

Knowing Better Than the AI

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

Lucie Kuiper 30-11-2022

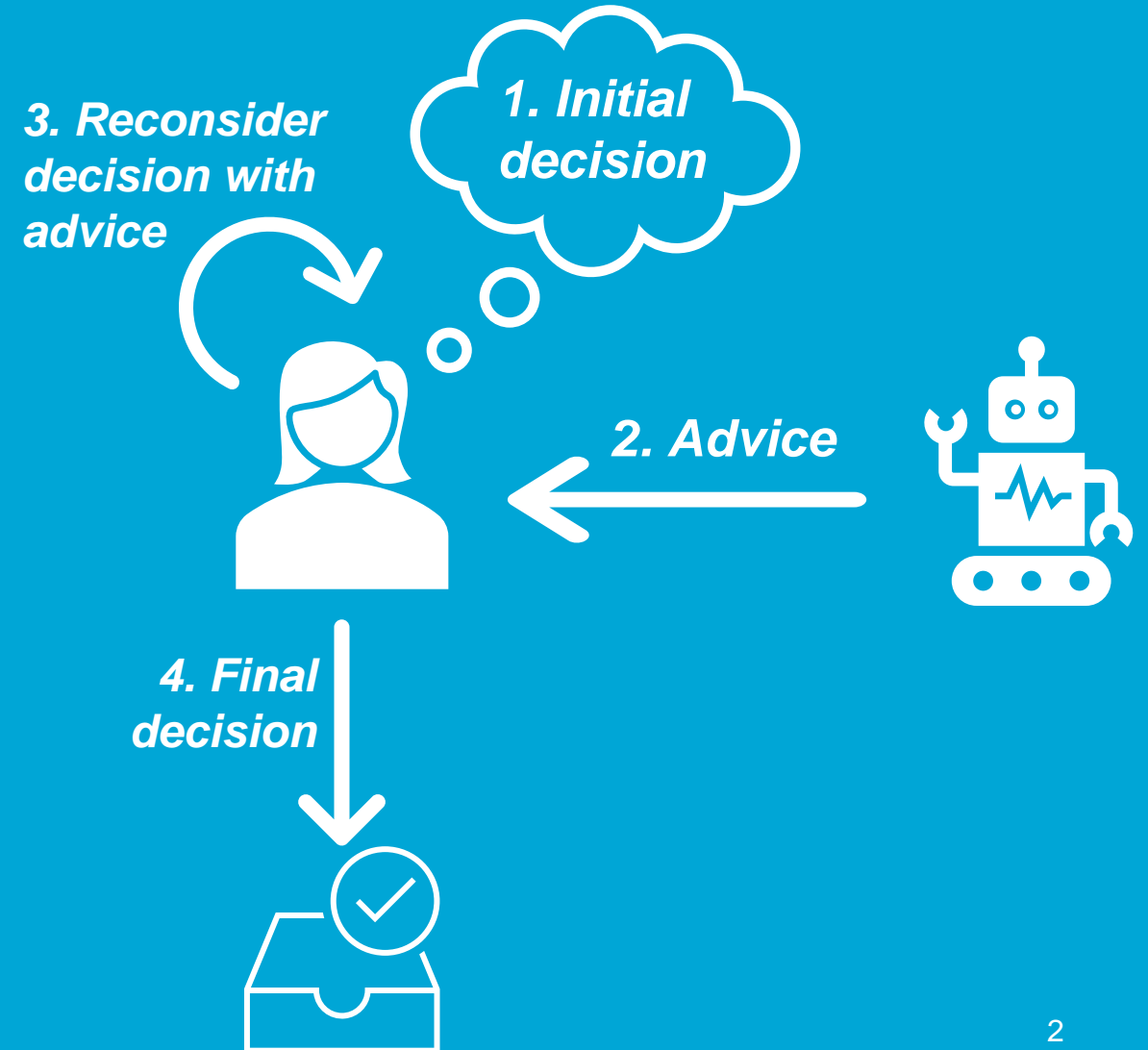
Chair:
Supervisor:
Thesis committee:
Daily supervisor:

