Human Interaction and Emerging Technologies (IHIET-AI), Vol. 201, 2026, 190–201
https://doi.org/10.54941/ahfe1007173

Human Autonomy Teaming and AI Metacognition in Maritime Threat Assessment

Kathryn J. Schulze,[1] Adèle Gallant,[2] Tanya Paul,[3] Cindy Chamberland,[2] Daniel Lafond,[3] Sébastien Tremblay,[2] and Heather F. Neyedli[1]

[1]Dalhousie University, Halifax, NS B3H 4R2, Canada
[2]Université Laval, Québec City, QC G1V 0A6, Canada
[3]Thales, cortAIx Labs, Quebec City, QC G1P 4P5, Canada

ABSTRACT

Human autonomy teaming (HAT) requires artificial agents not only to perform taskwork effectively but also to engage in adaptive teamwork processes such as transparency, shared learning, and metacognitive self-monitoring. This study presents baseline findings from a simulated maritime threat assessment task designed to support future integration of Cognitive Shadow (CS), an artificial intelligence (AI)-enabled decision-support system capable of modelling expert strategies and estimating its own reliability. Thirty-five participants completed three counterbalanced maritime surveillance scenarios and provided perceived workload ratings, self-confidence ratings, and situation awareness. Performance metrics of accuracy, recall, and precision were calculated, and confusion matrices were used to characterize classification behaviours. All scenarios showed consistent error patterns: the middle category, *uncertain,* was most frequently misclassified, and participants tended toward conservative decision strategies. These stable behavioural trends indicate room for performance improvement and provide essential groundwork for implementing CS and metacognitive capabilities. Future phases will evaluate whether CS reduces workload, maintains situation awareness, improves performance, and fosters appropriately calibrated trust in AI teammates through the implementation of metacognition.

Keywords: Human-autonomy teaming, Human-AI co-learning, Metacognition, Human factors

INTRODUCTION

Human autonomy teaming (HAT) is becoming increasingly important across a range of domains, including manufacturing, transportation, defence and security. This collaboration is essential for maintaining competitiveness, improving cost efficiency and enhancing safety in these sectors. While these systems are performing more advanced tasks, there is a need to develop their teamwork capabilities to provide better sociotechnical integration with their human teammates. In other words, there is not only a need to support human and artificial intelligence (AI)-based counterparts learning how to perform required taskwork, but they also need to learn how to engage together in the required teamwork processes.

Received August 6, 2025; Revised January 14, 2026; Accepted January 24, 2026; Available online April 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

For machines to become effective teammates, they must not only exchange information transparently and anticipate human actions but also adapt dynamically to changing contexts (Stowers et al., 2021). Integrating the concept of co-learning, this perspective suggests that effective human-machine teaming requires both parties to evolve together through shared experiences, mutual feedback, and continuous learning (Lu et al., 2025; Ramon Alaman et al., 2025). Co-learning fosters a reciprocal development of skills and understanding, enabling machines to refine their models based on human behaviour while humans adapt to the capabilities and limitations of their artificial counterparts. To support co-learning and adaptive teaming, automated agents must also develop metacognitive abilities such as monitoring their own performance, recognizing uncertainty, and adjusting strategies in real time, which are essential for fostering mutual understanding and improving team-level decision-making (McGrath et al., 2025).

Effective teamwork processes in HAT should lead to reduced cognitive workload for the human teammate without subsequent loss of shared situation awareness (SA), supported by transparent communication of the AI system's state, intent, and limitations (Endsley, 2023). Further, appropriate levels of trust in the automated teammate help foster the appropriate sharing of task work between human and automated teammates (Lee & See, 2004; Parasuraman & Riley, 1997). For the automated teammate, effective teamwork processes should lead to faster and more accurate taskwork learning and more effective co-learning. These teamwork processes mirror core elements of human–human teams, such as communication, shared mental models, and team SA, which have been identified as equally central for human–AI teaming (Berretta et al., 2023).

