Proceedings of the Human Factors and Ergonomics Society Annual Meeting
1–7


DOI: 10.1177/10711813251358254
journals.sagepub.com/home/pro
Sage

Inverse Counterfactual for AI-Assisted Decision Support: Enhancing Knowledge Elicitation for Capturing Aircraft Pilot Decisions

Jonay Ramon Alaman[1], Daniel Lafond[2], Alexandre Marois[1,3], and Sébastien Tremblay[1]

Abstract

Integrating AI into decision-support systems (DSS) for safety-critical domains like aviation requires aligning system behavior with pilot mental models to provide relevant information. Using the Cognitive Shadow—a DSS that models operator decisions and notifies discrepancies—we evaluated a novel knowledge-elicitation technique: the inverse counterfactual. After selecting their preferred option, users modified a single factor to make their second-best option preferable, creating paired cases across their decision boundary. In a simulated adverse-weather avoidance task, 44 participants completed 130 baseline trials and generated counterfactuals for 20 additional cases. Contrary to expectations, the current implementation of the technique did not enhance human-AI model similarity, as measured by the degree of agreement in a 20-case test phase. However, when counterfactuals involved minimal edits—remaining near the decision boundary—predictive accuracy improved and DSS recommendations were more often accepted. Larger edits degraded performance. These findings demonstrate the feasibility of counterfactual elicitation for improving model alignment with user mental models.

Keywords

policy capturing, cognitive engineering, decision support, machine learning, knowledge elicitation, counterfactual, mental model, decision making

Introduction

The integration of artificial intelligence (AI) into decision-support systems (DSS) for safety-critical tasks, such as aircraft piloting, requires AI models that not only assist decision making, but also align with expert mental models (Steyvers & Kumar, 2023). Effective human-AI teaming should benefit from shared models, as do human teams, by facilitating anticipation of actions and effective information exchange (O'Neil et al., 2023). However, current AI approaches often prioritize performance optimization while relying on developers' intuition of what "constitutes a 'good' explanation" (Miller, 2019, p. 1), rather than adapting to actual decision-making processes. Addressing this gap requires approaches informed by human factors research that move beyond predictive outputs to better capture the criteria behind expert decisions. Enhancing human-AI model similarity can foster mutual understanding, support situation awareness, and improve decision reliability.

One such system is the Cognitive Shadow, a prototype cognitive assistant that continuously models the decision-making pattern of the user and notifies them of any discrepancy between the model and the current decision of the user (i.e., when detecting potential errors due to distraction, fatigue, etc.; Lafond et al., 2020). The system has been applied in several contexts, including maritime and aviation decision-making domains (see Labonté et al., 2021; MacLean et al., 2024; Marois et al., 2023). For doing so, it uses policy capturing, a judgment analysis technique that models how individuals make decisions by identifying consistent patterns in their responses across a knowledge capture phase presenting sets of cases with varying attributes, filtering out random error to infer a representative decision-making rule (Marois et al., 2023; Nokes & Hodgkinson, 2018).

[1]École de Psychologie, Université Laval, Québec, QC, Canada
[2]Thales, Québec, QC, Canada
[3]School of Psychology and Humanities, University of Central Lancashire, Preston, UK

Corresponding Author:
Jonay Ramon Alaman, École de Psychologie, Université Laval, 2325 Rue des Bibliothèques, Québec, QC G1V 0A6, Canada.
Email: jonay.ramon@gmail.com

Based on the research of Nuñez et al. (2019), the selected use-case is in-flight adverse weather avoidance. This situation allows sufficient variability to generate the required number of cases and leaves a gray zone of decisions open to the discretion of the pilots, resulting in between-subjects variability that the DSS will need to capture. Weather avoidance typically involves evaluating current and forecasted weather data, anticipating hazardous meteorological conditions, and adjusting the flight path to maintain safety and operational efficiency. These decisions require pilots to weigh competing constraints such as turbulence risk, fuel availability, and schedule demands, often under time pressure and with incomplete information. The task was grounded in key elements identified in prior analyses of enroute adverse weather avoidance, such as weather severity and fuel consumption (Endsley & Jones, 2012; Ramon Alaman et al., 2024). The task was designed to capture operational decisions commonly encountered by pilots—for example, selecting the extent of a lateral or vertical deviation—while aligning with core objectives of flight operations of maintaining safety and aircraft integrity, and managing fuel constraints.