Prof. Dr. G.J.P.M. Houben
Dr. U.K. Gadiraju
Dr. M.L. Tielman
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)

A Decision Needs to Be Made

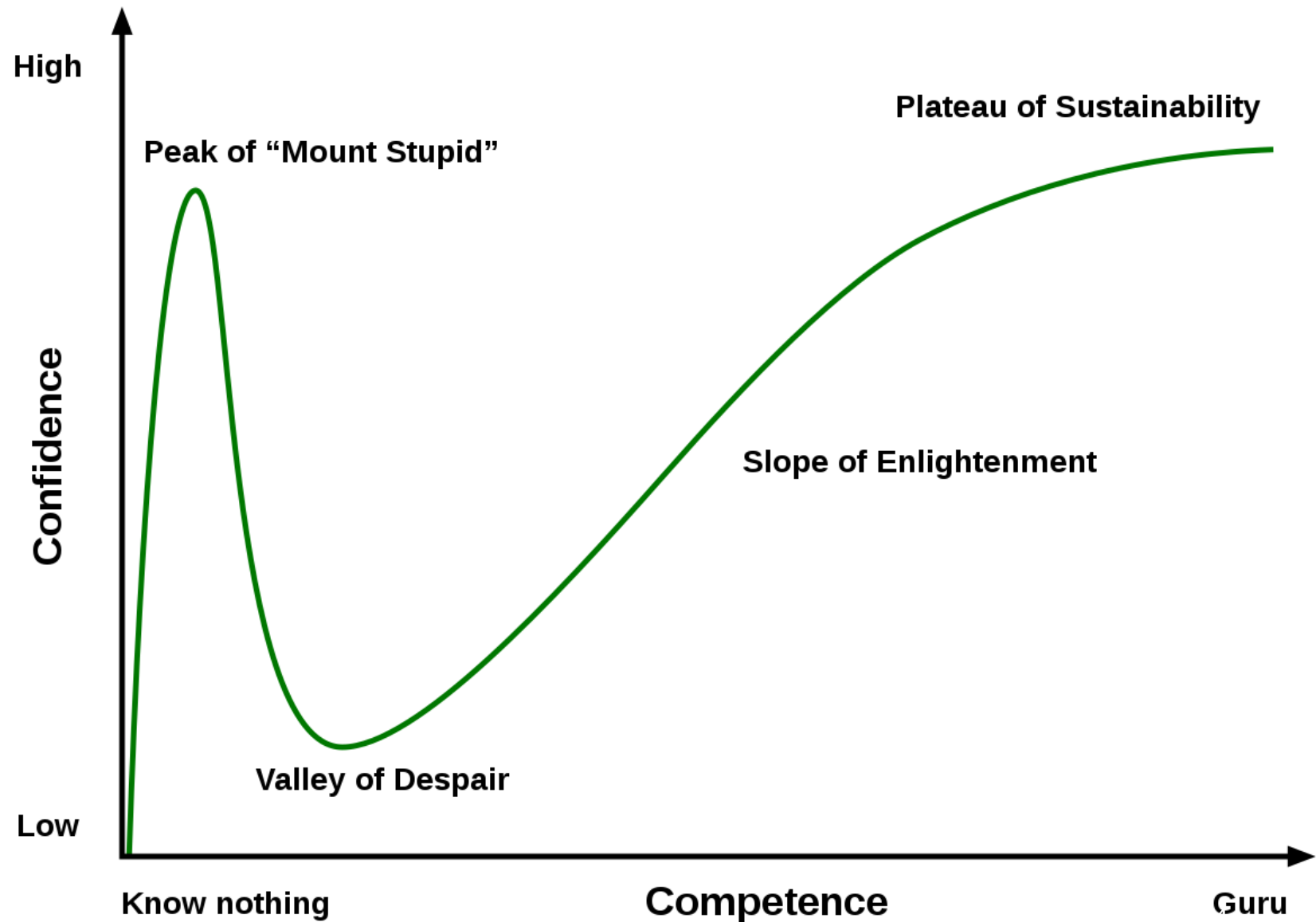


Reliance and Appropriate Reliance

- 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

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?