Cognitive Shadow (CS) is a toolkit that uses supervised machine learning to learn expert decision-making patterns (i.e., policy capturing) and support real-time HAT through a process of judgmental bootstrapping with either proactive or corrective feedback (Armstrong 2001; Lafond et al., 2020). It continuously refines its recommendations by dynamically adjusting models based on immediate user feedback, enhancing decision quality and human-autonomy teaming effectiveness (Marois et al., 2023). CS was recently extended to enable "meta-models" that monitor the performance of a trained model in order discover non-random error patterns and predict AI reliability in different situations, providing a new meta-cognitive function for trustworthiness assessment and human-AI teaming (Lafond et al., under review).

HAT systems have been successfully integrated into various industries, including aspects of national defence. In the Canadian Arctic, climate change continues to expand waterways, increasing available routes and, in turn, maritime traffic. For high-fidelity results in this field of research, it is common to implement simulation (Alaman et al., 2017; Lafond et al., 2017; Lafond et al., 2020; Marois et al., 2023a; Marois et al., 2023b). One area of interest for deploying CS is in the support of maritime surveillance, which relies on an integrated approach combining advanced sensors, platforms, and systems to ensure comprehensive SA to enable effective threat assessment.

Here, an adaptive command and control (C2) framework with adjustable human-autonomy collaboration is proposed to enhance threat assessment performance. This framework will be instantiated and evaluated in simulated human-in-the-loop maritime scenarios. Central to this approach is the integration of CS, a learning system that captures and models expert human decision patterns. New AI metacognition capabilities have expanded CS, using a recursive approach to model its own reliability based on situation attributes. This AI metacognition provides an empirically grounded reliability metric to help the human collaborator decide whether to rely on the AI or not.

The purpose of this study is to test this novel maritime surveillance simulation using three scenarios, based on a common template to ensure comparable levels of workload, SA, and self-confidence. An additional aim was to collect baseline data (no artificial support decision support) to train CS for future integration. Lastly, the baseline maritime surveillance task data will be compared with future phases of this study that will implement the use of CS and metacognition. This paper will present our baseline data and our predictions for the next phases with the integration of CS. We hypothesize that there will be no significant difference in workload, SA, or self-confidence across the three scenarios, which would be ideal to study the effects of human-AI co-learning over successive scenarios in a future phase.

METHODOLOGY

Participants

The research was approved by the human research ethics committee of Université Laval (approval number 2024–098 A-1/09–09–2025). Thirty-nine participants were initially recruited from Université Laval to take part in a single experimental session. The final analyzed sample consisted of 35 participants (age: 27.4 ± 10.51 years; gender: 60.0% female, 37.1% male, 2.9% other; occupation: 88.6% students, 8.6% employed, 2.9% other).

Task Environment and Scenarios

Participants completed a simulated maritime threat assessment task involving the continuous classification of entities displayed on a surveillance interface. The task required sustained monitoring, repeated information updating, and iterative decision making under dynamic conditions.

The experiment consisted of one training scenario followed by three experimental scenarios. Each scenario lasted 12 minutes and contained 30 entities that participants were required to classify as *friendly*, *uncertain*, or *suspect*. Classification decisions were based on each entity's observable features and the threat level associated with each feature (Figure 1). Features informed participants of characteristics such as an entity's automatic identification system (AIS) being *on* or *off,* and if there was known military intelligence (HumINT) about that entity. Specifically, participants classified entities by counting the number of features exhibiting *suspect* behaviour and applying predefined thresholds that mapped feature counts to the three threat categories (Table 1).

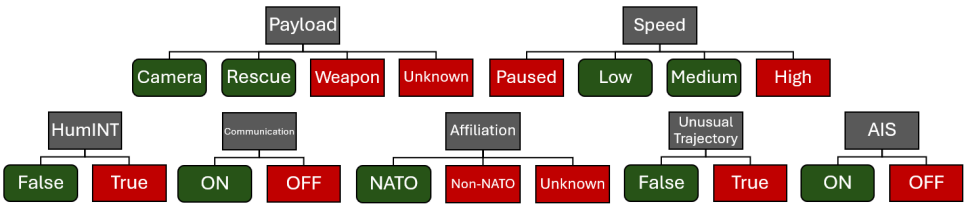

Figure 1: Feature characteristics for entity classification – see decision-rules (Table 1).

