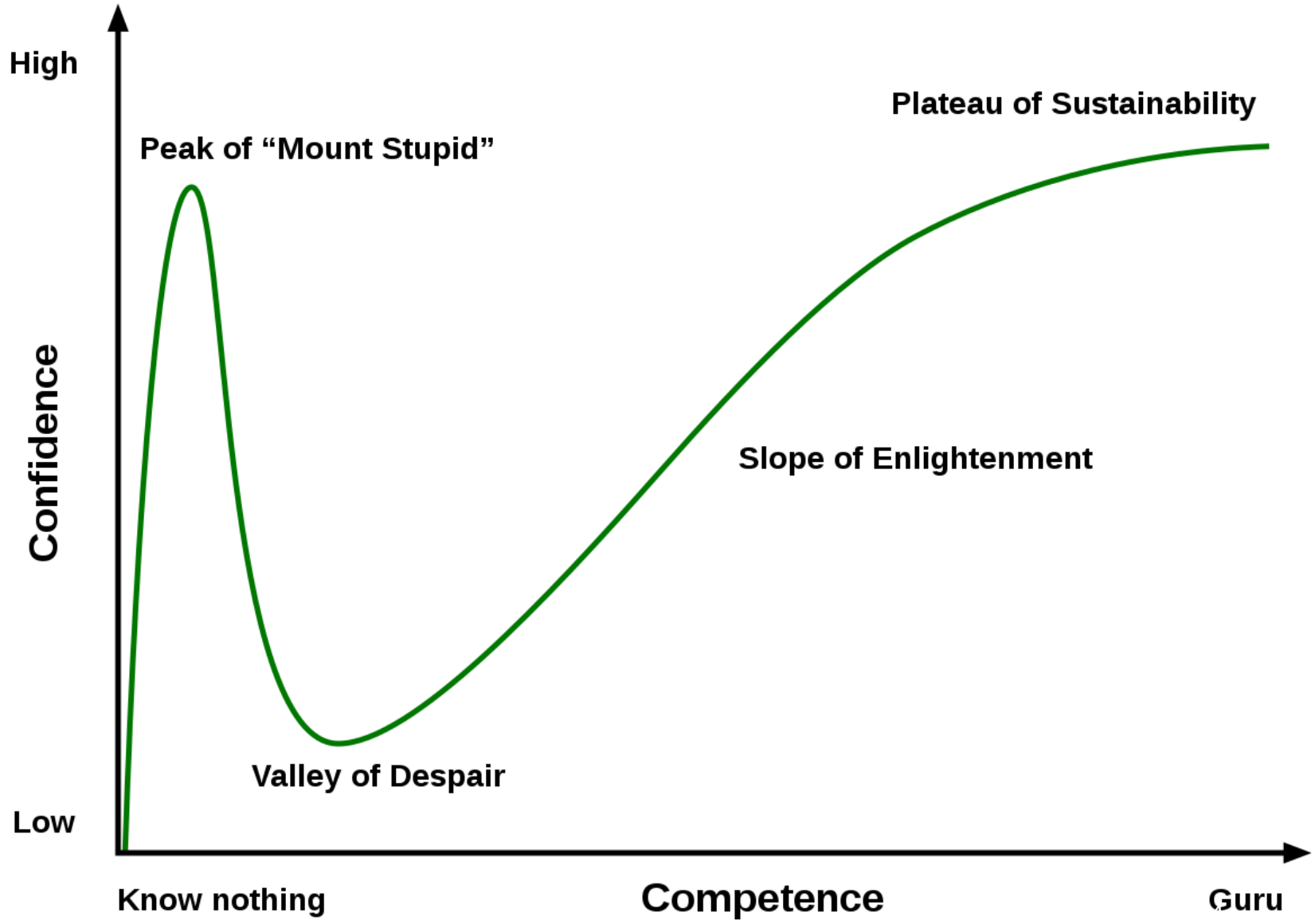
- Reliance on AI
- Appropriate reliance
  - Differentiating between correct and incorrect AI advice

- Self-assessment
  - How well do people think they perform
  - General tendency to overestimation

TU Delft

Dunning-Kruger Effect (DKE)

- People overestimate own performance when ability is low
- Effect on reliance on AI?
- How to mitigate DKE to an appropriate confidence?



TU Delft

Research questions:

How does the Dunning-Kruger effect shape reliance on AI systems?

How can this effect be mitigated?

TU Delft

Hypotheses

Hypothesis 1: *Participants overestimating their own performance rely less on AI systems*

Hypothesis 2: *Accurate self-assessment will result in more appropriate reliance on AI systems*

Hypothesis 3: *Making users aware of their miscalibrated self-assessment, will improve their self-assessment*

