AI-assistance to decision-makers: evaluating usability, induced cognitive load, and trust's impact

Alexis D. Souchet
IRT SystemX
contact@alexissouchet.com

Kahina Amokrane-Ferka
Irt Systemx,
kahina.amokrane-ferka@irt-systemx.fr
Jean-Marie Burkhardt
LaPEA, Université Gustave Eiffel et Université Paris Cité
jean-marie.burkhardt@univ-eiffel.fr

ABSTRACT

We designed a randomized-controlled study with 80 participants to investigate the effects of an AI assistant on the activity and processes implemented by users in their decision-making tasks. For that purpose, we will examine several aspects of decision-making in problem-solving situations in five experimental conditions resulting from the combination of the following factors: AI assistance (with vs. without), information related to the reliability of the assistant's proposals (yes vs no) and cognitive load induced variation through a dual task (with vs without). We plan to collect profile data, heart-rate variables, task efficiency and perceived usability, cognitive load, and trust. We are currently finalizing the prototype to conduct pre-tests.

CCS CONCEPTS

• : **Artificial intelligence**; • **Empirical studies in HCI**; • **User studies**; • **Decision support systems**;

KEYWORDS

Decision-making, Artificial intelligence, Teaming, Cognitive load

ACM Reference Format:
Alexis D. Souchet, Kahina Amokrane-Ferka, Irt Systemx, kahina.amokrane-ferka@irt-systemx.fr, and Jean-Marie Burkhardt. 2024. AI-assistance to decision-makers: evaluating usability, induced cognitive load, and trust's impact. In *European Conference on Cognitive Ergonomics (ECCE 2024), October 08–11, 2024, Paris, France*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3673805.3673845

1 INTRODUCTION

1.1 Context

The development and democratization of decision-making assistants based on Artificial Intelligence (AI) give rise to many issues regarding their effect(s) on the activity of human operators, i.e. explainability, interpretability, reproducibility, and human-centered AI as well as standardized measurements and hypothetical-deductive methodology (Rahimi et al., 2022; Lai et al., 2023). Providing users with AI-based assistance is based on strong assertions such as countering human reasoning bias, considering more data, reducing user cognitive load, and enhancing performance. Indeed, providing AI assistance improves decision-making (Lai et al., 2023). However, studies of human-IA decision-making little assess users' cognitive load while performing the task, hence not considering how much cognitive resources are mobilized (Steyvers & Kumar, 2023). This paper presented the work in progress on a study aiming at better understanding those interactions between systems' characteristics and human factors while making decisions with AIs.

1.2 Key AI-based assistant characteristics and human factors in decision making

The major characteristics of AI-based advisory systems are reliability (or accuracy), explainability, and transparency (Chancey et al., 2017; Gilpin et al., 2018). Depending on the algorithm, the scope, and the nature of the data processed by the AI, as well as the quality of its design, there may be a margin of uncertainty as to the relevance and suitability of the solution proposed to the user. In addition, the reasoning by the AI may not be visible or intelligible to their users. These characteristics strongly influence the trust that the user attributes to the AI and its proposal e.g. (Hoff & Bashir, 2015; Chiou & Lee, 2023). Overreliance - which describes humans trusting AI without questioning its suggestions enough, which can lead to unadapted decisions - can arise (Buçinca et al., 2021). For some authors, one way of calibrating the user's trust could be that the User Interface displays the degree of certainty/uncertainty of the assistant concerning what it proposes as a solution to help the user (Fügener et al., 2021; Hemmer et al., 2023; Schemmer et al., 2022; W. Xu et al., 2023). However, there is currently no consensus on the impact of trust and explainability on decision-making activity in supervisory tasks (Schemmer et al., 2022; W. Xu et al., 2023). Furthermore, the cognitive load variable is not addressed in those works about trust.

1.3 Needed contributions identified in previous works

The review by Lai et al. (2023) highlights several gaps between AI user studies and real-world decision support applications. Most previous work focuses on the effectiveness of AI in generating decision proposals. But, they rarely assess the factors affecting

ECCE 2024, October 08–11, 2024, Paris, France

ACM ISBN 979-8-4007-1824-3/24/10
https://doi.org/10.1145/3673805.3673845

AI-assisted human decision-making. There are also methodological limits related to the quality of procedures (e.g. missing information related to the context and participants' characteristics relevant to decision-making; Appelbaum et al., 2018; Orkin et al., 2021) and measurement tools (e.g. ad-hoc questionnaires; Lai et al. 2023) used in these studies. Finally, the effects of reliability and trust in decision-making still require research in most areas where AI is expected to be adopted (Rahimi et al., 2022). Therefore, contributions are required to better evaluate and understand the interactions between systems' characteristics and human factors such as cognitive load. We propose to address those issues through a randomized-controlled study. The study takes place in the context of the Cockpit and Bidirectional Assistant (CAB) that aims to develop, in partnership with Orange, RTE, SNCF, Flying Whales, and Dassault aviation, a collaborative AI (Berretta et al., 2023) to assist operators in their decision-making tasks.