Table 1: Decision-rules implemented for the classification of entities as *friendly*, *uncertain*, or *suspect* based on reported features.

Class	Description of the decision-rule
Friendly	0 - 2 features are red squares
Uncertain	3 - 4 features are red squares
Suspect	5 - 7 features are red squares

Entity features changed dynamically at three time points during each scenario: T0 at the start of the scenario, T1 at one-third of the scenario, and T2 at two-thirds of the scenario, necessitating continuous monitoring and frequent inspection of entities throughout each scenario.

Although the scenarios differed in narrative context, they shared identical task structure, decision-rules, timing, performance demands, and simulated maritime surveillance interface (Figure 2). This design ensured that the scenarios were well-matched and suitable for evaluating the effects of CS introduced in subsequent phases of the experiment.

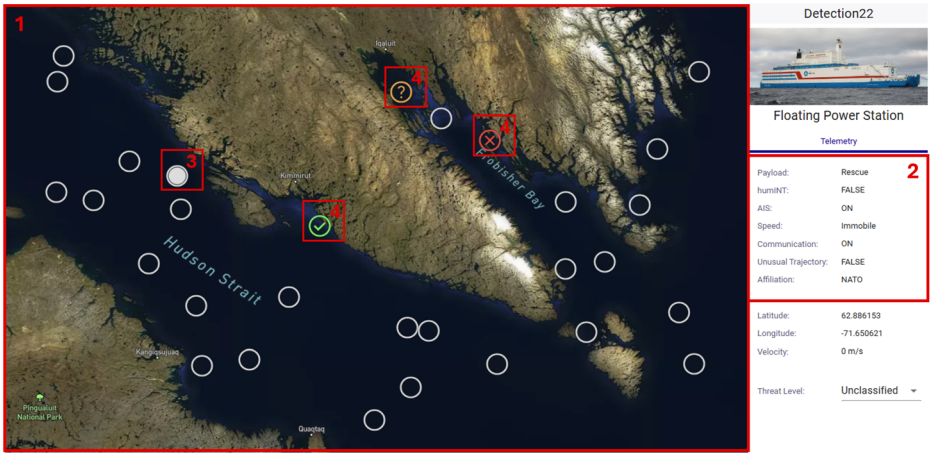

Figure 2: The visual interface of the simulated maritime threat assessment task: **(1)** a map displaying all entities, **(2)** a list of features related to the entity, **(3)** an entity when selected, **(4)** once an entity is classified, its corresponding circle changes colour to reflect its assigned threat category: *friendly*, *uncertain*, or *suspect*.

Training Scenario - Hostile Payload During Regular Shipping Route Operations: Regular shipping route operations are underway; some hostile vessels are carrying surveillance equipment or weapons. Operators need to identify hostile vessels with limited information.
Scenario A – After Impact/Attack - Sensor Damage Limits Detection: Following an attack, some sensors are damaged, creating undetectable zones in the surveillance area. Operators need to make decisions with limited information.
Scenario B – Blockage from Enemy: Hostile forces plan to create blockades, obstructing navigation and restricting operational zones. The objective is to identify hostile actors and maintain/restore safe access.
Scenario C – Infiltration During Search and Rescue: During a search and rescue operation involving commercial and humanitarian vessels, hostile actors attempt to infiltrate under the guise of legitimate activities.