The Inverse Counterfactual

Counterfactual reasoning, which explores how outcomes would change under different conditions, has long been central to human cognition and causal inference (Miller, 2019). In explainable AI, counterfactual generation has been proposed as a way to produce interpretable, and potentially causal, explanations by identifying the minimal modifications needed to alter the prediction of a model (Chou et al., 2022). Inspired by such a use of counterfactuals, this study proposes a new knowledge-elicitation method called the "inverse counterfactual," which is human-driven rather than AI-driven. This technique was developed with the objective of training AI models more efficiently. Here, for each case presented in the knowledge capture session, the user generates a new case on the opposite side of their decision boundary, providing the DSS with additional information to more precisely model the decision boundary of the user. This higher precision on the decision boundary may, however, require a higher number of training cases to effectively sample the whole problem space for achieving similar model accuracy. Additionally, similarly to the benefit of process tracing (another human-centered data modeling approach relying on explicit feature information acquisition; Labonté et al., 2021), we hypothesize that self-reflection about those decision boundaries could enhance mental model quality, leading to more consistent decision making.

The approach was initially introduced by Marois et al. (2023) for a binary classification task to support frugal learning (efficient cognitive modeling using a relatively small number of examples), though its impact was not assessed. MacLean et al. (2024) later evaluated the impact of self-reflection with a sonar range prediction credibility assessment task, where participants could modify situational features to generate counterfactual cases. Their findings revealed that generating counterfactuals posed significant challenges for participants, due to the unrestricted nature of the counterfactual modifications that could be performed. Unlike natural counterfactual reasoning, which is typically limited to a small set of plausible alternatives under human control, the experimental task required participants to consider numerous alternatives with the possibility to modify factors beyond their control, increasing difficulty (MacLean et al., 2024). This highlights a broader challenge in aligning AI systems with human cognition—specifically, how to elicit and represent human reasoning patterns in a way that supports mutual understanding (Wang & Chen, 2024).

Study Objectives

The primary objective of the study was to assess whether inverse counterfactuals can enhance the ability of the DSS to generate an individualized model which aligns with the operator's (individual) decision-making model. Additionally, the study seeks to identify the characteristics of an effective inverse counterfactual case, providing insights into how effective human-DSS information sharing might enhance the similarity between human and AI decision-making models.

Method

Participants

Forty-four students or employees at Université Laval (24 women, 20 men, $M_{age} = 26.62 \pm 8.81$ years) took part in the study, which consisted of a 2-hr experimental session. They received CAD $20 as compensation.

Apparatus and Material

For each weather avoidance case presented, participants were given four avoidance options: stay on track or choose one of three diversion options (two lateral and one vertical). They had to select the option they considered best, based on how important each factor was to them personally. The custom-made interface provided all necessary information, enabled the generation of inverse counterfactuals, and managed communication with the DSS. As shown in Figure 1, participants were presented with an interface that included two standard flight deck displays: the Primary Flight Display (PFD, left) and the Navigation Display (ND, right). While the PFD was present for realism, it did not display task-relevant information. The ND, in contrast, presented a weather radar image along with two rhomboids indicating lateral avoidance options and their position relative to the aircraft (yellow cross). Tabular data provided additional information, including the movement (i.e., speed and direction) of the