Research questions:

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

How can this effect be mitigated?

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*

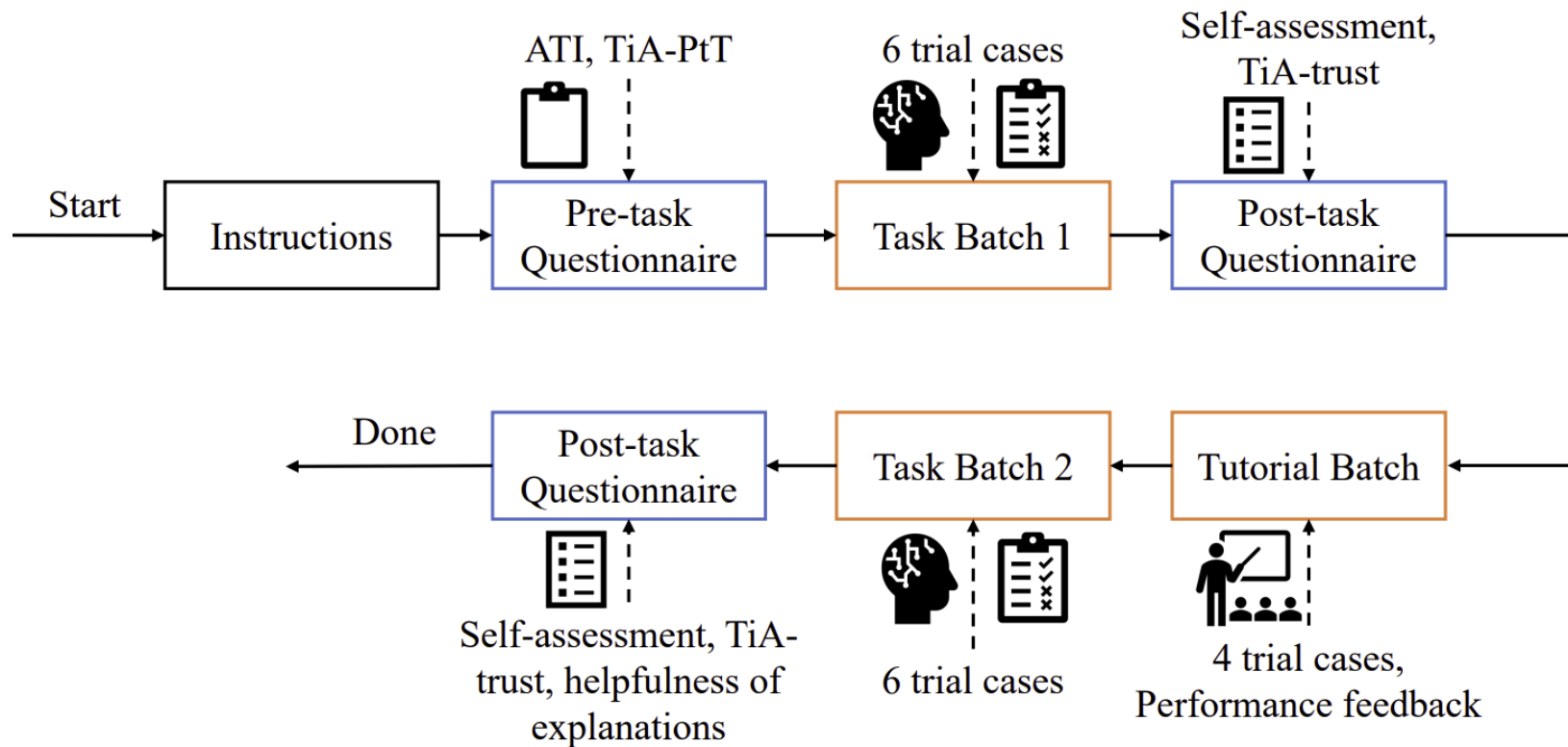
Hypothesis 3: *Making users aware of their miscalibrated self-assessment, will improve their self-assessment*