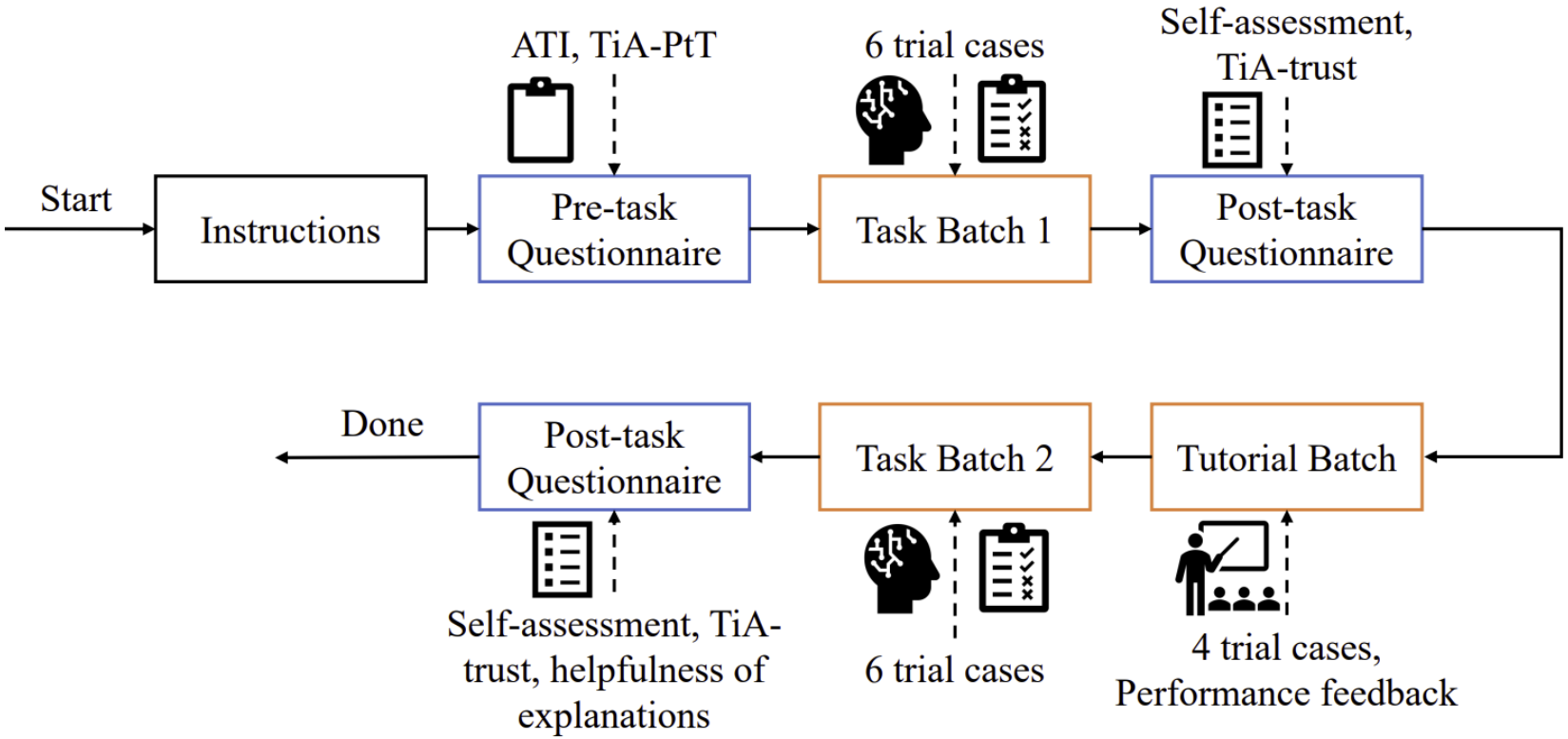
Hypothesis 4*: Making users aware of their miscalibrated self-assessment will result in more appropriate reliance on AI systems*