Procedure

Participants completed one 12-minute Training scenario followed by three 12-minute experimental scenarios presented in a counterbalanced order (Figure 3). During the Training scenario, participants were provided with a visual aid combining the feature characteristics decision tree (Figure 1) and the classification decision-rule table (Table 1). During the experimental scenarios, this visual aid was removed, and participants were required to apply the same classification rules from memory. After completing each scenario, participants completed a set of psychometric questionnaires assessing workload, self-confidence, and SA. The full experimental session lasted approximately 90 minutes.

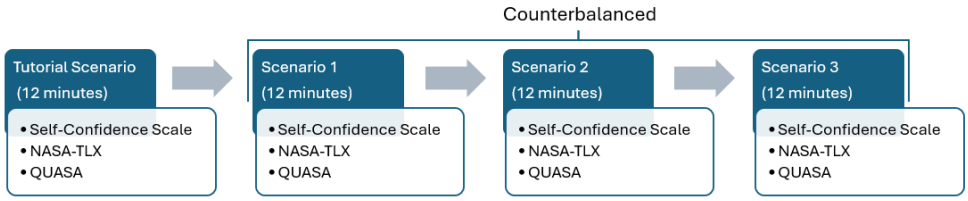

Figure 3: Study protocol.

Measures

Psychometric Measures

Self-confidence – Participants rated their overall confidence in their classification performance for each completed scenario on a 0-100 scale (Rittenberg et al., 2024).

Workload – Perceived workload was assessed using a modified version of the NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988), consisting of five dimensions (mental demand, temporal demand, effort, frustration, and performance), each rated on a 0–100 scale, with the physical demand dimension omitted. An overall workload score was computed by summing the five-dimension scores, yielding a composite score ranging from 0 to 500.

Situation Awareness – SA was measured using an adapted version of the Quantitative Analysis of Situation Awareness (QUASA; McGuinness, 2004). At the end of each scenario, participants responded to a set of true or false probes designed to assess their awareness and comprehension of key elements of the preceding scenario. For each probe, participants also rated their confidence in their response using an ordinal confidence scale. QUASA accuracy was calculated as the proportion of correct responses across the true or false probes associated with the completed scenario. QUASA calibration bias was calculated to quantify the correspondence between participants' confidence ratings and their QUASA accuracy, with higher values indicating overconfidence and lower values indicating underconfidence.

Performance Measures

Participants' classification responses were recorded continuously throughout each scenario to characterize task performance: *i) Classification accuracy* – Proportion of correctly classified entities relative to ground truth, averaged per participant and scenario; *ii) Confusion matrices* – Confusion matrices were generated for each scenario to characterize the distribution and direction of classification errors across threat classes (*friendly, uncertain, suspect);* and *iii) Class-wise precision and recall* – Precision and recall were computed for each threat class to characterize detection accuracy and systematic error tendencies across scenarios, based on ground-truth and participant response distributions, respectively.

Statistical Analysis

One-way repeated-measures ANOVAs compared psychometric and performance measures across the Training scenario and the three experimental scenarios. Post-hoc pairwise comparisons were conducted using estimated marginal means with Holm-adjusted p-values. Mean values and standard deviations are reported for all measures, and statistical significance was evaluated using an alpha level of .05. Outliers were identified post-hoc using the interquartile range method ($Q1 - 1.5 \times IQR$; $Q3 + 1.5 \times IQR$). All four identified outliers fell below the lower bound and were excluded from analysis.

To assess potential learning via performance accuracy and workload, additional one-way repeated-measures ANOVAs were conducted using scenario position (positions 1 to 4) as a within-subjects factor, where position 1 was always the Training scenario, and positions 2–4 corresponded to the three experimental scenarios in order of presentation as they were counterbalanced. Mauchly's test of sphericity assessed sphericity; if $p > .05$, Greenhouse-Geisser corrections were applied.

Furthermore, planned paired-samples t-tests compared position 1 (Training Scenario) to the average of positions 2–4 for both workload and accuracy. As participants had a visual aid during the Training trial (Figure 1 and Table 1), this comparison allowed further assessment of the possible effect of this aid on performance or perceived workload.