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

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



Tasks

Context

Professor Beckstein: American Sign Language is the native language of many North Americans. Therefore, it is not a foreign language, and for that reason alone, no student should be permitted to satisfy the university's foreign language requirement by learning it. Professor Sedley: According to your argument, students should not be allowed to satisfy the university's foreign language requirement by learning French or Spanish either, since they too are the native languages of many North Americans. Yet many students currently satisfy the requirement by studying French or Spanish, and it would be ridiculous to begin prohibiting them from doing so.

Task 2/16

Their statements commit Professors Beckstein and Sedley to disagreeing about which one of the following?

- | | |
|--|---|
| A Whether the fact that a language is the native language of many North Americans justifies prohibiting its use to fulfill the university's foreign language requirement. | B Whether any other universities in North America permit their students to fulfill a foreign language requirement by learning American Sign Language. |
| C Whether American Sign Language is the native language of a significant number of North Americans. | D Whether any North American whose native language is not English should be allowed to fulfill the university's foreign language requirement by studying his or her own native language. |

AI advice

A

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.

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.

Task 9/16

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

- 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.

Correct answer

B

AI advice

B

Your choice

A

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.

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]



Tasks

Context

Selena: Asteroid impact on the Earth caused the extinction of the dinosaurs by raising vast clouds of dust, thus blocking the Sun's rays and cooling the planet beyond the capacity of the dinosaurs, or perhaps the vegetation that supported them, to adapt. A worldwide dust layer provides evidence of asteroid impact at approximately the correct time, and a huge crater exists on the edge of the Yucatan peninsula in Mexico. Trent: That asteroid crater is not large enough for the requisite amount of dust to have been produced. Besides, the extinction of dinosaur species took many years, not just one or two. So the extinctions must have been due not to asteroid impact on the Earth but to some other kind of cause.

Task 1/16

Trent's argument assumes that:

- | | | | |
|----------|--|----------|--|
| A | Dinosaurs in the neighborhood of an asteroid impact but not within the zone of direct impact would have survived such an impact. | B | Dust from the impact of an asteroid on the Earth would not have had any cooling effect on the climate. |
| C | No more than one large asteroid struck the Earth during the period when the dinosaurs were becoming extinct. | D | Any collision of an asteroid with the Earth would have occurred on a land area rather than an ocean. |