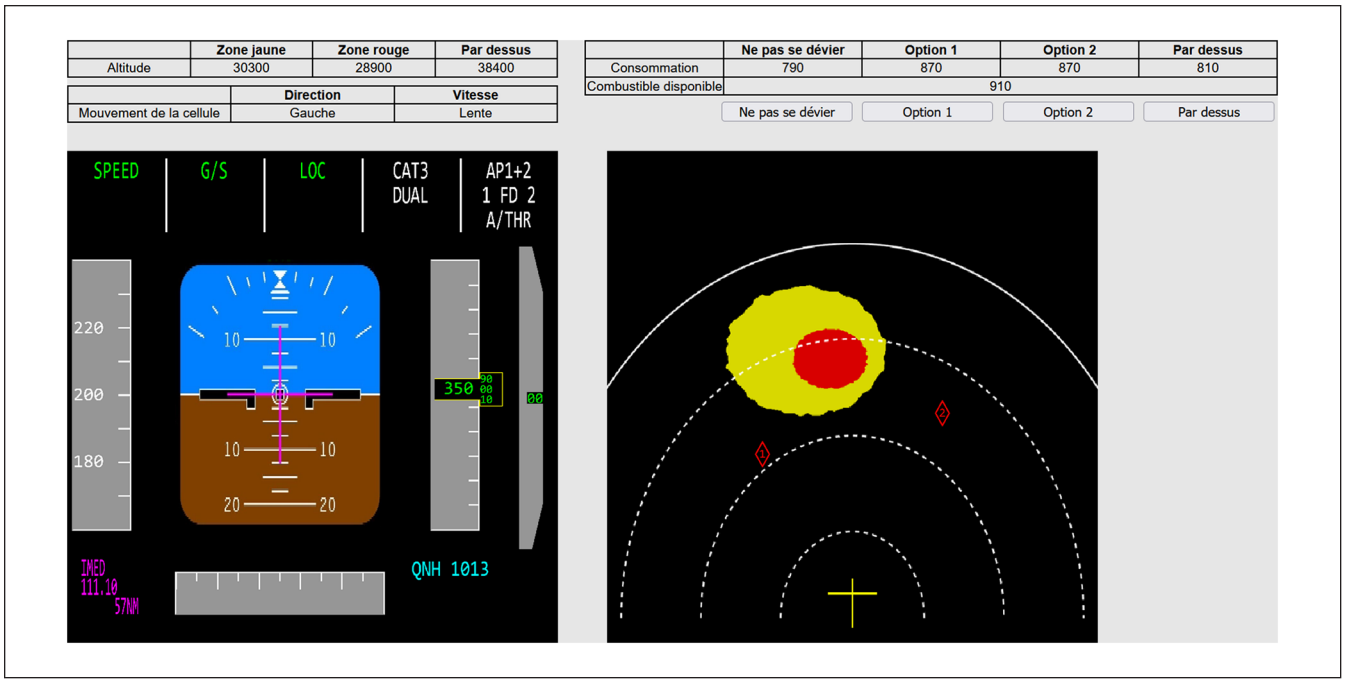

Figure 1. Task interface used by participants, displaying all task-relevant information. PFD on the left under the tables. ND on the right under the tables. French terms in the image are translated as follows: "Zone jaune" = "Yellow zone," "Zone rouge" = "Red zone," "Par dessus" = "Overfly," "Mouvement de la cellule" = "[Adverse weather] cell movement," "Vitesse" = "Speed," "Consommation" = "Consumption," "Combustible disponible" = "Available fuel."

adverse weather region, the fuel consumption associated with each option, available fuel, the altitude of the top of the adverse weather region, and the service ceiling (i.e., maximum operational flight altitude), which constrained the vertical avoidance option. Features were selected for the task to be accessible to novices, simplified enough to be learnable, yet realistic enough to reflect the kinds of trade-offs that pilots typically face in operational settings.

Procedure

After signing the informed consent form, participants received task instructions. The instructions stated that their goal was to minimize fuel consumption while ensuring passenger safety and aircraft integrity by maximizing lateral or vertical distance from the adverse-weather region and strictly avoiding at all costs crossing the zone of highest intensity displayed in red. Instructions clearly stated that the importance assigned to each factor was determined solely by the participant. This importance was reflected in the decisions they made. In practice, this meant that participants were free to prioritize or compromise on any criterion as they saw fit, provided that they respected the fuel limitations and the requirement to avoid the highest intensity zone.

