A protocol to study the impact of XAI on the performance of AI-assisted decision making

Jules Leguy

Postdoc @ SyCoIA, IMT Mines Ales, Ales, France

Postdoc context

The ENFIELD Project

- European Network of Excellence
- Gathering of 30 institutions from 18 UE countries (including the IMT)

Postdoc hosting

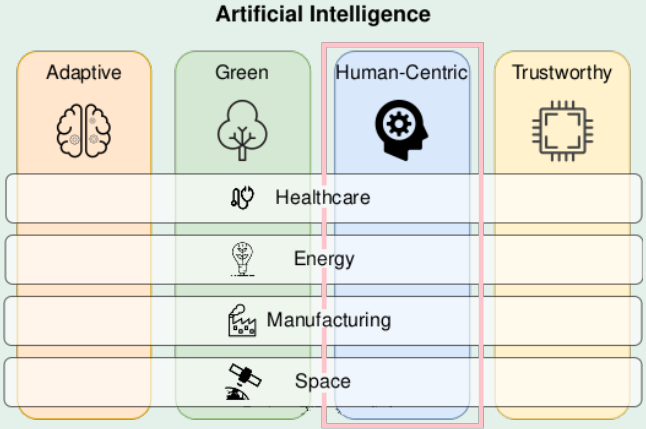
- IMT Mines Alès (Sébastien Harispe, PR)
- Links with IMT Business School (Nicolas Soulié, MCF)

Personal domains of interest

AI/Combinatorial optimization → XAI

ENFIELD

« European Lighthouse to Manifest Trustworthy and Green AI »



Introduction

Context

- Study of the impact of XAI on AI-assisted decision making
- In particular, what is the impact of XAI on the performance of human operators ?
- Work in progress

Explainable AI (XAI)

Domain of study and development of AI models whose decision-making processes can be explained, in a way that is understandable to human users.

AI-assisted decision making task

Task that must be accomplished by a human operator, with the help of an AI system which gives a suggestion on the decision to be made.

Motivation

- Still limited validation of XAI techniques in a realistic human-centric context.
- Studies of human-XAI interaction show contradictory and sometimes disappointing results regarding the benefits of XAI. [a,b,c]
- Some aspects are only scarcely studied in the literature : impact of time pressure, impact of task difficulty.
- Few studies propose a direct comparison of several explainability paradigms.

[a] Romy Müller. "How Explainable AI Affects Human Performance: A Systematic Review of the Behavioural Consequences of Saliency Maps". In: *International Journal of Human–Computer Interaction* 4 (2025)

[b] Rosina O Weber et al. "XAI is in trouble". In: *AI Magazine* 45 (2024)

[b] Raymond Fok and Daniel S. Weld. "In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making". In: *AI Magazine* (2024)

Introduction

Research questions

- What is the impact of explainability on the performance of humans solving a AI-assisted decision task ?
- What are the relative effects of various XAI paradigms in this context ?
- What is the impact of time pressure and task difficulty in this context ?
- What is the impact of XAI on reliance and overreliance to the model in this context ?

Research program

- Definition of an experimental framework to study human-XAI interaction (decision task).
- Design of the experimental protocol to answer the research questions.
- Testing the protocol with small cohorts. ← Current step
- Realization of the experiment (400 participants).

Experimental framework (Decision task)

Design of a synthetic decision task that:

- Does not require prior knowledge.
- Is non-trivial, justifying the help of a machine learning assistant.
- Allows generating explanations that "make sense" to human participants.

Task definition

Identification of the presence of patterns (yes/no decision) in randomly generated images of symbols.
Decisions in limited time.
Example of pattern question : is there at least one row containing triangles only ?

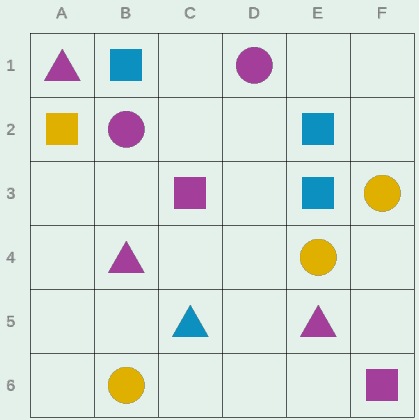


Figure: Example of image (6x6)

Experimental framework (Decision task)

Design of a synthetic decision task that:

- Does not require prior knowledge.
- Is non-trivial, justifying the help of a machine learning assistant.
- Allows generating explanations that "make sense" to human participants.

Task definition

Identification of the presence of patterns (yes/no decision) in randomly generated images of symbols.
Decisions in limited time.
Example of pattern question : is there at least one row containing triangles only ?