Effects will be amplified when participants are supplied with XAI

TU Delft

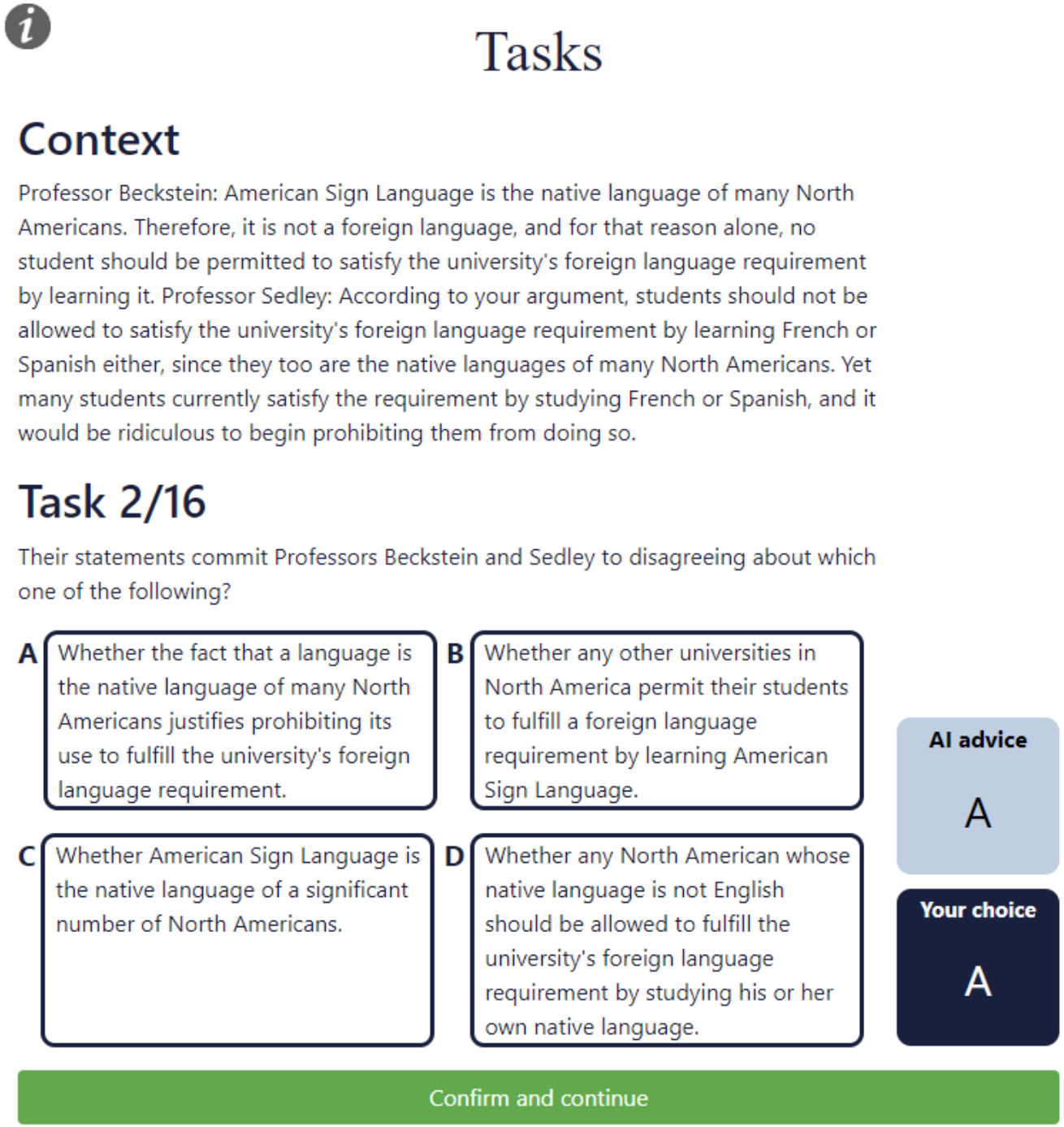
Experiment

249 participants



Task

- Logical reasoning questions
  - Layman friendly
  - Realistic human-AI team scenario
- Two stage decision-making
  - First participant alone only after initial answer AI advice provided



TU Delft

Tasks

Context

Professor Hartley's new book on moral philosophy contains numerous passages that can be found verbatim in an earlier published work by Hartley's colleague, Professor Lawrence. Therefore, in view of the fact that these passages were unattributed in Hartley's book, Hartley has been dishonest in not acknowledging the intellectual debt owed to Lawrence.

Task 2/16

Which one of the following is an assumption on which the argument is based?

A Hartley considered the passages in question to be the best possible expressions of the ideas they contain.

B Hartley could not have written the new book without the passages in question.

C A book on moral philosophy should contain only material representing the author's own convictions.

D Lawrence did not get the ideas in the passages in question or did not get their formulations originally from Hartley.

Confirm and continue

Tasks

Context

Professor Hartley's new book on moral philosophy contains numerous passages that can be found verbatim in an earlier published work by Hartley's colleague, Professor Lawrence. Therefore, in view of the fact that these passages were unattributed in Hartley's book, Hartley has been dishonest in not acknowledging the intellectual debt owed to Lawrence.

Task 2/16

Which one of the following is an assumption on which the argument is based?

A Hartley considered the passages in question to be the best possible expressions of the ideas they contain.

B Hartley could not have written the new book without the passages in question.

C A book on moral philosophy should contain only material representing the author's own convictions.

D Lawrence did not get the ideas in the passages in question or did not get their formulations originally from Hartley.

Confirm and continue

Tasks

Context

Professor Hartley's new book on moral philosophy contains numerous passages that can be found verbatim in an earlier published work by Hartley's colleague, Professor Lawrence. Therefore, in view of the fact that these passages were unattributed in Hartley's book, Hartley has been dishonest in not acknowledging the intellectual debt owed to Lawrence.

Task 2/16

Which one of the following is an assumption on which the argument is based?

A Hartley considered the passages in question to be the best possible expressions of the ideas they contain.

B Hartley could not have written the new book without the passages in question.

C A book on moral philosophy should contain only material representing the author's own convictions.

D Lawrence did not get the ideas in the passages in question or did not get their formulations originally from Hartley.

AI advice

D

Your choice

A

Confirm and continue

Tasks

Context

Professor Hartley's new book on moral philosophy contains numerous passages that can be found verbatim in an earlier published work by Hartley's colleague, Professor Lawrence. Therefore, in view of the fact that these passages were unattributed in Hartley's book, Hartley has been dishonest in not acknowledging the intellectual debt owed to Lawrence.

Task 2/16

Which one of the following is an assumption on which the argument is based?

A Hartley considered the passages in question to be the best possible expressions of the ideas they contain.

B Hartley could not have written the new book without the passages in question.

C A book on moral philosophy should contain only material representing the author's own convictions.

D Lawrence did not get the ideas in the passages in question or did not get their formulations originally from Hartley.

AI advice

D

Your choice

A

Confirm and continue

Tutorial

- Reduce miscalibration
  - Make aware of their mistakes
  - Show complexity of task
- Contrastive explanations

Tasks

Context

Because visual inspection cannot reliably distinguish certain skin discolorations from skin cancers, dermatologists at clinics have needed to perform tests of skin tissue taken from patients. At Westville Hospital, dermatological diagnostic costs were reduced by the purchase of a new imaging machine that diagnoses skin cancer in such cases as reliably as the tissue tests do. Consequently, even though the machine is expensive, a dermatological clinic in Westville is considering buying one to reduce diagnostic costs.

Correct answer

B

Task 9/16

Which of the following would it be most useful for the clinic to establish in order to make its decision?

AI advice

B

Your choice

A

A Whether the visits of patients who require diagnosis of skin discolorations tend to be shorter in duration at the clinic than at the hospital.

B Whether the machine at the clinic would get significantly less heavy use than the machine at the hospital does.

C Whether the principles on which the machine operates have been known to science for a long time.