1.4 Objectives of the study

Our study aims to investigate the effects of an AI assistant on the activity and processes implemented by users in their decision-making tasks. For that purpose, we examine several aspects of decision-making in problem-solving situations in 5 conditions resulting from the combination of the following factors: AI assistance (with vs. without), information related to the reliability of the assistant's proposals (yes vs no) and cognitive load induced variation through a dual task (with vs without). The measured variables concern the level and evolution of cognitive load during the trials, performance in decision-making by the user regarding expert judgment, and the evolution of confidence in the assistant.

1.5 Hypotheses

We will test five hypotheses: H1: Cognitive load is reduced when humans are assisted by AI in decision-making. H2: High confidence in AI improves performance. H3: Low cognitive load improves trust in AI. H4: Displaying AI reliability increases performance in AI-assisted decision-making. H5: Displaying AI reliability reduces cognitive load.

2 MATERIAL AND METHOD

2.1 Participant recruitment and sample characteristics

80 participants with half men, and half women (defined with G*Power (Faul et al., 2007) for 5 experimental conditions, average effect size (of 0.5) expected based on previous work, to perform ANOVA or MANOVA) between 18-60 years, with normal or corrected vision (glasses, contacts), and no medical treatment or pathology likely to influence cognition and cardiac function. Participants aren't experts in railway networks but from the general population. Participants are recruited by Eurosyn in their user base. They receive 30 € (in the form of an Amazon voucher). The protocol is currently assessed for ethical approval by the CER U Paris Cité.

2.2 Experimental conditions

We created five experimental conditions:

- Control: primary task *without* simulated AI
- AI-support: primary task *with* simulated AI
- Dual-Task: primary task *without* simulated AI + secondary task
- Dual-Task-AIsupport: primary task *with* simulated AI + secondary task
- Dual-Task-AIsupport-Accuracy: primary task *with* simulated AI diplaying % accuracy + secondary task

2.3 Material

Participants will be seated in front of a PC monitor in an ergonomic chair. They will use a mouse to interact with the prototype. The interface has been developed by experts from rail management, plane pilots, and ICT system management dedicated to telecommunication and power grid management. The human-centered procedure has been followed. Although the initial system has been developed for operators, we created a sub-design for a better understanding by the general population based only on the rail management use case.

2.4 Primary Task

Each trial represents a train incident on the Bordeaux-Paris line, for which they must decide which of 4 actions to apply, taking into account the cost and number of passengers affected. The 4 actions are: hold the train at the station, cancel the train, delay the train, and change the train's route. The following information is displayed on the user interface: the number of users impacted, the cost of each action, and a map of the train traffic. Subjects are prompted to consult the information displayed and decide on the optimal action. They have 1 minute to make their decision. To do so, they have to multiply the number of users impacted and the cost of each action, memorize it for each possible action, and then decide based on the result which action is the best. Conditions with simulated AI assistance are similar, but suggest actions by displaying and cost results that can be wrong and displaying or not displaying AI confidence. If the time for making decisions on the train number has elapsed, the next case automatically replaces the previous one. The subject receives no feedback as to whether its action decision is "right or wrong." Subjects are not aware that the AI is simulated. They will be informed after the study and the questionnaires. 1 displays the interface for condition E, the prototype for the other conditions consists of having or not part of the UI.

2.5 Secondary task for the dual-task conditions

In the dual-task conditions, subjects are asked to answer a simulated phone call to answer an audio call (real human voices) in which they are given a train number. They must memorize this train number and then communicate it in turn via the interface within 15 seconds. This secondary task is repeated 10 times in each experimental condition.

2.6 Procedure

We use a between-subjects design. Each participant is randomly assigned to one out of the five experimental conditions. 1 displays the procedure for a trial for one participant.

