A protocol to study the impact of XAI on 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)

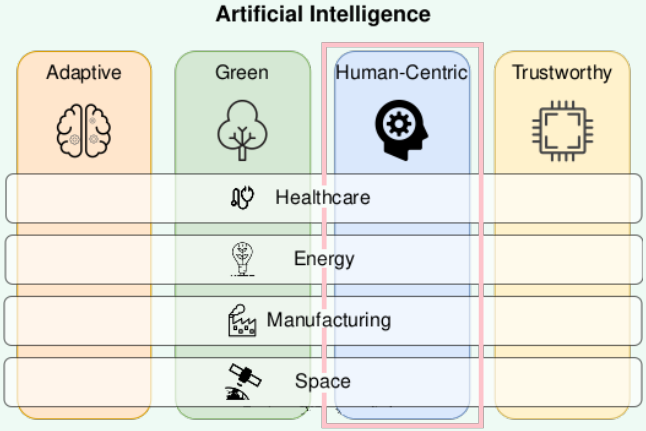
Postdoc hosting

- IMT Mines Alès (Sébastien Harispe, Andon Tchechmedjiev, Jacky Montmain)
- Links with IMT Business School (Nicolas Soulié)

Personal domains of interest

AI/Combinatorial optimization → XAI

« European Lighthouse to Manifest Trustworthy and Green AI »



Introduction

Context

- Study of the impact of XAI on AI-assisted decision making
- In particular, study of performance, trust and reliance in a time-pressure context.

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.

Introduction

Motivation

- Studies of human-XAI interaction show contradictory and sometimes disappointing results regarding the benefits of XAI. [a,b,c]
- Hot topics in the literature : overreliance and appropriate reliance [d], impact of time pressure[e].
- 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* (Feb. 2025)

[b] Julian Senoner et al. "Explainable AI improves task performance in human–AI collaboration". In: *Scientific Reports* (Dec. 2024)

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

[d] Helena Vasconcelos et al. "Explanations Can Reduce Overreliance on AI Systems During Decision-Making". In: *Proceedings of the ACM on Human-Computer Interaction* (2022)

[e] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. "How Time Pressure in Different Phases of Decision-Making Influences Human-AI Collaboration". In: *Proc. ACM Hum.-Comput. Interact.* (Oct. 2023)

Introduction

Research questions

In the context of a (X)AI-assisted decision task...

- What are the relative effects of various explainability paradigms ?
- What is the impact of time pressure and task difficulty ?
- What is the impact of XAI on reliance and overreliance to the model ?

Program

1. Definition of the decision task.
2. Design of the protocol.
3. Calibration of the protocol (thank you to the PhD students!).
4. Pilot study with 40 participants recruited online. ← Current step
5. Large scale realization of the experiment (500+ 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 ?

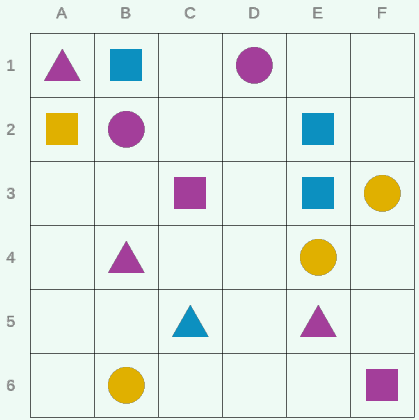


Figure: Example of image (6x6)

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 ?

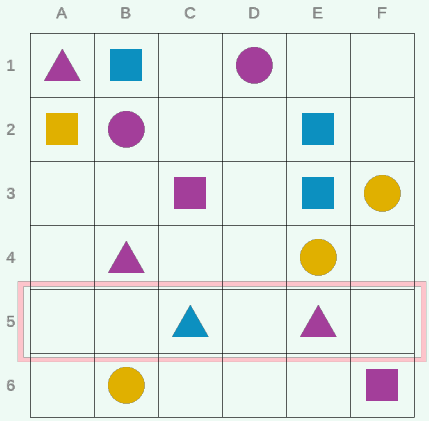


Figure: Example of image (6x6)