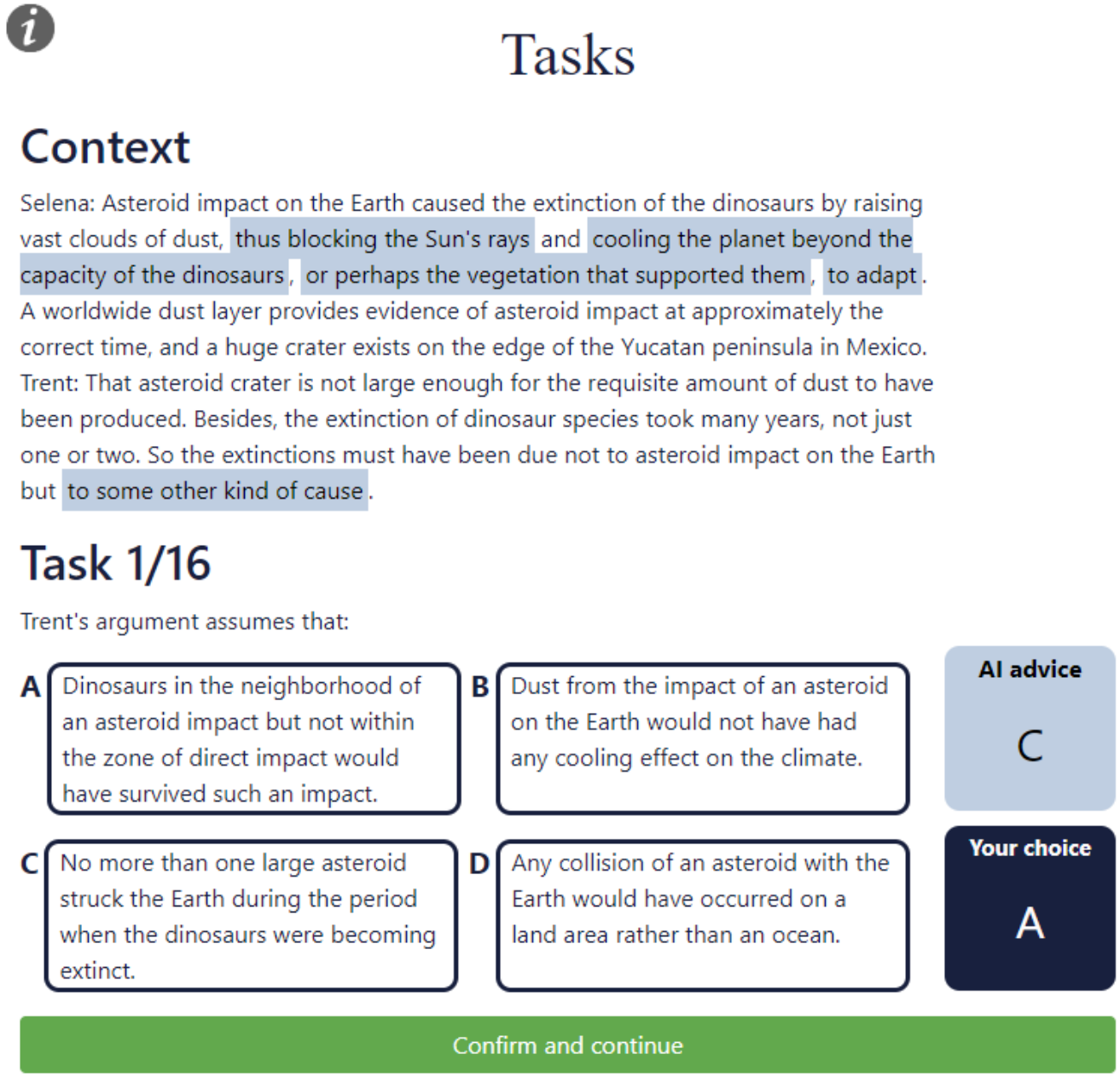
D Whether in certain cases of skin discoloration, visual inspection is sufficient to make a diagnosis of skin cancer.

The answer is B rather than A. A provides information about the visit duration, but does not directly consider cost, the clinic wants to base its decision based on reducing the costs. As the hospital managed to reduce costs with the machine, it is most important to consider this in the answer. Therefore it is most important to know the amount of use the machine would get in comparison to the machine in the hospital, making B the best answer.

TU Delft

XAI

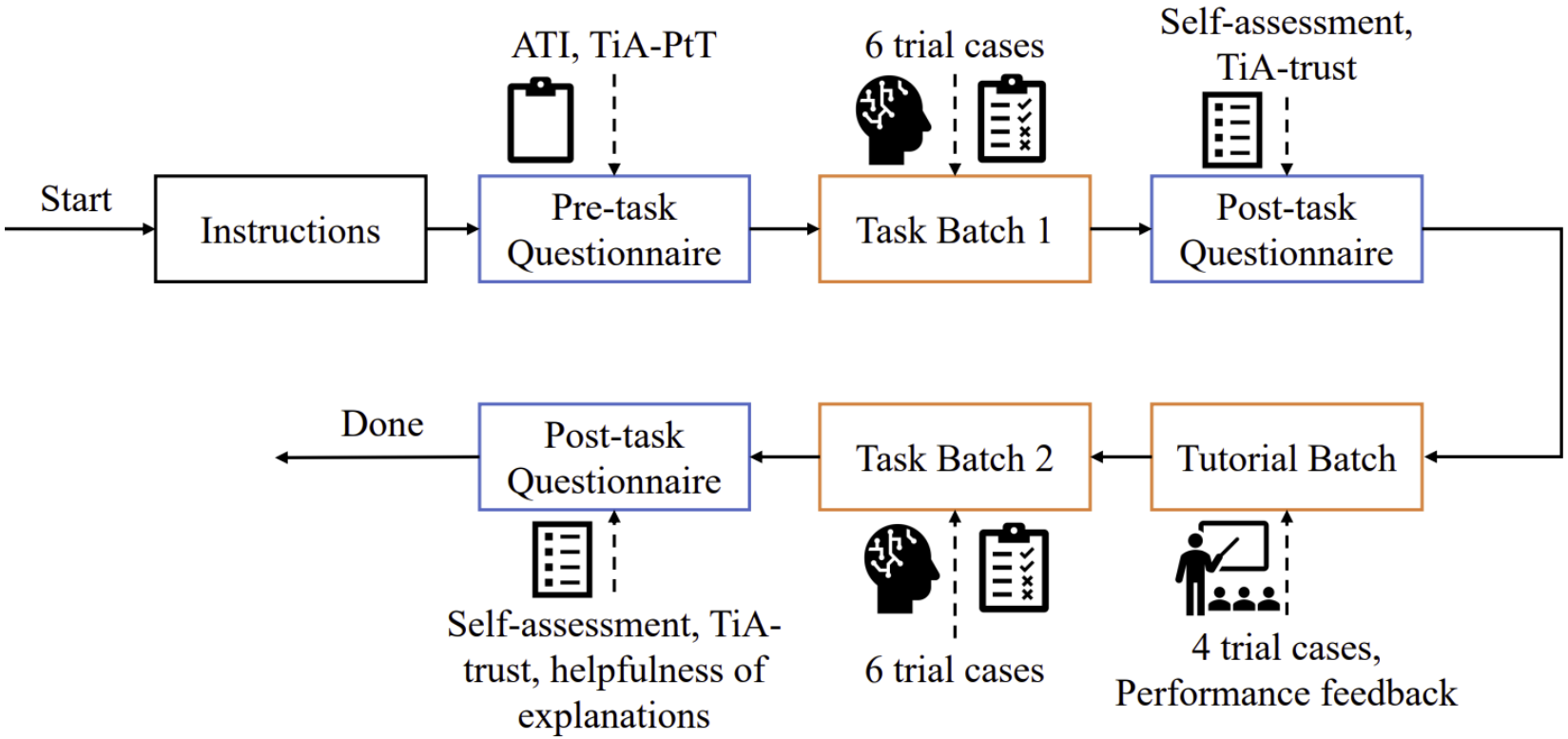
- To increase understanding of AI
- Logic units-based explanations
- Most important ones for AI advice highlighted
  - Top 5 displayed to participants
- LogiFormer model [1]