RESULTS

Self-confidence ratings ranged from 17 to 100 ($M = 74.04$, $SD = 19.38$) and did not differ significantly between scenarios, $F(3, 132) = 0.21$, $p = .89$. Summed NASA-TLX scores, of perceived workload, ranged from 101 to 432 ($M = 258.12$, $SD = 62.89$) and did not differ significantly between scenarios, $F(3, 132) = 0.33$, $p = .81$. SA, assessed using QUASA, did not differ significantly between scenarios, $F(3,132) = 1.01$, $p = 0.89$ QUASA calibration bias did not differ significantly between scenarios, $F(3, 132) = 0.61$, $p = .61$.

Table 2: QUASA accuracy and calibration bias for all scenarios (mean ± SD).

	Training	Scenario A	Scenario B	Scenario C
Accuracy	0.57 ± 0.21	0.61 ± 0.24	0.67 ± 0.21	0.59 ± 0.24
Calibration bias	0.11 ± 0.24	0.10 ± 0.26	0.08 ± 0.20	0.14 ± 0.26

Turning to task performance, classification accuracy differed between scenarios ($M = 80.80$, $SD = 7.90$), with a significant effect of scenario, $F(3, 102) = 9.14$, $p < .001$. Mean classification accuracy was 82.51% for the Training Scenario, 79.60% for Scenario A, 76.97% for B, and 84.92% for C. Post hoc paired comparisons with Holm correction indicated that accuracy in Scenario C was significantly higher than in Scenario A ($p < .01$) and Scenario B ($p < .001$). In addition, accuracy in the Training scenario was significantly higher than in Scenario B ($p < .01$). No significant differences were observed between Scenarios A and B, or between the Training scenario and Scenario C (all $ps > .05$).

To further characterize classification behaviour, confusion matrices revealed consistent error patterns across scenarios (Figure 4). *Friendly* entities were most frequently classified correctly, whereas misclassifications were primarily concentrated around the *uncertain* category. Errors most often involved confusion between *uncertain* and the *friendly* or *suspect* categories, whereas direct misclassifications between *friendly* and *suspect* were rare. This error structure was consistent across scenarios, including the Training scenario, indicating stable classification behaviour across task conditions.

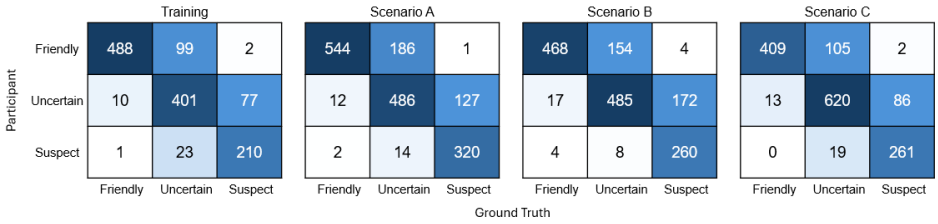

Figure 4: Confusion matrices for detections and error patterns of *friendly*, *uncertain*, and *suspect* distributions for the ground truth (y-axis) and participant entity classifications (x-axis) for each experimental scenario and the Training scenario.

Class-wise precision and recall analyses showed that *friendly* entities yielded the highest recall, whereas suspect entities yielded the highest precision (Table 3). *Uncertain* entities demonstrated moderate precision and recall across scenarios, reflecting their role as an intermediate classification category.

Table 3: Percentages of precision (P) and recall (R) metrics summarizing participant classification performance for each threat class across the four scenarios.