Participants underwent a familiarization phase during which they received feedback from the experimenter. As was the case with the instructions, to allow for between-participants variability in the decision-making models the DSS would capture, the feedback was limited to indications of non-compliance with rules—for example, crossing the most intense region—or clarifications on the process. Participants then completed 130 cases, randomly drawn from a pool of 170, where they had to choose the best option for each presented situation. In the next phase, involving 20 new cases, drawn from the same pool, the inverse counterfactual was produced. After selecting the best option, participants were required to choose the second-best option and then modify only one feature to make this second-best option preferable, thus generating an inverse counterfactual case.

Participants were randomly and equally assigned to either the control or the experimental condition. As shown in Figure 2, for participants in the control condition, the DSS was trained only on those cases where participants only selected the first option, that is did not generate counterfactual case. In contrast, for participants in the experimental condition, the DSS was trained with all cases where participants needed to choose the best option, substituting the last 20 cases with 20 inverse counterfactual cases. This allowed to isolate the effect of the inverse counterfactual on the DSS while generalizing any potential effect it might have on the subjects performing it, including self-reflection. A final set of 20 new cases, drawn from the same 170-cases pool, was performed by participants, which involved receiving DSS recommendations when their decisions did not align with the output of the model. As shown in Figure 3, after their initial choice, participants received a recommendation if it diverged from the

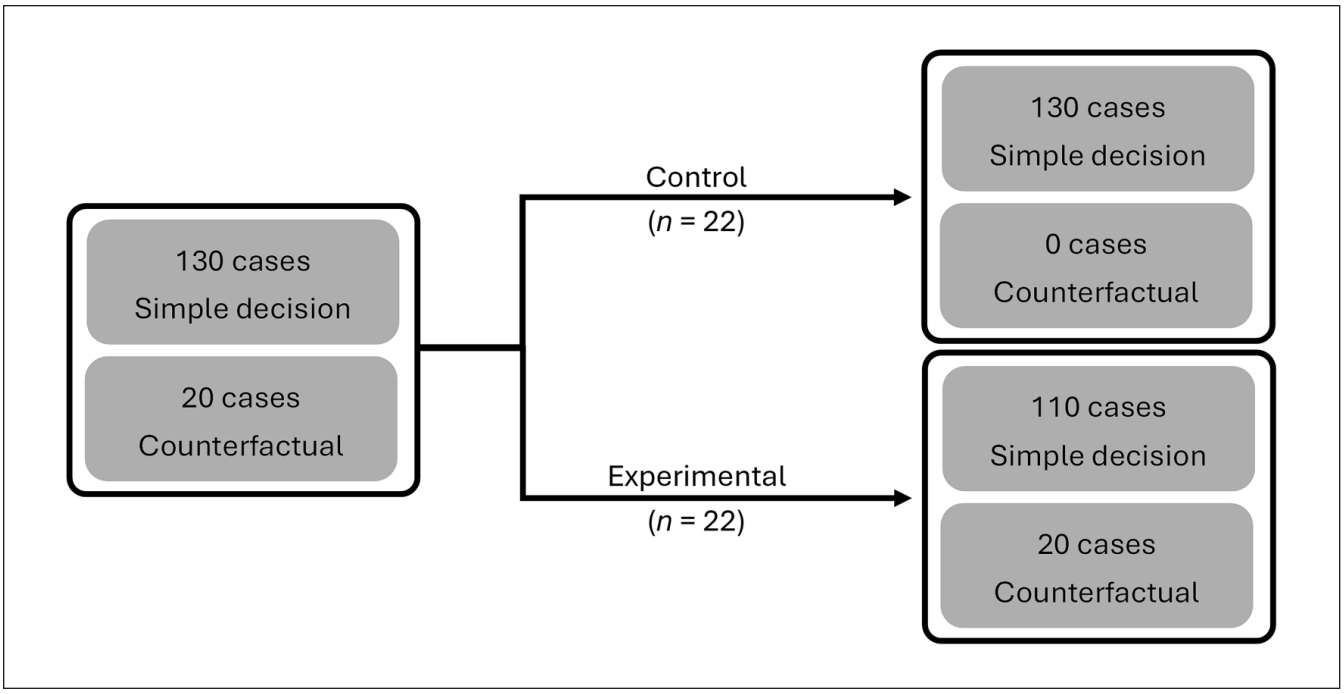

Figure 2. Training case selection for each condition based on the 130 simple decision cases and 20 counterfactual cases completed during the knowledge-elicitation phase.