TU Delft

Experimental design

Four conditions (× tutorial, × XAI) , (✓ tutorial, × XAI), (× tutorial, ✓ XAI) and (✓ tutorial, ✓ XAI)



Measures

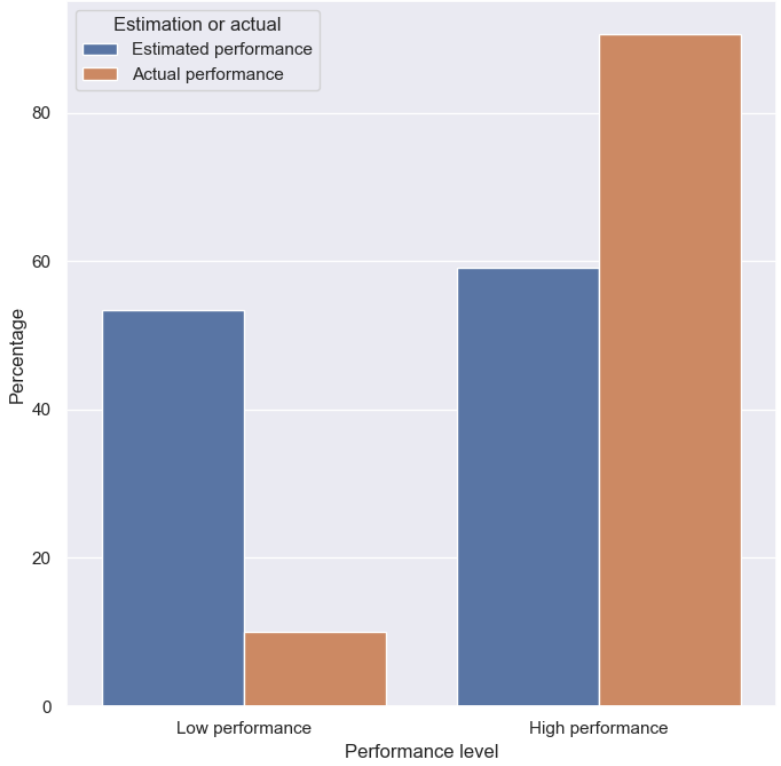
- Reliance
  - Agreement fraction $= \frac{Number\ of\ decisions\ same\ as\ the\ system}{Total\ number\ of\ decisions}$
  - Switch fraction $= \frac{Number\ of\ decisions\ where\ the\ user\ switched\ to\ agree\ with\ the\ system}{Total\ number\ of\ decisions\ with\ initial\ disagreement}$
- Appropriate reliance
  - RAIR $= \frac{Postivie\ AI\ reliance}{Positive\ AI\ reliance + Negative\ self-reliance}$
  - RSR $= \frac{Positive\ self-reliance}{Positive\ self-reliance + Negative\ AI\ reliance}$
- Accuracy $= \frac{Number\ of\ final\ decisions\ correct}{Total\ number\ of\ decisions}$
- Self-estimation
  - Degree of Miscalibration = (Number estimated correct answers – Actual number of correct answers)/6

TU Delft

Establishing the Dunning-Kruger Effect

- Bottom 20% put themselves at 53%
- Top 19% put themselves at 59%

- Lower performance participants overestimate themselves
  - DKE is present in experiments



Hypothesis 1

- Overestimating → degree of miscalibration > 0
  - 101 overestimation
  - 148 no overestimation
- Reliance on AI *Wilcoxon-Mann Whitney test*

Dependent Variables	U	p	$M \pm SD$(Overestimation)	$M \pm SD$(Other)
Agreement Fraction	5683.0	**.0010**	0.594 ± 0.237	0.695 ± 0.198
Switch Fraction	4809.5	**<.0001**	0.325 ± 0.323	0.515 ± 0.298

- No significant difference XAI *Wilcoxon-Mann Whitney test*

Participants	With Overestimation				Without Overestimation			
Dependent Variables	U	p	$M \pm SD$(XAI)	$M \pm SD$(no XAI)	U	p	$M \pm SD$(XAI)	$M \pm SD$(No XAI)
Agreement Fraction	1071.5	.1710	0.562 ± 0.243	0.631 ± 0.228	2624.5	.6778	0.688 ± 0.198	0.701 ± 0.198
Switch Fraction	1087.5	.2071	0.285 ± 0.300	0.371 ± 0.345	2471.5	.3184	0.489 ± 0.308	0.538 ± 0.289