Threat Class	Training		Scenario A		Scenario B		Scenario C	
	P	R	P	R	P	R	P	R
Friendly	82.85	97.80	74.42	97.49	74.76	95.71	79.26	96.92
Uncertain	82.17	76.67	77.76	70.85	71.96	74.96	86.23	83.33
Suspect	89.74	72.66	95.24	71.43	95.59	59.63	93.21	74.79

The repeated-measures ANOVA determined a significant effect of scenario position (1–4) on workload, $F(3, 102) = 3.75$, $p = .013$, $\eta^2 = .020$. Greenhouse–Geisser correction was applied ($W = 0.365$, $p < .001$), and the effect remained significant ($p = .029$). However, Holm-adjusted post-hoc comparisons between positions were not significant (all ps $> .05$), possibly reflecting reduced power of post-hoc tests (Chen, 2018). Furthermore, the planned paired-samples t-test showed significantly lower workload during Training compared to the average of the other scenarios, $t(34) = -2.48$, $p = .018$, with a mean difference of 19.4 units (95% CI [−35.27, −3.53]).

The repeated-measures ANOVA determined a no significant effect of scenario position (1–4) on performance accuracy, $F(3, 102) = 0.83$, $p = .48$. The paired t-test comparing Training to the average of the experimental scenarios was also non-significant, $t(34) = 1.44$, $p = .158$, 95% CI [−0.009, 0.052].

DISCUSSION

The primary objective of this study was to evaluate a novel maritime surveillance simulation designed to support the investigation of human–automation interaction in future experimental phases. Specifically, this baseline phase of the study aimed to determine whether multiple scenarios constructed from a common template would elicit comparable levels of workload, self-confidence, and SA, while remaining sensitive to performance-related variation. Establishing equivalence in subjective and cognitive demands, alongside measurable differences in task performance, is essential to ensure that any effects observed in subsequent phases can be attributed to the introduction of CS and metacognitive support rather than to uncontrolled differences in task difficulty or structure.

There were no significant differences in workload, self-confidence, or SA. Participant performance was best during scenario C; however, it did not significantly differ between scenario A and B, and no scenario performance was nearing ceiling. These measures indicated variability across participants and

scenarios, with room for improvement, which allows for the implementation of CS, and the further implementation of metacognition in future phases to serve as aids and lead to improvement.

As indicated by the confusion matrices and precision/recall analyses by threat class, participants demonstrated a consistent pattern in how they approached the classification task across all scenarios, including Training. For incorrect classifications, participants tend to select a less *threatening* class: selecting *uncertain* for *suspect* entities, and *friendly* for *uncertain* entities. Additionally, class-specific recall indicated that participants were more reliable at identifying the presence of a *friendly* entity, whereas class-specific precision indicated that participants had fewer false alarms when identifying *suspect* entities. These trends were consistent across all scenarios, suggesting a conservative decision strategy, characterized by a preference for safer classifications that minimize the risk of overclassifying potential threats. This pattern indicates reliance on stable decision heuristics rather than scenario-specific cues and aligns with the Recognition-Primed Decision framework, in which operators default to low-risk interpretations unless evidence strongly indicates otherwise (Klein, 1993). Furthermore, this aligns with uncertainty promoting conservative decision thresholds to avoid high-cost errors (Wickens, 2002).

Overall, participants' performance suggests that learning occurred rapidly within the Training scenario, indicating that, despite the critical and cognitively demanding nature of the task, the decision-rule was relatively straightforward to acquire and apply consistently across scenarios.

Further analysis of scenario order revealed no evidence of systematic learning or fatigue effects. Classification accuracy did not vary significantly by scenario position, and no performance advantage was observed in the initial Training scenario compared to the experimental scenarios. Although workload was lower during the Training scenario, this likely reflects the temporary cognitive support provided by the visual aids rather than a meaningful difference in task demands. The absence of a position-based performance effect reinforces the conclusion that participants adopted a consistent classification strategy and that scenario content, rather than presentation order, accounted for observed performance differences.

Considerations for Future Phases

In addition to the measures reported in this baseline phase, several performance-related metrics were collected to support analyses across subsequent phases of the study. Throughout each scenario, measures of critical change detection, time to classify, and classification omissions were recorded. Critical change detection was operationalized separately for *friendly* and *suspect* transitions, reflecting the time elapsed between a change in ground truth status and the participant's accurate classification of the entity.