Machine learning models

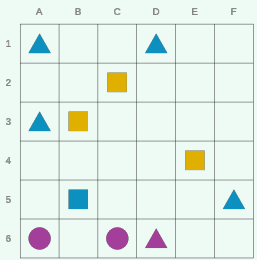
AI models

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

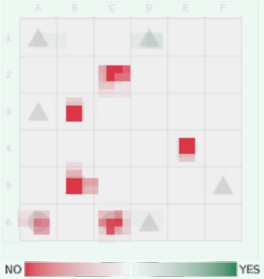
[a] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016

Explainability

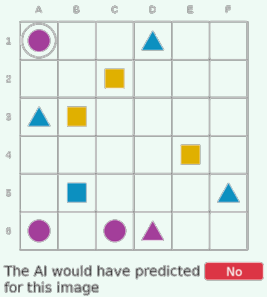
XAI paradigms considered



Source



SHAP explanation



Counterfactual explanation

The AI predicts Yes because there is at least one row which contains only triangles:
Row 1 contains only triangles located at A1, D1

LLM explanation

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

Explainability

XAI paradigms that might be considered

- Multimodal LLM text + image.
- GradCAM.[b]
- Weak baselines (highlighting random symbols).

[a] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *International Journal of Computer Vision* (Feb. 2020)

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.

Main hypotheses

Standard hypotheses

- XAI increases trust and reliance.
- Time pressure increases reliance.

Study-specific hypotheses

- Explanations increase the cognitive load when the task is easy
- LLM leads to highest trust, followed by SHAP and then counterfactuals.
- LLM leads to lowest overreliance, followed by SHAP and then counterfactuals.
- Only XAI techniques which enable verifiability decrease overreliance.
- XAI increases overreliance in case of false negatives.

Study implementation

Recruitment of participants

- Through Prolific.
- 500+ participants expected.
- Fixed remuneration + bonus remuneration depending on performance.

Web interface implementation

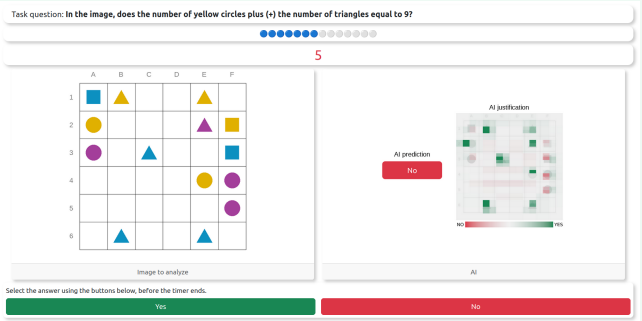


Figure: Implementation of the protocol using WebXAII.[a]

[a] Jules Leguy et al. *WebXAII: an open-source web framework to study human-XAI interaction*. en. May 2025. URL: https://arxiv.org/abs/2506.14777v1 (visited on 06/20/2025)

Calibration phase

Experimental conditions

- A total of about 50 participations (PhD and Master's students from IMT Mines Alès).
- 4 sessions from October-December.

Main objectives

- Calibrate the difficulty of the tasks.
- Calibrate the duration of the experiment.

First sessions' outcomes

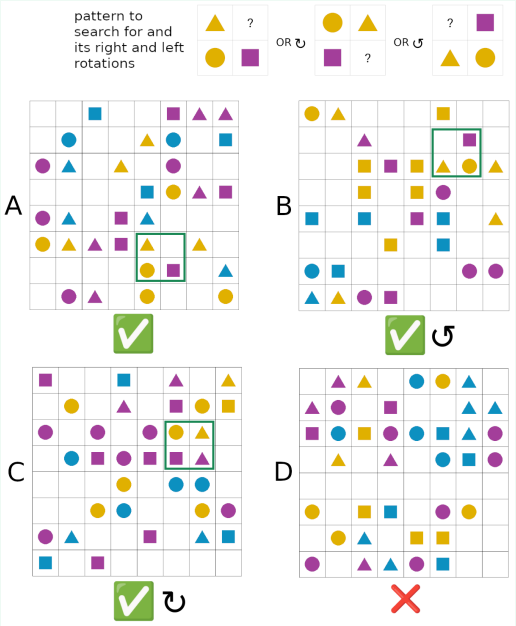
Qualitative

- 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.