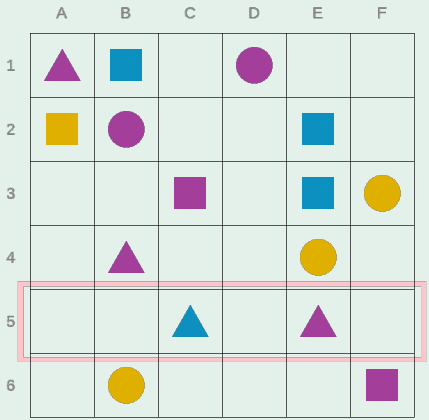


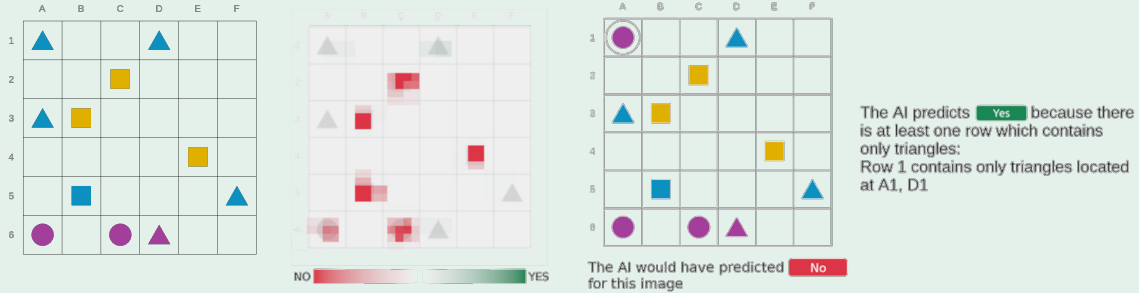
Figure: Example of image (6x6)

AI models

- Trained to perform the decision tasks.
- Constant 85% accuracy rate.
- ResNet-18 provides satisfying results for all tasks we considered. [a]

[a] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016

XAI paradigms considered in the study



Source | SHAP explanation | Counterfactual explanation | LLM explanation

Figure: Example of image and explanations for the question "is there at least one row containing triangles only?".

Study design

Independent variables

- AI condition (no AI, AI model).
- XAI condition (no XAI, XAI_1, XAI_2, ...).
- Task difficulty (low, high).
- Time pressure (mild, strong).

Mixed between-within design

- AI and XAI conditions are assigned to separate groups of participants (*between*).
- All participants are presented sequentially with low-high difficulty tasks and mild-strong time pressure conditions (*within*).

Between design for AI and XAI conditions

Name	**AI**	**XAI**
Human (control)	✗	✗
Human + AI	✓	✗
Human + AI + XAI_1	✓	XAI_1
...		
Human + AI + XAI_n	✓	XAI_n

Table: AI and XAI conditions for the disjoint groups of participants.

Study design

Timeline

- Introduction (instructions and training tasks).
- Main experiment (see Figure).
- Closing phase (surveys)

Within design for difficulty and time pressure

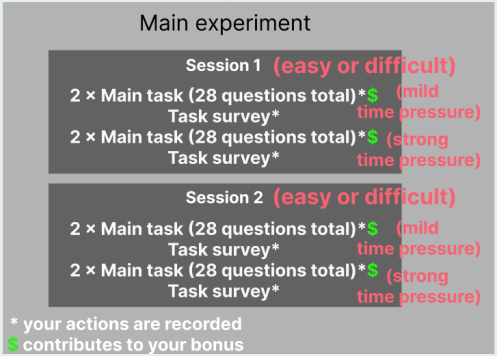


Figure: Timeline of the main experiment as presented to the participants (additional information in pink).

Main dependant variables

Variable	Type of measure	Scope
Accuracy (score)	Objective	Global
Reliance	Objective	Each condition
Overreliance	Objective	Each condition
Declared reliance	Declarative	Each condition
Declared trust	Declarative	Each condition
Declared XAI reliance	Declarative	Each condition
Cognitive load[a]	Standard survey	Each condition
Need for cognition[b]	Standard survey	Global

Table: Dependent variables, their type of measure, and scope (global or assessed for each condition).

[a] Sandra G. Hart and Lowell E. Staveland. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". In: *Advances in Psychology*. 1988

[b] Gabriel Lins de Holanda Coelho, Paul H P Hanel, and Lukas J Wolf. "The Very Efficient Assessment of Need for Cognition: Developing a Six-Item Version". In: *Assessment* (2020)

Study implementation

Recruitment of participants

- Through the platform Prolific.
- 400 participants expected (about 80 per cohort).
- Fixed remuneration + bonus remuneration depending on performance.

Try it yourself!

~~Connect to ...~~

If you are interested in trying the protocol, you can contact me by email.