Building on these baseline findings, phases two and three will incorporate CS and metacognitive features into the task. In phase two, CS will provide a classification recommendation for each entity upon selection, allowing us to examine how AI-generated suggestions influence participants' established

decision tendencies. Phase three will extend this by presenting each recommendation with a confidence value, a key aspect of metacognition, that reflects the AI system's certainty of that classification. This addition will allow us to evaluate whether transparent communication of AI confidence helps operators interpret recommendations more effectively and calibrate their own overconfidence more appropriately.

Psychometric measures collected during phases two and three will mirror those used in the baseline condition, including assessments of self-confidence, SA, and workload, thereby enabling direct comparison across all phases of the study. In addition, phases two and three will include a measure of trust in AI, assessed after each scenario using the Trust in Automation scale (Rittenberg et al., 2024). Tracking trust alongside performance and psychometric measures will allow evaluation of whether increased transparency through confidence displays promotes appropriate reliance while preserving SA, maintaining a manageable workload, and supporting task performance. This focus is motivated by recent findings showing that the benefits of transparency are not uniform but depend on how confidence information is interpreted and integrated by human operators, with evidence that such information can shape not only trust but also users' understanding and reliance strategies depending on how it is presented (Gegoff et al., 2025; McGrath et al., 2025).

This design approach is consistent with recent theoretical perspectives emphasizing that effective human–AI teaming should rely on support and assistance, rather than replacing human judgment, particularly in dynamic and uncertain environments (Schulke & Reiman, 2025). Their framework highlights the importance of preserving operator agency while reducing cognitive demands. The stable decision strategies observed in the baseline phase, therefore, provide a critical point of comparison for determining whether CS recommendations and metacognitive confidence displays strengthen, refine, or alter how participants approach the classification task across subsequent phases.

CONCLUSION

Overall, this baseline phase has established that the maritime surveillance task will provide a well-controlled experimental foundation, characterized by equivalence in subjective and cognitive demands alongside meaningful sensitivity in performance and structured decision-making behaviour. These properties are essential for isolating the effects of implementing AI recommendations in later phases. Although full co-learning mechanisms are not yet implemented, the planned introduction of metacognitive confidence displays represents a foundational step toward adaptive human–AI teaming by supporting mutual understanding, calibrated reliance, and early components of co-learning, in which humans and artificial agents progressively adapt to one another through interaction and feedback (Lu et al., 2025; Ramon Alaman et al., 2025; Stowers et al., 2021).

ACKNOWLEDGMENT

We are thankful to Coralie Bureau and Jeanne Nicole for assistance with data collection and to Dr Aren Hunter and Dr Mary MacLean of DRDC Atlantic for their advice on the development of the simulation scenarios. Thanks are also due to Steven Thomas and Sylvain Comtois for the simulation software development and data processing. This work is part of a research program supported by grants from the Natural Sciences and Engineering Research Council of Canada [ALLRP 580533-22] and Prompt Québec [185_2022.12] awarded to Sébastien Tremblay and Heather Neyedli with DRDC Atlantic and Thales Canada.

REFERENCES

Armstrong, J. S. (2001). Judgmental bootstrapping: Inferring experts' rules for forecasting. In *Principles of forecasting: A handbook for researchers and practitioners* (pp. 171–192). Boston, MA: Springer US.

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://doi.org/10.3389/frai.2023.1250725

Chen, T., Xu, M., Tu, J., Wang, H., & Niu, X. (2018). Relationship between Omnibus and Post-hoc Tests: An Investigation of performance of the F test in ANOVA. *Shanghai Archives of Psychiatry*, *30*(1), 60–64. https://doi.org/10.11919/j.issn.1002-0829.218014

Endsley, M. R. (2023). Supporting Human-AI Teams:Transparency, explainability, and situation awareness. *Computers in Human Behavior*, *140*, 107574. https://doi.org/10.1016/j.chb.2022.107574

Gegoff, I., Tatasciore, M., Bowden, V. K., & Loft, S. (2025). Deciphering Automation Transparency: Do the Benefits of Transparency Differ Based on Whether Decision Recommendations Are Provided? Human Factors, 00187208251318465. https://doi.org/10.1177/00187208251318465

Hart, S., Staveland, E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. Advances in Psychology. 52, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Klein, G. (1993). A Recognition Primed Decision (RPD) Model of Rapid Decision Making. In Decision making in action: Models and methods.