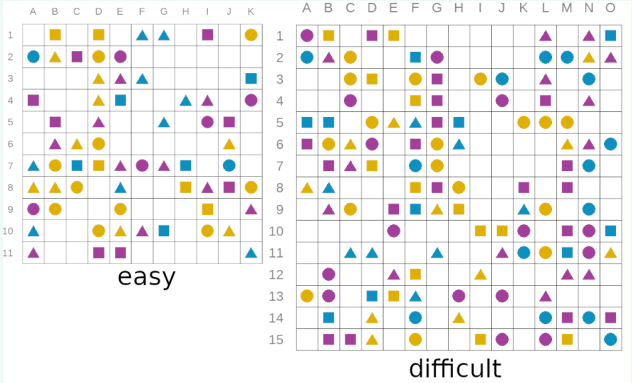
Quantitative

- Score : $\geq 90\%$ (too high).
- Total time : between 25 and 45 minutes (too high).

Final protocol



Difficulty

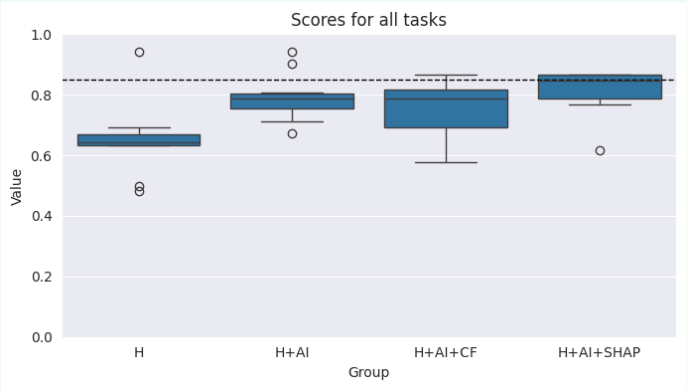


Time pressure

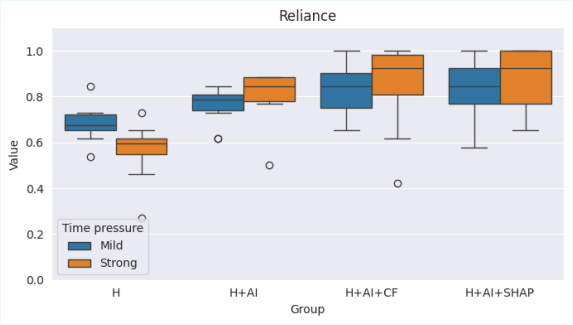
Mild: 20s; strong: 10s

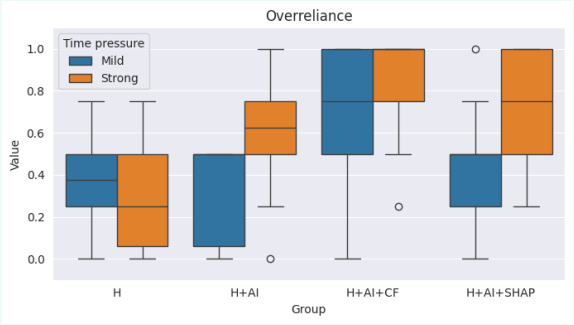
Pilot study

- 40 Prolific participants took the experiment this week.
- Reward : £2 fixed + £0-£2.6 bonus for 20 minutes.
- Groups assigned randomly
  - H : Human alone
  - H+AI : Human + AI predictions
  - H+AI+SHAP : Shap explanations
  - H+AI+CF : Counterfactual explanations



Pilot study


Reliance
Value
1.0
0.8
0.6
0.4
0.2
0.0
Time pressure
Mild
Strong
H
H+AI
H+AI+CF
H+AI+SHAP
Group


Overreliance
Value
1.0
0.8
0.6
0.4
0.2
0.0
Time pressure
Mild
Strong
H
H+AI
H+AI+CF
H+AI+SHAP
Group

Thank you for your attention.
Any question?

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*

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