TU Delft

Hypothesis 2

- Accurate self-assessment → degree of miscalibration = 0
  - 76 accurate
  - 173 inaccurate
- Appropriate reliance on AI *Wilcoxon-Mann Whitney test*

Dependent Variables	U	p	$M \pm SD$(Accurate)	$M \pm SD$(Inaccurate)
RAIR	7966.5	**.0062**	0.577 ± 0.375	0.431 ± 0.393
RSR	6473.0	.8284	0.421 ± 0.476	0.436 ± 0.481

- No significant difference XAI *Wilcoxon-Mann Whitney test*

Participants	Accurate self-estimation				Inaccurate self-estimation			
Dependent Variables	U	p	$M \pm SD$(XAI)	$M \pm SD$(no XAI)	U	p	$M \pm SD$(XAI)	$M \pm SD$(No XAI)
RAIR	568.0	.1282	0.500 ± 0.384	0.636 ± 0.361	1330.5	**.0022**	0.412 ± 0.393	0.451 ± 0.396
RSR	712.0	.9812	0.424 ± 0.486	0.419 ± 0.475	3350.0	.1912	0.390 ± 0.470	0.489 ± 0.490

TU Delft

Hypothesis 3

- Comparison before and after tutorial
  - 87 participants
- Miscalibration *Wilcoxon signed ranks test*

Dependent Variable	T	p	$M \pm SD$(First batch)	$M \pm SD$(Last batch)
Mis-calibration	365.0	**.0005**	0.278 ± 0.152	0.190 ± 0.175

- No significant difference XAI *Wilcoxon-Mann Whitney test*

Dependent Variable	U	p	$M \pm SD$(XAI)	$M \pm SD$(No XAI)
Mis-calibration last batch of tasks	868.5	.4803	0.167 ± 0.120	0.213 ± 0.210
Difference in mis-calibration	862.0	.4601	-0.114 ± 0.203	-0.062 ± 0.236

- No significant difference to no tutorial *Wilcoxon-Mann Whitney test*

Dependent Variable	U	p	$M \pm SD$(Tutorial)	$M \pm SD$(No tutorial)
Mis-calibration last batch of tasks	3386.5	.252	0.190 ± 0.175	0.203 ± 0.150
Difference in mis-calibration	3644.0	.7615	-0.088 ± 0.220	-0.079 ± 0.202

TU Delft

Hypothesis 4

- Comparison before and after tutorial
  - 87 participants
- Miscalibration *Wilcoxon signed ranks test*

Dependent Variables	T	p	$M \pm SD$(First six)	$M \pm SD$(Last six)
RAIR	899.0	.5860	0.431 ± 0.405	0.400 ± 0.373
RSR	579.0	.3967	0.477 ± 0.494	0.425 ± 0.479

- No significant difference XAI *Wilcoxon-Mann Whitney test*

Participants	First Batch of Questions				Last Batch of Questions			
Dependent Variables	U	p	$M \pm SD$(XAI)	$M \pm SD$(no XAI)	H	p	$M \pm SD$(XAI)	$M \pm SD$(No XAI)
RAIR	904.5	.7179	0.419 ± 0.412	0.444 ± 0.402	915.5	.7931	0.386 ± 0.365	0.415 ± 0.385
RSR	788.5	.1296	0.398 ± 0.492	0.558 ± 0.470	1002.5	.5909	0.455 ± 0.492	0.395 ± 0.470

- No significant difference to no tutorial *Wilcoxon-Mann Whitney test*

Dependent Variables	U	p	$M \pm SD$(Tutorial)	$M \pm SD$(No Tutorial)
RAIR	3094.0	.0443	0.400 ± 0.373	0.516 ± 0.382
RSR	3847.0	.7143	0.421 ± 0.476	0.401 ± 0.484

TU Delft

Summary of results

Hypothesis 1: *Participants overestimating their own performance rely less on AI systems*	Supported by results found
Hypothesis 2: *Accurate self-assessment will result in more appropriate reliance on AI systems*	Partially supported by results found
Hypothesis 3: *Making users aware of their miscalibrated self-assessment, will improve their self-assessment*	Supported by results found
Hypothesis 4*: Making users aware of their miscalibrated self-assessment will result in more appropriate reliance on AI systems*	Not supported by results found

TU Delft

Discussion

- Biases
  - self-interest bias
- Tutorial subject to inconsistency
- How will results transfer
  - Domain with experts
  - Different type of tasks

Conclusions

- *How does the Dunning-Kruger effect shape reliance on AI systems?*
  - Group with DKE rely less on the AI system
- *How can this effect be mitigated?*
  - Tutorial group reduced miscalibration of self-assessment

- Future work
  - How to promote appropriate reliance
  - Test other means of XAI