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: Ahram, T., Taiar, R., Colson, S., Choplin, A. (eds) *Human Interaction and Emerging Technologies*. IHIET 2019. Advances in Intelligent Systems and Computing, vol 1018. Springer, Cham. https://doi.org/10.1007/978-3-030-25629-6_13

Lafond, D., Paul, T.S., & Auouy, A. (under review). Metacognitive Skills for Trustworthy Intelligent Agents in Collaborative Situation Assessment. IEEE Transactions on Human Machine Systems.

Lafond, D., Roberge-Vallières, B., Vachon, F., & Tremblay, S. (2017). Judgment Analysis in a Dynamic Multitask Environment: Capturing Nonlinear Policies Using Decision Trees. Journal of Cognitive Engineering and Decision Making, 11(2), 122–135. https://doi.org/10.1177/1555343416661889

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lu, J., Yan, Y., Huang, K., Yin, M., & Zhang, F. (2025). Do We Learn From Each Other: Understanding the Human-AI Co-Learning Process Embedded in Human-AI Collaboration. *Group Decision and Negotiation*, *34*(2), 235–271. https://doi.org/10.1007/s10726-024-09912-x

Marois, A., Labonté, K., Lafond, D., Neyedli, H. F., & Tremblay, S. (2023a). Cognitive and behavioral impacts of two decision-support modes for judgmental bootstrapping. *Journal of Cognitive Engineering and Decision Making*, *17*(3), 215–235.

Marois, A., Lafond, D., Audouy, A., Boronat, H., & Mazoyer, P. (2023b). Policy Capturing to Support Pilot Decision-Making: A Proof of Concept Study. Aviation Psychology and Applied Human Factors, 13(1), 26–38. https://doi.org/10.1027/2192-0923/a000237

McGrath, M. J., Duenser, A., Lacey, J., & Paris, C. (2024). Collaborative human-AI trust (CHAI-T): A process framework for active management of trust in human-AI collaboration. *arXiv preprint arXiv:2404.01615*.

McGuinness, B. (2004). Quantitative analysis of situational awareness (QUASA): Applying signal detection theory to true/false probes and self-ratings. Bae Systems Bristol Advanced Technology Center. https://apps.dtic.mil/sti/pdfs/ADA465817.pdf

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors: The Journal of the Human Factors and Ergonomics Society, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Ramon Alaman, J., Lafond, D., Marois, A., & Tremblay, S. (2025). Inverse Counterfactual for AI-Assisted Decision Support: Enhancing Knowledge Elicitation for Capturing Aircraft Pilot Decisions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. https://doi.org/10.1177/10711813251358254

Rittenberg, B. S. P., Holland, C. W., Barnhart, G. E., Gaudreau, S. M., & Neyedli, H. F. (2024). Trust with increasing and decreasing reliability. *Human Factors*, *66*(12), 2569–2589. https://doi.org/10.1177/00187208241228636

Schilke, O., & Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. Organizational Behavior and Human Decision Processes, 188, 104405. https://doi.org/10.1016/j.obhdp.2025.104405

Stowers, K., Brady, L. L., MacLellan, C., Wohleber, R., & Salas, E. (2021). Improving Teamwork Competencies in Human-Machine Teams: Perspectives From Team Science. *Frontiers in Psychology*, *12*, 590290. https://doi.org/10.3389/fpsyg.2021.590290

Wickens, C. D. (2002). Multiple resources and performance prediction. Theoretical Issues in Ergonomics Science, 3(2), 159–177. https://doi.org/10.1080/14639220210123806