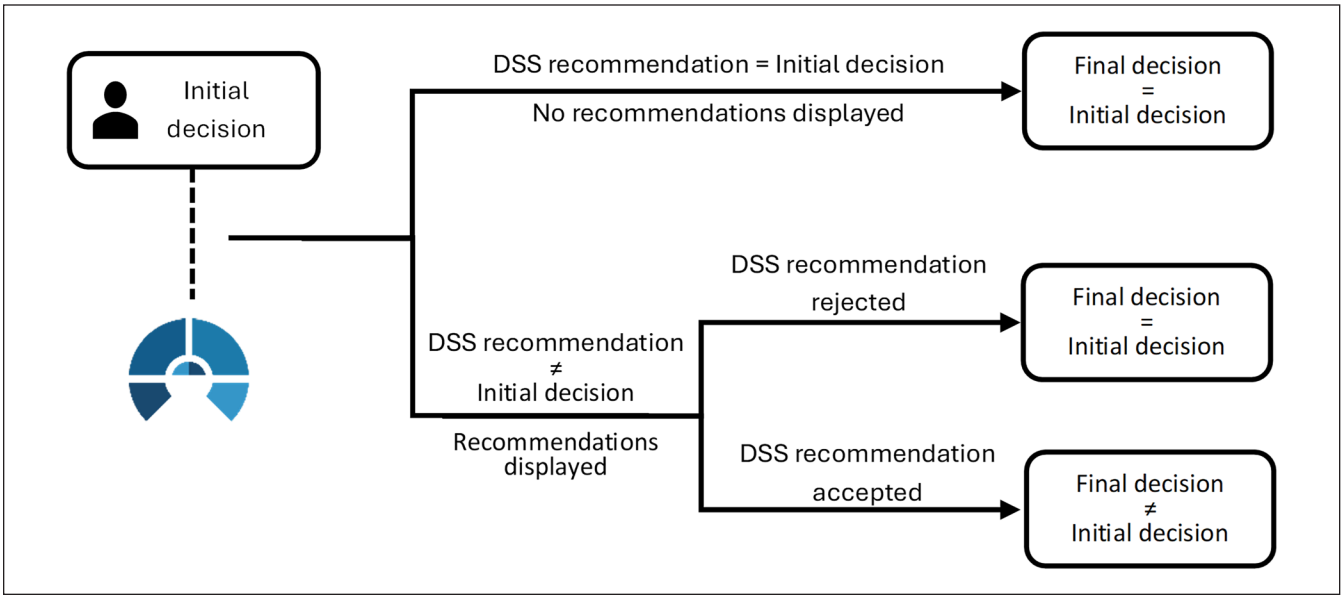

Figure 3. Interaction between the DSS and the participant for each case presented on at a time.

prediction of the DSS, which they could accept or reject (final choice).

Analysis

The performance of each training method was measured by the degree of agreement between the initial and final choice of the participant and the recommendation of the DSS (cf. Figure 3). As previously discussed, the inverse counterfactual method is intended to provide information on the decision boundary of the operator, enabling the DSS to accurately model it. For an inverse counterfactual case to be informative and fully leverage the technique, it must be close to the decision boundary of the user. To estimate this, we computed the