AI advice

C

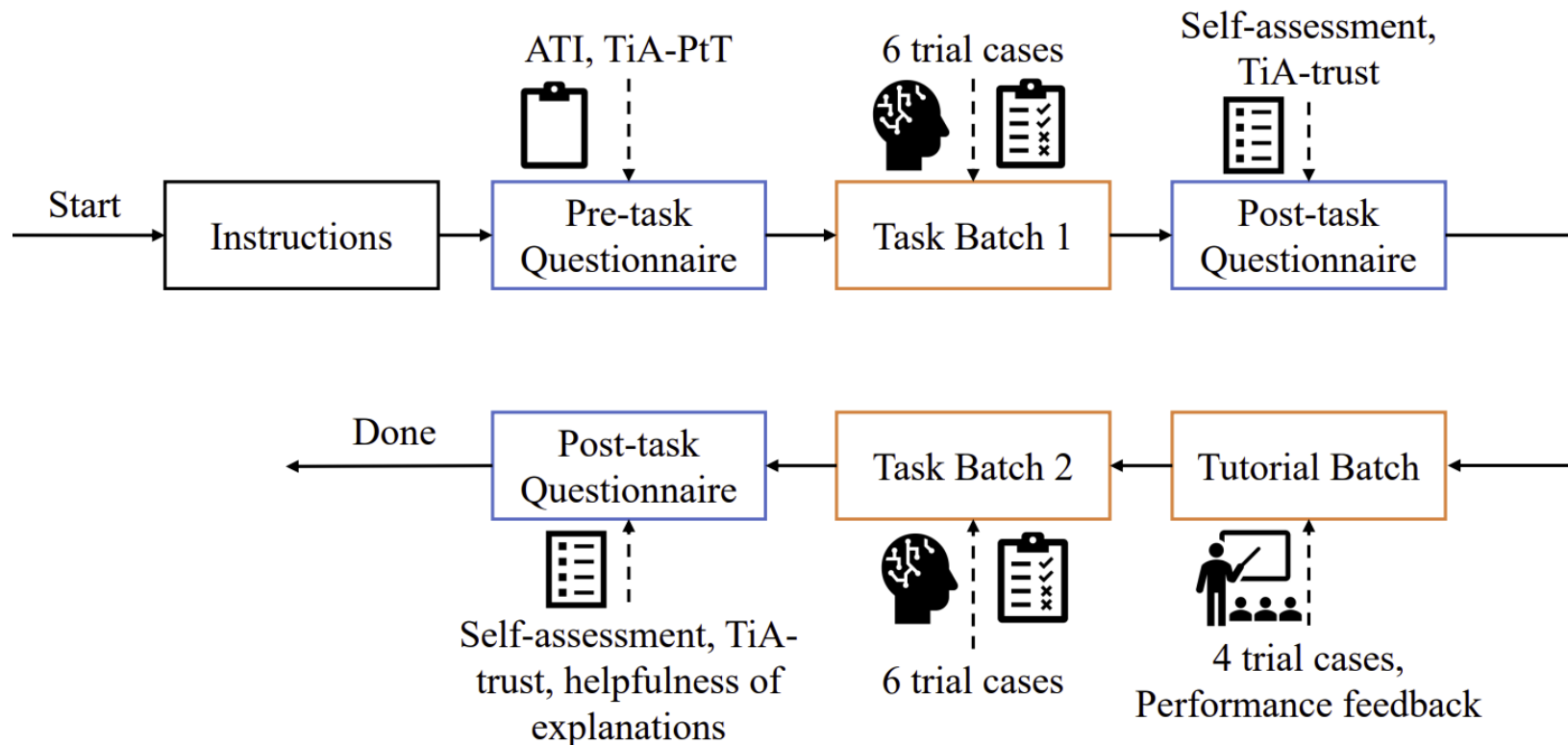
Your choice

A

Confirm and continue

Experimental design

Four conditions (\times tutorial, \times XAI) , (\checkmark tutorial, \times XAI), (\times tutorial, \checkmark XAI) and (\checkmark tutorial, \checkmark XAI)



Measures

- Reliance

- Agreement fraction = $\frac{\text{Number of decisions same as the system}}{\text{Total number of decisions}}$

- Switch fraction = $\frac{\text{Number of decisions where the user switched to agree with the system}}{\text{Total number of decisions with initial disagreement}}$

- Appropriate reliance

- RAIR = $\frac{\text{Positive AI reliance}}{\text{Positive AI reliance} + \text{Negative self-reliance}}$

- RSR = $\frac{\text{Positive self-reliance}}{\text{Positive self-reliance} + \text{Negative AI reliance}}$