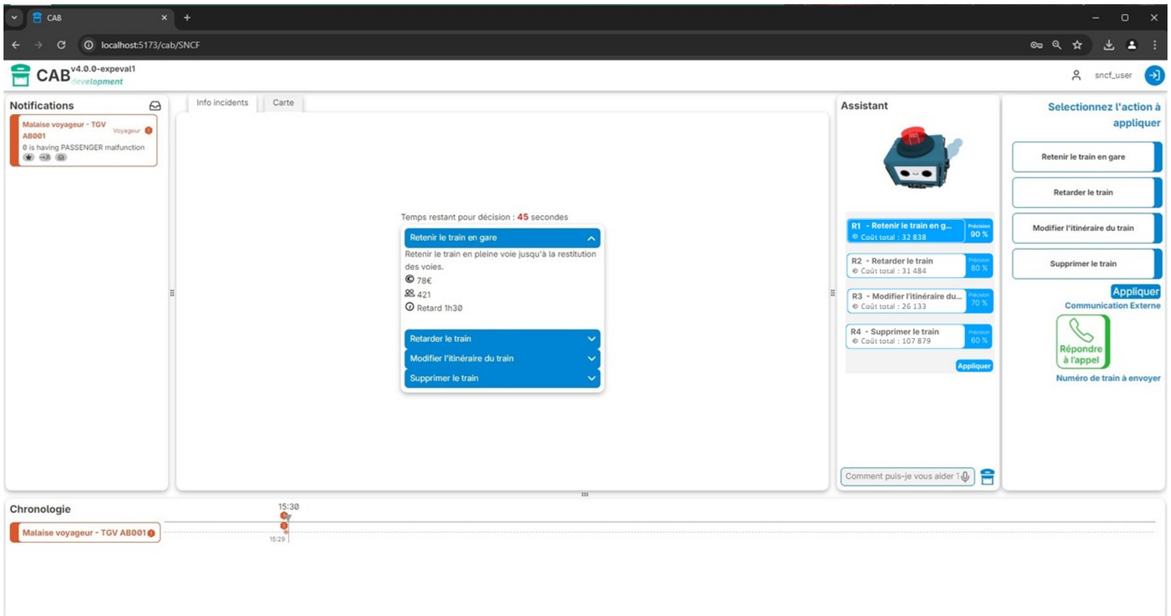

Figure 1: Condition E prototype, participants can check on the traffic map, see AI recommendations with accuracy, and answer a phone call to then communicate a number

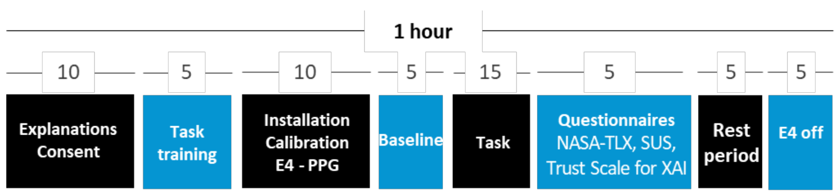

Figure 2: Procedure for each participant

2.7 Collected data

Profil questionnaires: the socio-demographic questionnaire includes questions about Age, Gender, Highest degree, and Socio-professional category (INSEE grid). The AI Experience Questionnaire (Wang & Peng, 2023) is used to assess participants' prior experience with AI through their daily life. The General Attitudes Towards Artificial Intelligence Scale (Schepman & Rodway, 2020) measures the general positive or negative acceptance of AI.

Physiological data: Photoplethysmography (Hughes et al., 2019) - E4 wristband Empatica - is used to assess physiological indicators of cognitive load (Heart rate variability (HRV), Heart period (HP), Heart rate (HR).

Behavioral data: Task performance, Time on task.

Questionnaire after the task: The System Usability Scale (Gronier & Baudet, 2021) assesses the perceived usability. The NASA-Task Load index (Hart & Staveland, 1988; Cegarra & Morgado, 2009) assesses the perceived cognitive load induced by the task. The Trust Scale for Explainable AI (Hoffman et al., 2023; Perrig et al., 2023) assesses the perceived trust in the system.

3 EXPECTED RESULTS AND DISCUSSION

The data are to be collected in July. We are currently finalizing the prototype and are about to perform pre-tests. We predict statistically significant differences between experimental conditions on each dependent variable measured with average effect sizes. The hypotheses tested will enable us to support or not: the decrease in cognitive load during AI-assisted decision-making; the link between Confidence in AI and improved performance; the link between low cognitive load and improved confidence in AI; the link between displaying AI reliability and increased performance in AI-assisted decision making; the link between displaying AI reliability and reduced cognitive load.

ACKNOWLEDGMENTS

This work has been supported by the French government under the "France 2030" program, as part of the SystemX Technological Research Institute. The project was funded by Orange, RTE, SNCF, Flying Whales, and Dassault aviation.