Euclidean distance (Equation 1)—a common practice in proximity analysis in counterfactuals (cf. Keane et al., 2021)—between the feature vector of the initially selected option and that of the newly preferred option, after participant modifications, in each case. Let $\boldsymbol{x}^{(A)} = (x_1^{(A)}, x_2^{(A)}, \ldots, x_n^{(A)})$ denote the normalized feature vector of the initially selected option (Option A), and let $\boldsymbol{x}_c^{(B)} = (x_{c,1}^{(B)}, x_{c,2}^{(B)}, \ldots, x_{c,n}^{(B)})$ represent the normalized feature vector of an initially non-preferred option (Option B), after it was modified by the participant during the counterfactual phase. All features were normalized to the [0, 1] range by dividing each value by the maximum value of its corresponding feature. The Euclidean distance between the two vectors was then calculated as:

$$d\left(x^{(A)}, x_c^{(B)}\right) = \sqrt{\sum_{i=1}^{n}\left(x_i^{(A)} - x_{c,i}^{(B)}\right)^2}$$

where n is the number of features. These distances were then averaged across all cases for each participant to serve as an index of proximity to their individual decision boundary.

To further characterize the informational value of the cases produced by participants in the counterfactual phase, each edited option was also checked for dominance. An option is nondominated when no other feasible alternative is strictly better on all of its features—there is always at least one dimension on which the option matches or surpasses every other alternative (Nayak, 2020). For every participant we then computed the proportion of nondominated counterfactuals they produced during training. This proportion served as an index of how often the participant generated cases where no compromise was made for any of the variables, so a higher share of them should give the DSS clearer information about where they draw the line between acceptable and unacceptable choices.

Results

Decision Prediction

The DSS demonstrated overall a superior performance in predicting the initial choices of participants in the control condition ($M = 71.82$, $SD = 11.19$), compared to the experimental condition ($M = 58.86$, $SD = 17.45$), Mann–Whitney $U = 335.50$, $p = .017$, $r = .39$. A similar pattern was observed for final choices ($M = 85.91$, $SD = 9.21$ for control; $M = 72.27$, $SD = 19.32$ for experimental), Mann–Whitney $U = 350.50$, $p = .003$, $r = .45$. Since all participants generated counterfactual cases, which were then excluded from model training in the control condition, we were able to create alternative models for each participant, simulating the conditions of the opposite group (cf. Figure 2). We generated models with counterfactual cases for control participants and models without counterfactual cases for experimental participants. Since participants did not receive system recommendations from these ad hoc models, only initial choice coincidence could be measured, not the final choice. The control model achieved a mean coincidence of 67.49% ($SD = 12.64$), while the counterfactual model achieved 63.63% ($SD = 15.63$). A Wilcoxon Single-Ranked test comparing coincidence of initial choices between both models did not reach significance ($W = 241.5$, $p = .060$, $Z = -1.88$), suggesting that neither technique was definitively superior. This result points to variability across participants, suggesting that some were able to fully leverage the counterfactual technique, generating better-performing models as a result.

Counterfactual Case Quality

Considering these results, we aimed to identify the factors contributing to the observed differences and understand why some participants were able to effectively utilize the technique. The Euclidian distance between the original case and the inverse counterfactual generated was found to be negatively and significantly correlated with the performance improvement from the model without counterfactuals to the one with counterfactuals, $r(42) = -.34$, $p = .023$. That is, as the distance between initial and alternative decisions increased, the performance prediction gain decreased. Additionally, greater distance was associated with lower alignment between participants' final choices and the DSS recommendations—whether that alignment was based on initial agreement or a change following the DSS suggestion—$r_s(20) = -.50$, $p = .021$. Greater distance also correlated positively with a higher rate of DSS recommendation rejections, $r_s(20) = .50$, $p = .020$.

To assess whether the quality of the counterfactual cases influenced model performance, we tested whether the proportion of nondominated options produced during training predicted the extent of model improvement. A significant positive correlation was found, $r(42) = .42$, $p = .005$, indicating that participants who more consistently generated nondominated counterfactuals tended to benefit more from their inclusion in the DSS training dataset.

Discussion

Drawing on the pivotal role shared mental models play in human teams—enhancing performance through similarity (Cooke et al., 2003; Hanna et al., 2013)—the current study aimed to evaluate the impact of the inverse counterfactual technique on human-AI model similarity and to identify the core characteristics of an effective inverse counterfactual case. Contrary to our initial expectations, the current implementation of the technique did not enhance human-AI model similarity, as measured by the degree of agreement in decision choices. On average, between-groups comparisons showed that DSS models trained with inverse counterfactual cases did not outperform those trained with conventional cases. However, the absence of a significant difference in within-participant comparisons

suggests that the observed between-groups differences may not be due solely to the use of the inverse counterfactual technique. This prompted further analysis to explore whether variation in counterfactual case quality could account for differences in model performance.