TU Delft

Thank you

TU Delft

References

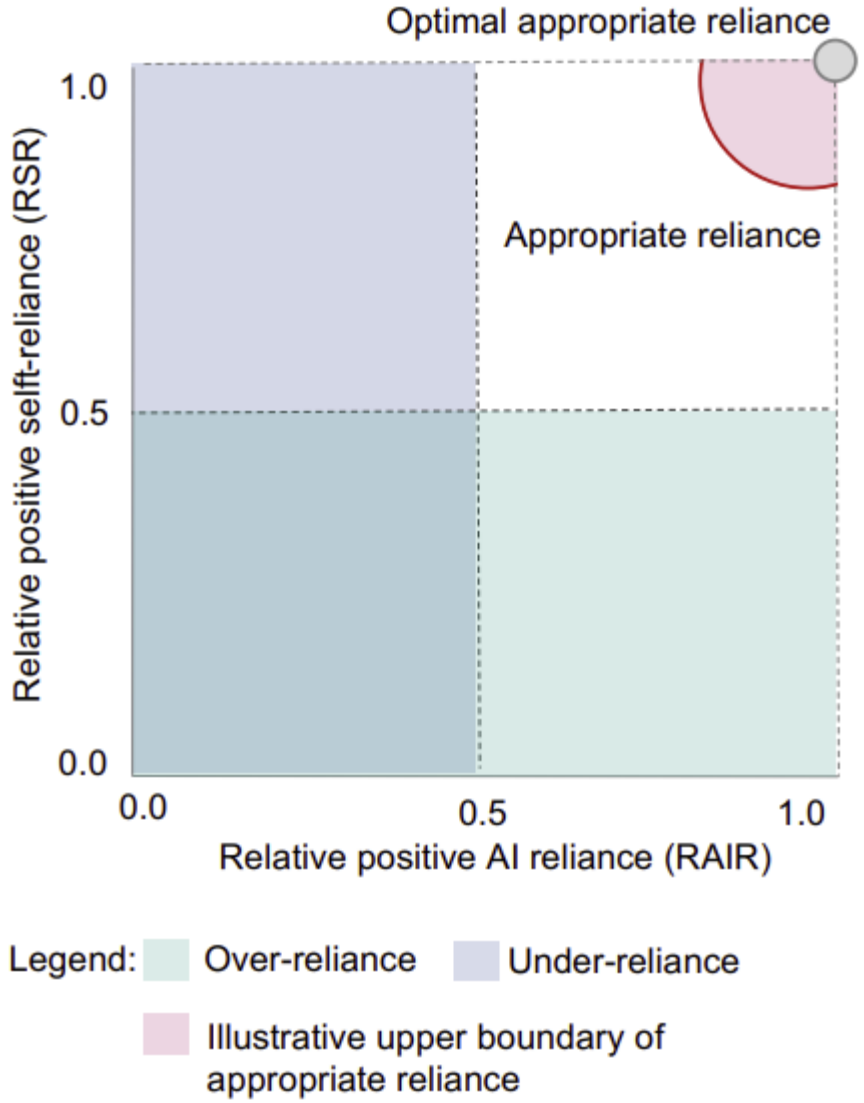
- [1] F. Xu, J. Liu, Q. Lin, Y. Pan, and L. Zhang, Logiformer: A two-branch graph transformer network for interpretable logical reasoning, in SIGIR ’22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, edited by E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, and G. Kazai (ACM, 2022) pp. 1055–1065.
- [2] M. Schemmer, P. Hemmer, N. Kühl, C. Benz, and G. Satzger, Should i follow ai-based advice? Measuring appropriate reliance in human-ai decision-making, in ACM Conference on Human Factors in Computing Systems (CHI’22), Workshop on Trust and Reliance in AI-Human Teams (trAIt) (2022).

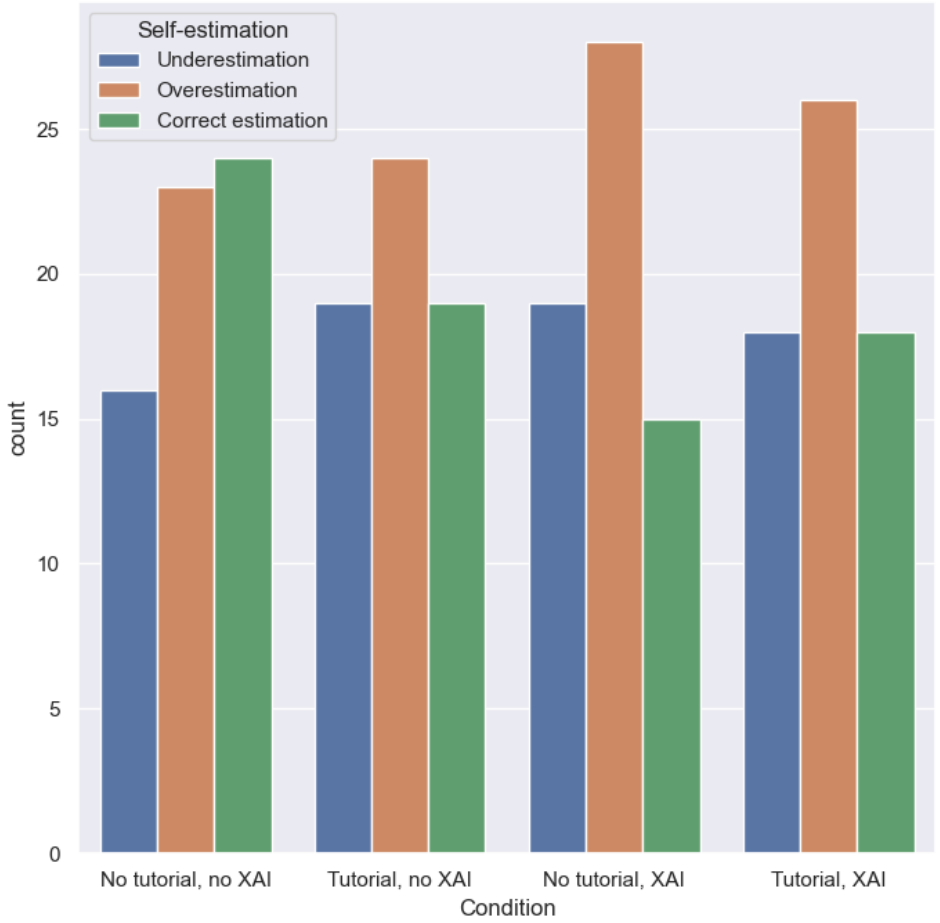
Variables

Variable Type	Variable Name	Value Type	Value Scale
Performance (DV)	Accuracy	Continuous, Interval	[0.0, 1.0]
Reliance (DV)	Agreement Fraction	Continuous, Interval	[0.0, 1.0]
	Switch Fraction	Continuous, Interval	[0.0, 1.0]
	RAIR	Continuous, Interval	[0.0, 1.0]
	RSR	Continuous, Interval	[0.0, 1.0]
Assessment (DV)	Degree of Miscalibration	Continuous, Interval	[-6,6]
	Self-Estimation	Continuous, Interval	[0,6]
	Estimation of Others	Continuous, Interval	[0,6]
	Comparison of Self to Others	Continuous, Interval	[0,100]
Trust (DV)	TiA-Trust	Likert	5-point, 1:strong distrust, 5: strong trust
Covariates	ATI	Likert	6-point, 1: low, 6: high
	TiA-PtT	Likert	5-point, 1: tend to distrust, 5: tend to trust
Other	Helpfulness of Explanation	Likert	5-point, 1: not helpful, 5: very helpful