Web interface implementation

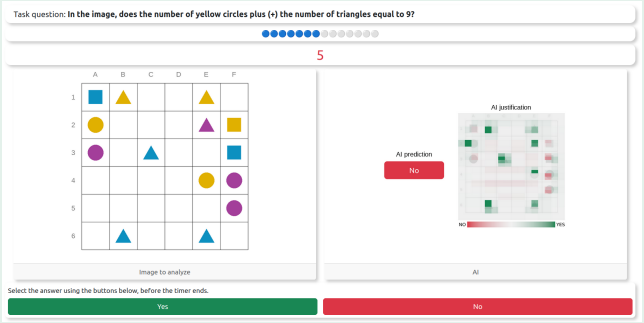


Figure: Implementation of the protocol using the WebXAII web platform[b].

[a] Jules Leguy et al. *WebXAII: an open-source web framework to study human-XAI interaction*. 2025. URL: https://arxiv.org/abs/2506.14777v1

Tests of the protocol

Experimental conditions

- About 20 participants (PhD students and Master's students from IMT Mines Alès).
- Times for mild/strong time pressure : 12s/7s to answer each question.
- 6×6 images.

Tasks considered

Easy:

- In the image, are there exactly 6 X symbols (X=color)?
- In the image, is there at least one row (1, ..., 6) containing only X (X=shape)?

Difficult:

- In the image, does the number of X plus (+) the number of Y equal to 8 (X=color+shape, Y=shape)?
- In the image, does the number of X multiplied by 2 (×2) equal to the number of Y (X=color+shape, Y=color)?

Metrics assessed and main outcomes

Qualitative

- Comprehension : no issues.
- Perceived difficulty : progression from easy to very difficult.
- Main frustration cause : not enough time to answer some questions.
- Actual use of AI : low.

Quantitative

- Score : $\geq 90\%$ (very high).
- Total time : between 25 and 40 minutes.

Adaptations for next phase

Levers to increase difficulty

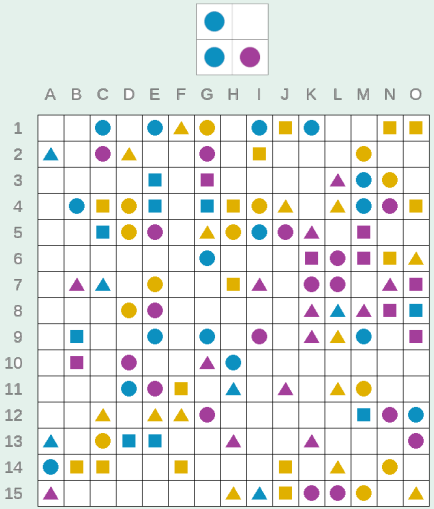
- Increased cognitive demand.
- More information to process.

Levers to promote the use of AI and XAI

- Higher task difficulty.
- Explanations must provide relevant information to solve the tasks[a].

New tasks

- Does the pattern appear in the image?
- Does the pattern appear exactly 3 times in the image?
- Does the pattern or a rotation of the pattern appear in the image?



[a] Raymond Fok and Daniel S. Weld. "In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making". In: *AI Magazine* (2024)

Conclusion

Main research question

- What is the impact of XAI on the performance of humans solving a AI-assisted decision task?

Main contributions

- Definition of a synthetic decision task.
- Design of an experimental protocol to study the research questions.

Next step

- Implementation of the protocol into a real-world experiment.

Come chat with me!

jules.leguy@mines-ales.fr

https://www.linkedin.com/in/jules-leguy/

https://jules-leguy.info/

Thank you for your attention.
Any question?

Main hypotheses

Performance

- AI improves the accuracy (all data or with high difficulty and high time pressure).
- XAI improves the accuracy (all data or with high difficulty and high time pressure).

Reliance and overreliance

- XAI increases reliance and overreliance to the model.
- High difficulty and high time pressure increase reliance and overreliance to the model.
- High cognitive load is correlated with higher reliance and overreliance to the model.
- XAI increases declared trust, declared reliance and estimated AI accuracy.
- High need for cognition leads to lower overreliance to the model.

Note : hypotheses related to XAI can be tested independently for every type of XAI technique.

Declared reliance and trust to AI and explanations

1. I relied on the AI to make my decisions.
   (1) Strongly Disagree (2) Disagree (3) Somewhat Disagree (4) Neutral (5) Somewhat Agree (6) Agree (7) Strongly Agree

2. I trusted the AI's decisions.
   (1) Strongly Disagree (2) Disagree (3) Somewhat Disagree (4) Neutral (5) Somewhat Agree (6) Agree (7) Strongly Agree

3. Estimate the accuracy of the AI model's predictions
   0-100 slider from 0% good decisions to 100% good decisions

4. The justifications gave me relevant insights about the AI's decisions.
   (1) Strongly Disagree (2) Disagree (3) Somewhat Disagree (4) Neutral (5) Somewhat Agree (6) Agree (7) Strongly Agree

5. The justifications had an impact on my decisions.
   (1) Strongly Disagree (2) Disagree (3) Somewhat Disagree (4) Neutral (5) Somewhat Agree (6) Agree (7) Strongly Agree

Trust and distrust in AI and XAI.