The distance analysis indicated that performance depended on the extent of participant modifications: when edits were minimal—that is, when the average Euclidean distance between the original and the paired case was low—the counterfactual model outperformed the conventional model for that participant, reflected in higher user-DSS agreement. In contrast, larger modifications pushed the counterfactual example further from the user's true decision boundary, degrading model performance. Minimal modifications are likely to reveal borderline regions between classes, offering insight into where class transitions occur and where participant decisions tend to be more variable (Pascual-Triana et al., 2025). Larger modifications, on the other hand, may introduce noise or reinforce already well-sampled problem-space regions, leading to overrepresentation. In short, operator-driven counterfactuals appear most informative when they remain close to the factual case, aligning with the "proximity" principle that underlies algorithmic counterfactual generators in explainable AI (e.g., Suffian et al., 2024).

These difficulties are consistent with earlier findings. MacLean et al. (2024) reported that sonar analysts struggled when they were allowed to freely modify many variables. In the current study, guidance provided on counterfactual case generation and restrictions to the number of allowed modifications improved performance for some participants by helping them to generate more informative cases. However, most participants still had difficulty producing counterfactual cases near their individual decision boundary, as reflected in the large modifications they made, and the experimental group performed worse than the control group. However, the results suggest an effect of model similarity on user-DSS interaction as measured by an increase in coincidence between the final choice of the participant and the recommendation of the system. Specifically, smaller counterfactual modifications—that is cases closer to the participant's decision boundary—were associated with fewer rejections of DSS recommendations. In other words, as counterfactual cases became more informative, the DSS not only predicted more accurately the initial choice of participants but also generated recommendations that were more likely to be accepted.

The current study demonstrates both the potential and technical feasibility of inverse counterfactuals for enhancing knowledge elicitation in training an AI-based DSS, particularly for participants who were able to fully leverage the technique. The guidance provided for generating inverse counterfactual cases was improved based on the previous findings by MacLean et al. (2024). However, the results also highlight the need for further refinement of the method to better align it with human cognitive processes, ensuring that the generated counterfactual cases not only accurately represent the decision model of the operator but also provide information about boundary regions. Future research will address these challenges by leveraging the ability of AI to generate counterfactuals that users can validate. Building on these findings, upcoming work will focus on improving AI integration into DSS, with the goal of supporting more reliable and effective human-AI teaming.

Acknowledgments

Thanks are due to Denis Ouellet for his technical support in the deployment of the experimental platform, and to Coralie Bureau for assistance in data collection. We also thank the Thales development team for software implementation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Mitacs Canada through the Mitacs Accelerate program.

ORCID iDs

Jonay Ramon Alaman https://orcid.org/0000-0002-8642-0422
Daniel Lafond https://orcid.org/0000-0002-1669-353X
Alexandre Marois https://orcid.org/0000-0002-4127-4134
Sébastien Tremblay https://orcid.org/0000-0002-7030-5534

References

Chou, Y. L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, *81*, 5983. https://doi.org/10.1016/j.inffus.2021.11.003

Cooke, N. J., Salas, E., Kiekel, P. A., Stout, R., Bowers, C., & Cannon-Bowers, J. (2003). Measuring team knowledge: A window to the cognitive underpinnings of team performance. *Group Dynamics*, *7*(3), 179–199. https://doi.org/10.1037/1089-2699.7.3.179

Endsley, M. R., & Jones, D. G. (2012). *Designing for situation awareness: An approach to user-centered design* (2nd ed.). CRC Press. https://doi.org/10.1201/b11371

Hanna, N., Richards, D., & Hitchens, M. (2013). Evaluating the impact of the human-agent teamwork communication model (HAT-CoM) on the development of a shared mental model. *Lecture Notes in Computer Science*, *8291*, 453–460. https://doi.org/10.1007/978-3-642-44927-7_34