TU Delft

Initial user decision	AI advice	Final decision after AI advice	Reliance
Incorrect	Correct	Correct	Positive AI reliance
Incorrect	Correct	Incorrect	Negative self reliance
Correct	Incorrect	Correct	Positive self reliance
Correct	Incorrect	Incorrect	Negative AI reliance




Self-estimation
Underestimation
Overestimation
Correct estimation
count
0
5
10
15
20
25
No tutorial, no XAI
Tutorial, no XAI
No tutorial, XAI
Tutorial, XAI
Condition

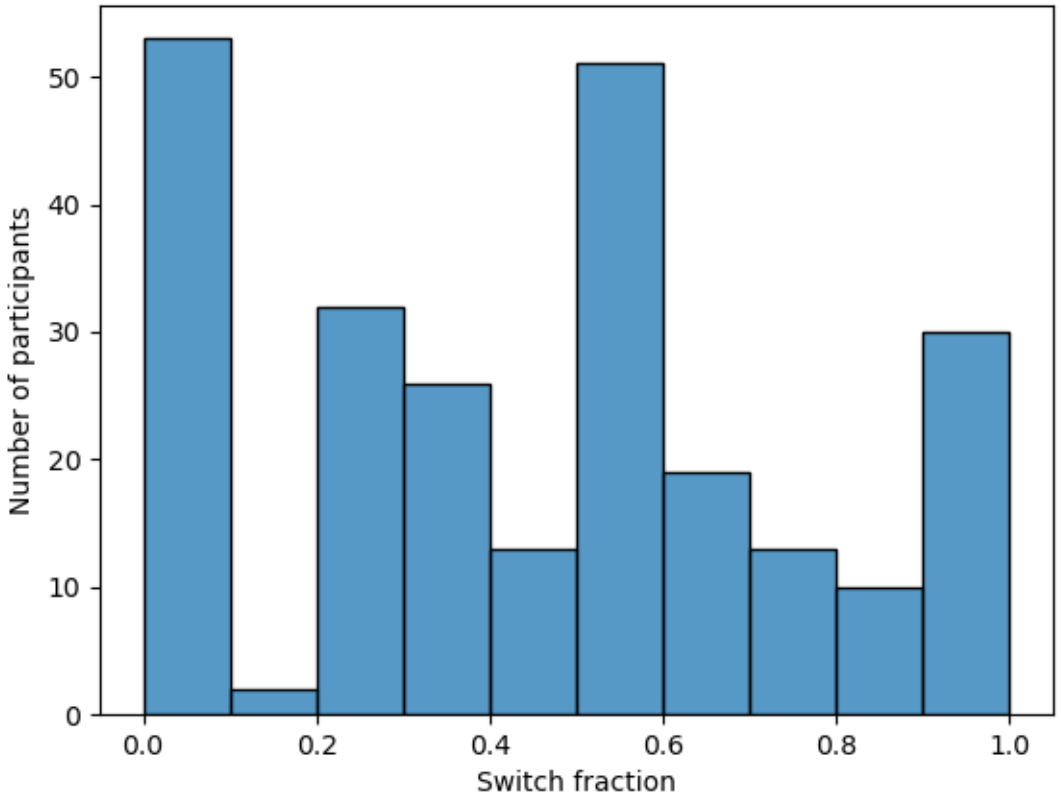
TU Delft

Dependent Variables	U	p	$M \pm SD$(Overestimation)	$M \pm SD$(Other)
Accuracy	3229.0	**<.0001**	0.449 ± 0.190	0.660 ± 0.168
Agreement Fraction	5683.0	**.0010**	0.594 ± 0.237	0.695 ± 0.198
Switch Fraction	4809.5	**<.0001**	0.325 ± 0.323	0.515 ± 0.298
RAIR	3811.0	**<.0001**	0.273 ± 0.332	0.613 ± 0.372
RSR	5241.0	**<.0001**	0.267 ± 0.433	0.544 ± 0.477

Dependent Variables	U	p	$M \pm SD$(Accurate)	$M \pm SD$(Inaccurate)
Accuracy	7269.0	.1703	0.605 ± 0.153	0.561 ± 0.223
Agreement Fraction	7393.0	.1090	0.691 ± 0.214	0.638 ± 0.221
Switch Fraction	8087.5	**.0035**	0.526 ± 0.311	0.399 ± 0.319
RAIR	7966.5	**.0062**	0.577 ± 0.375	0.431 ± 0.393
RSR	6473.0	.8284	0.421 ± 0.476	0.436 ± 0.481

TU Delft

Dependent Variables	T	p	$M \pm SD$(First six)	$M \pm SD$(Last six)
Accuracy	942.5	.1066	0.575 ± 0.223	0.527 ± 0.217
Agreement Fraction	1222.0	.9041	0.630 ± 0.222	0.625 ± 0.240
Switch Fraction	1251.5	.3591	0.378 ± 0.327	0.414 ± 0.337
RAIR	899.0	.5860	0.431 ± 0.405	0.400 ± 0.373
RSR	579.0	.3967	0.477 ± 0.494	0.425 ± 0.479


Number of participants
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
Switch fraction

TU Delft