1. I earned trust in the AI thanks to the rightness of its predictions.
   (1) Strongly Disagree (2) Disagree (3) Somewhat Disagree (4) Neutral (5) Somewhat Agree (6) Agree (7) Strongly Agree

2. I earned trust in the AI thanks to the relevance of the justifications.
   (1) Strongly Disagree (2) Disagree (3) Somewhat Disagree (4) Neutral (5) Somewhat Agree (6) Agree (7) Strongly Agree

3. I lost trust in the AI because of its errors.
   (1) Strongly Disagree (2) Disagree (3) Somewhat Disagree (4) Neutral (5) Somewhat Agree (6) Agree (7) Strongly Agree

4. I lost trust in the AI because the justifications were not convincing or did not make sense.
   (1) Strongly Disagree (2) Disagree (3) Somewhat Disagree (4) Neutral (5) Somewhat Agree (6) Agree (7) Strongly Agree

Cognitive load NASA-TSX defined originally in [5]. Modified to a 7-points scale in [7] and also used in [4]

1. **Mental Demand** – How mentally demanding was the task?
   7-point scale from *Very Low* to *Very High*
2. **Physical Demand** – How physically demanding was the task?
   7-point scale from *Very Low* to *Very High*
3. **Temporal Demand** – How hurried or rushed was the pace of the task?
   7-point scale from *Very Low* to *Very High*
4. **Performance** – How successful were you in accomplishing what you were asked to do?
   7-point scale from *Perfect* to *Failure*
5. **Effort** – How hard did you have to work to accomplish your level of performance?
   7-point scale from *Very Low* to *Very High*
6. **Frustration** – How insecure, discouraged, irritated, stressed, and annoyed were you?
   7-point scale from *Very Low* to *Very High*

Need for cognition First defined in [2]. We are using the "very efficient" version which only uses 6 questions from [6]. Questions 4 and 5 are reverse coded.

1. I would prefer complex to simple problems.
   (1) Extremely Uncharacteristic (2) Somewhat Uncharacteristic (3) Uncertain (4) Somewhat Characteristic (5) Extremely Characteristic

2. I like to have the responsibility of handling a situation that requires a lot of thinking.
   (1) Extremely Uncharacteristic (2) Somewhat Uncharacteristic (3) Uncertain (4) Somewhat Characteristic (5) Extremely Characteristic

3. Thinking is not my idea of fun.
   (1) Extremely Uncharacteristic (2) Somewhat Uncharacteristic (3) Uncertain (4) Somewhat Characteristic (5) Extremely Characteristic

4. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.
   (1) Extremely Uncharacteristic (2) Somewhat Uncharacteristic (3) Uncertain (4) Somewhat Characteristic (5) Extremely Characteristic

5. I really enjoy a task that involves coming up with new solutions to problems.
   (1) Extremely Uncharacteristic (2) Somewhat Uncharacteristic (3) Uncertain (4) Somewhat Characteristic (5) Extremely Characteristic

6. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.
   (1) Extremely Uncharacteristic (2) Somewhat Uncharacteristic (3) Uncertain (4) Somewhat Characteristic (5) Extremely Characteristic

Sensibility to monetary incentive Question : Which sentence would best describe the strategy you used to maximize your score and monetary bonus ?

1. I largely relied on the AI's predictions, because I think they were correct all the time or almost all the time.
2. I largely relied on the AI's predictions, because I perceived them as imperfect but sufficiently accurate, and not worth the effort to surpass.
3. I sometimes or often relied on the AI's predictions, but due to limited trust in the AI, I made efforts to respond independently or to verify its predictions for many questions.
4. I barely relied or did not rely at all on the AI's predictions, because I had a very limited trust in the predictions.
5. I barely relied or did not rely at all on the AI's predictions, because I wanted to do the task by myself, independently of my assessment of the reliability of the AI.
6. I did not have a consistent strategy, or my strategy was not described in the propositions above.
7. Input text field to describe the strategy if last option was checked.

Use of explanations

1. I think the justifications were helpful to verify the answers to the questions.
   (1) Strongly Disagree (2) Disagree (3) Somewhat Disagree (4) Neutral (5) Somewhat Agree (6) Agree (7) Strongly Agree
2. I think the justifications were helpful to understand the AI's decision processes.
   (1) Strongly Disagree (2) Disagree (3) Somewhat Disagree (4) Neutral (5) Somewhat Agree (6) Agree (7) Strongly Agree
3. I think the justifications were helpful to detect the AI's errors.
   (1) Strongly Disagree (2) Disagree (3) Somewhat Disagree (4) Neutral (5) Somewhat Agree (6) Agree (7) Strongly Agree