Keane, M. T., Kenny, E. M., Delaney, E., & Smyth, B. (2021). If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. arXiv preprint arXiv:2103.01035. https://doi.org/10.48550/arXiv.2103.01035

Labonté, K., Lafond, D., Chatelais, B., Hunter, A., Akpan, F., Neyedli, H. F., & Tremblay, S. (2021). Combining process tracing and policy capturing techniques for judgment analysis in an anti-submarine warfare simulation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 1557–1561. https://doi.org/10.1177/1071181321651113

Lafond, D., Labonté, K., Hunter, A., Neyedli, H. F., & Tremblay, S. (2020). Judgment analysis for real-time decision support using the Cognitive Shadow policy-capturing system. In T. Ahram, R. Taiar, S. Colson, & A. Choplin (Eds.), *Human interaction and emerging technologies. IHIET 2019. Advances in intelligent systems and computing* (Vol. 1018, pp. 78–83). Springer.

MacLean, M., Lafond, D., & Li, J. (2024). *Capturing expert judgment policies for assessing automated sonar range prediction credibility*. 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS), Toronto, ON, Canada, pp. 1–5. https://doi.org/10.1109/ICHMS59971.2024.10555599

Marois, A., Lafond, D., Audouy, A., Boronat, H., & Mazoyer, P. (2023). Policy capturing to support pilot decision-making. *Aviation Psychology and Applied Human Factors*, *13*(1), 26–38. https://doi.org/10.1027/2192-0923/a000237

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Nayak, S. (2020). *Fundamentals of optimization techniques with algorithms*. Academic Press. https://doi.org/10.1016/C2019-1-02539-9

Nokes, K., & Hodgkinson, G. P. (2018). Policy-capturing: An ingenious technique for exploring the cognitive bases of work-related decisions. In R. J. Galavan, K. J. Sund, & G. P. Hodgkinson (Eds.), *Methodological challenges and advances in managerial and organizational cognition* (pp. 95–121). Emerald Publishing Limited, UK.

Nuñez, J., de la Hogue, T., Duchevet, A., Bonelli, S., & Manuel, M. J. (2019). *D2.1: Analysis of potential cognitive computing-aided tasks* (Report No. D2.1, Edition 00.01.00). HARVIS Project. https://www.harvis-project.eu/wp-con-tent/uploads/2020/03/HARVIS-D2.1_Analysis-of-Potential-Cognitive-Computing-Aided-Tasks_ed00.01.00.pdf

O'Neill, T. A., Flathmann, C., McNeese, N. J., & Salas, E. (2023). 21st century teaming and beyond: Advances in human-autonomy teamwork. *Computers in Human Behavior*, *147*, Article 107865. https://doi.org/10.1016/j.chb.2023.107865

Pascual-Triana, J. D., Fernández, A., Del Ser, J., & Herrera, F. (2025). Overlap Number of Balls Model-Agnostic Counter Factuals (ONB-MACF): A data-morphology-based counterfactual generation method for trustworthy artificial intelligence. *Information Sciences*, *701*, Article 121844. https://doi.org/10.1016/j.ins.2024.121844

Ramon Alaman, J., Lafond, D., & Tremblay, S. (2024, May). *A counterfactual knowledge elicitation method for modeling pilot decision making* [Conference session]. The International Conference on Cognitive Aircraft Systems, Toulouse, France.

Steyvers, M., & Kumar, A. (2023). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, *19*(5), 722–734. https://doi.org/10.1177/17456916231181102

Suffian, M., Alonso-Moral, J. M., & Bogliolo, A. (2024). Introducing user feedback-based counterfactual explanations (UFCE). *International Journal of Computational Intelligence Systems*, *17*(1), 123. https://doi.org/10.1007/s44196-024-00508-6

Wang, X., & Chen, X. (2024). Towards human-AI mutual learning – A new research paradigm. arXiv preprint arXiv:2405.04687. https://doi.org/10.48550/arXiv.2405.04687