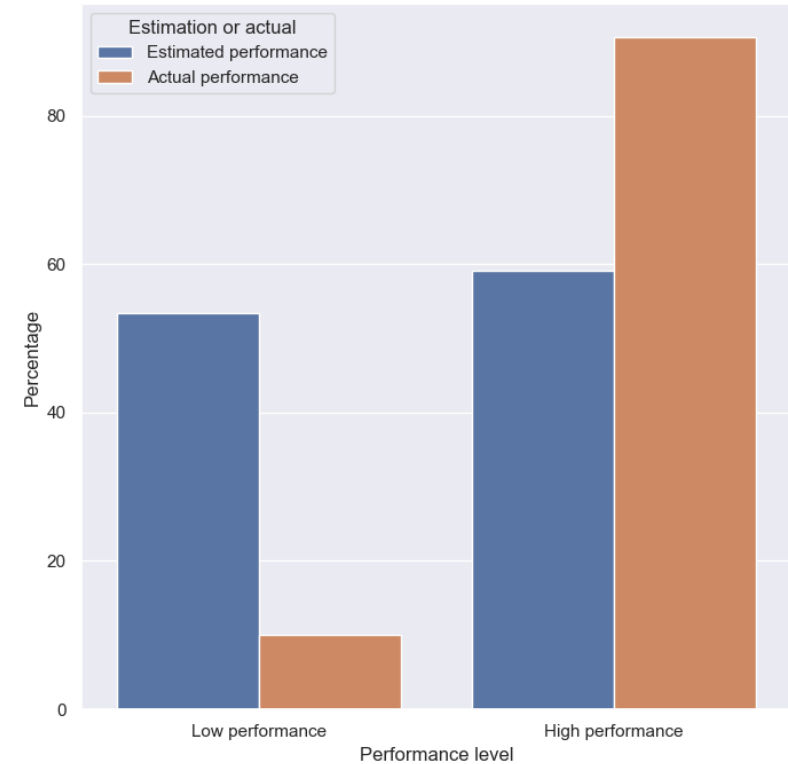
- Accuracy = $\frac{\text{Number of final decisions correct}}{\text{Total number of decisions}}$

- Self-estimation

- Degree of Miscalibration = $(\text{Number estimated correct answers} - \text{Actual number of correct answers})/6$

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	<i>U</i>	<i>p</i>	$M \pm SD(\text{Overestimation})$	$M \pm SD(\text{Other})$
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	<i>U</i>	<i>p</i>	$M \pm SD(\text{XAI})$	$M \pm SD(\text{no XAI})$	<i>U</i>	<i>p</i>	$M \pm SD(\text{XAI})$	$M \pm SD(\text{No XAI})$
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

Hypothesis 2

- Accurate self-assessment → degree of miscalibration = 0

- 76 accurate
- 173 inaccurate

- Appropriate reliance on AI

Wilcoxon-Mann Whitney test

Dependent Variables	<i>U</i>	<i>p</i>	$M \pm SD(\text{Accurate})$	$M \pm SD(\text{Inaccurate})$
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	<i>U</i>	<i>p</i>	$M \pm SD(\text{XAI})$	$M \pm SD(\text{no XAI})$	<i>U</i>	<i>p</i>	$M \pm SD(\text{XAI})$	$M \pm SD(\text{No XAI})$
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

Hypothesis 3

- Comparison before and after tutorial
 - 87 participants
- Miscalibration

Wilcoxon signed ranks test

Dependent Variable	<i>T</i>	<i>p</i>	$M \pm SD(\text{First batch})$	$M \pm SD(\text{Last batch})$
Mis-calibration	365.0	.0005	0.278 ± 0.152	0.190 ± 0.175

- No significant difference XAI

Wilcoxon-Mann Whitney test

Dependent Variable	<i>U</i>	<i>p</i>	$M \pm SD(\text{XAI})$	$M \pm SD(\text{No XAI})$
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	<i>U</i>	<i>p</i>	$M \pm SD(\text{Tutorial})$	$M \pm SD(\text{No tutorial})$
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

Hypothesis 4

- Comparison before and after tutorial
 - 87 participants
- Miscalibration

Wilcoxon signed ranks test

Dependent Variables	<i>T</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (First six)	<i>M</i> ± <i>SD</i> (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	<i>U</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (XAI)	<i>M</i> ± <i>SD</i> (no XAI)	<i>H</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (XAI)	<i>M</i> ± <i>SD</i> (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	<i>U</i>	<i>p</i>	<i>M</i> ± <i>SD</i> (Tutorial)	<i>M</i> ± <i>SD</i> (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

Summary of results

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

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

Thank you

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 (trAlt) (2022).
- Picture slide 4: https://commons.wikimedia.org/wiki/File:Dunning%E2%80%93Kruger_Effect_01.svg