REFERENCES

[1] Appelbaum, M., Cooper, H., Kline, R., Mayo-Wilson, E., Nezu, A., & Rao, S. (2018). Journal Article Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report. *American Psychologist*, *73*, 3-25. https://doi.org/10.1037/amp0000389

[2] Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., & Kluge, A. (2023). Defining human-AI teaming the human-centered way: A scoping review and network analysis. *Frontiers in Artificial Intelligence*, *6*. https://www.frontiersin.org/articles/10.3389/frai.2023.1250725

[3] Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 188:1-188:21. https://doi.org/10.1145/3449287

[4] Cegarra, J., & Morgado, N. (2009). Étude des propriétés de la version francophone du NASA-TLX. *EPIQUE 2009: 5ème Colloque de Psychologie Ergonomique*, 233-239.

[5] Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. H. (2017). Trust and the Compliance–Reliance Paradigm: The Effects of Risk, Error Bias, and Reliability on Trust and Dependence. *Human Factors*, *59*(3), 333-345. https://doi.org/10.1177/0018720816682648

[6] Chiou, E. K., & Lee, J. D. (2023). Trusting Automation: Designing for Responsivity and Resilience. *Human Factors*, *65*(1), 137-165. https://doi.org/10.1177/00187208211009995

[7] Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

[8] Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research*. https://doi.org/10.1287/isre.2021.1079

[9] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80-89. https://doi.org/10.1109/DSAA.2018.00018

[10] Gronier, G., & Baudet, A. (2021). Psychometric Evaluation of the F-SUS: Creation and Validation of the French Version of the System Usability Scale. *International Journal of Human–Computer Interaction*, *37*(16), 1571-1582. https://doi.org/10.1080/10447318.2021.1898828

[11] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Éds.), *Advances in Psychology* (Vol. 52, p. 139-183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

[12] Hemmer, P., Westphal, M., Schemmer, M., Vetter, S., Vössing, M., & Satzger, G. (2023). Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 453-463. https://doi.org/10.1145/3581641.3584052

[13] Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, *57*(3), 407-434. https://doi.org/10.1177/0018720814547570

[14] Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, *5*. https://www.frontiersin.org/articles/10.3389/fcomp.2023.1096257

[15] Hughes, A. M., Hancock, G. M., Marlow, S. L., Stowers, K., & Salas, E. (2019). Cardiac Measures of Cognitive Workload: A Meta-Analysis. *Human Factors*, *61*(3), 393-414. https://doi.org/10.1177/0018720819830553

[16] Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1369-1385. https://doi.org/10.1145/3593013.3594087

[17] Orkin, A. M., Nicoll, G., Persaud, N., & Pinto, A. D. (2021). Reporting of Sociodemographic Variables in Randomized Clinical Trials, 2014-2020. *JAMA Network Open*, *4*(6), e2110700. https://doi.org/10.1001/jamanetworkopen.2021.10700

[18] Perrig, S. A. C., Scharowski, N., & Brühlmann, F. (2023). Trust Issues with Trust Scales: Examining the Psychometric Quality of Trust Measures in the Context of AI. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-7. https://doi.org/10.1145/3544549.3585808

[19] Rahimi, S. A., Cwintal, M., Huang, Y., Ghadiri, P., Grad, R., Poenaru, D., Gore, G., Zomahoun, H. T. V., Légaré, F., & Pluye, P. (2022). Application of Artificial Intelligence in Shared Decision Making: Scoping Review. *JMIR Medical Informatics*, *10*(8), e36199. https://doi.org/10.2196/36199

[20] Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 617-626. https://doi.org/10.1145/3514094.3534128

[21] Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards Artificial Intelligence Scale. *Computers in Human Behavior Reports*, *1*, 100014. https://doi.org/10.1016/j.chbr.2020.100014

[22] Steyvers, M., & Kumar, A. (2023). Three Challenges for AI-Assisted Decision-Making. *Perspectives on Psychological Science*, 17456916231181102. https://doi.org/10.1177/17456916231181102

[23] Wang, C., & Peng, K. (2023). AI Experience Predicts Identification with Humankind. *Behavioral Sciences*, *13*(2), Article 2. https://doi.org/10.3390/bs13020089

[24] Xu, C., Lien, K.-C., & Höllerer, T. (2023). Comparing Zealous and Restrained AI Recommendations in a Real-World Human-AI Collaboration Task. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-15. https://doi.org/10.1145/3544548.3581282

[25] Xu, W., Dainoff, M. J., Ge, L., & Gao, Z. (2023). Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI. *International Journal of Human–Computer Interaction*, *39*(3), 494-518. https://doi.org/10.1080/10447318.2022.2041900