The Effects of Inaccurate Decision-Support Systems on Structured Shared Decision-Making for Human-Robot Teams

Jack Kolb*, Divya Srivastava*, Karen M. Feigh

***Abstract*— Human-robot teams can leverage a human's expertise and a robot's computational power to meaningfully improve mission outcomes. In command and control domains, the robot teammate can also act as a *decision-support system* to advise human users. However, decision-support systems are susceptible to human factors issues including miscalibrated trust and degraded team performance. Recent work has mitigated these issues by using *cognitive forcing functions* to structure shared decision-making systems and place users as proactive *on-the-loop* actors. We bring this approach to a human-robot teaming domain, and investigate how Type I and Type II errors in the robot's recommendation affects team performance and user rational trust. We present the architecture of our decision-making process and a Mars rover landing experiment domain. Results from a comprehensive user study demonstrate that the error type of the robot's recommendation forms a trade-off between team performance and rational trust.**

I. Introduction

Robots have strong potential as *decision-support systems* in human-robot teams [22]. While AI-driven decision-support systems are established across numerous industries [18], [10], only recently has the human-robot interaction (HRI) community applied robots as decision-support in real-world applications. In human-robot teaming domains, robot partners can use their understanding of the world state and other teammates to present recommendations on taskwork division, team strategy, and world state interpretation. In recent years researchers have sought to provide robots with an active role in team decision-making across domains including space operation [6], command and control [11], [12], medical teamwork [9], emergency evacuation [17], and industrial production [13].

A critical challenge with decision-support systems in human-robot teams is their effect on the team's human factors. Prior work has indicated the importance of calibrating an appropriate trust in the robot system [20], [14], and human tendencies to *overtrust* (lazily accept) or *undertrust* (ignore) the robot's suggestion. An open problem for the HRI community is mitigating the reduction of user situational awareness and team performance in decision-support robots.

Researchers have approached this challenge in several ways: providing explanations of the robot or AI system's decision [3], [4], studying how humans react to incorrect recommendations [17], and improving the contextual understanding of decision-support systems [19], [21].

Recent work has also explored how the *structure* of a shared decision-making process affects the team's outcomes [2], [14]. Buçinca et. al details three *cognitive forcing functions* that can provide such structure [2]: requiring users to make a preliminary decision before seeing the AI's recommendation [7], time-extending the decision-making process [15], and requiring users to solicit the AI's recommendation [5]. We are interested in applying the first function to adapt the user's decision as they are exposed to additional information and the AI's decision.

Researchers have also studied the effect of inaccurate robot decision-support systems on user trust. Azevedo-Sa et. al studied how user trust is affected by Type I and Type II errors in a vehicle's obstacle detection system [1]. However, little work has addressed human-robot teaming domains where *both* error types are important to the team's performance – i.e. when "missed alarms" are as important as "false alarms".

To address the above challenges and improve the HRI community's understanding of structured shared decision-making systems, we investigate the effects of an inaccurate decision-support robot in a human-robot teaming domain. We evaluate the effects of the robot recommendation's Type I error (*false positive*) and Type II error (*false negative*) on user trust and mission performance. To the authors' knowledge, no prior work has addressed the effect of a decision-support system's error type in an HRI domains.

Concretely, our work explores the research question: *is human-robot team performance and user trust affected by the error type of a structured shared decision-making system?* We address this problem through a space operations human-robot teaming domain, where a human user and a Mars rover decide whether it is safe for the rover to land.

We conducted a user study involving 135 participants. Notably, we found that error type presents a trade-off between *mission performance* and *rational trust*. Participants exposed to *false positive* error type conditions (the robot says it's safe to land when it is not) were more likely to believe the robot's incorrect recommendations and trust the robot. Meanwhile, participants in *false negative* error type conditions (the robot says to abort when it is safe) were more likely to disagree with the robot's incorrect recommendations, but at the cost of reduced rational trust.

All source code used in our experiment can be found at `https://github.com/gt-cec/error-types-in-structured-sdm`.

* denotes equal contribution

This work was funded by Sandia National Laboratory (SNL) with Dr. Paul Schutte serving as Program Manager. This work is solely that of the authors and does not represent an official SNL position.

Jack Kolb, Divya Srivastava, and Karen Feigh are with the Guggenheim School of Aerospace Engineering at the Georgia Institute of Technology, North Avenue, Atlanta, GA 30332, USA {kolb, divya.srivastava, karen.feigh}@gatech.edu

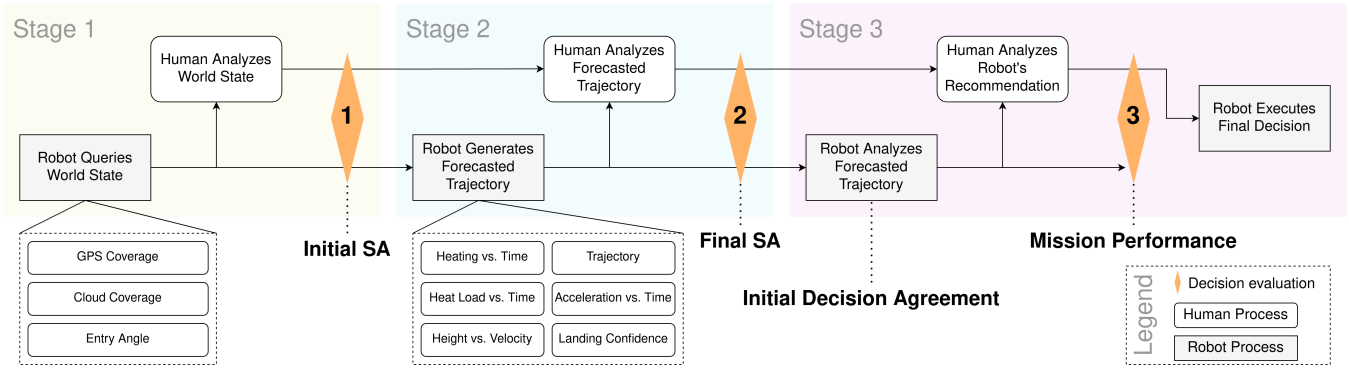

Fig. 1: Diagram overview of our structured decision-making architecture. In each stage, the user reviews an information source (derived from the previous stage's information) and updates their decision as to whether surface and entry conditions are safe for the robot to land. The final stage (Stage 3) presents the robot's recommendation from the world state conditions.

II. Experiment Design

We evaluate how error types affect our shared decision-making architecture through a between-subjects user study with a 3×3 design. In our study, participants support a simulated Mars rover's landing on the Mars surface. Participants play the role of a remote ground commander tasked with making the final decision of whether it is safe for the robot to land. A Mars landing domain was selected because it justifies the robot teammate's relevance and importance, and the tasks encourages users to consider the robot's feedback.

Participants assess three information sources to inform their decision. The *world state information* of the Mars surface, evaluations of the *forecasted trajectory* of the robot's entry, and the *robot's recommendation* of whether it is safe to land. We structure the decision-making process by separating the three information sources and by adding decision points for participants, as shown in Fig. 1.

Participants are randomly assigned to a study condition represented by two independent variables – the participant's access to world state information (*world state awareness*, 3 levels), and the error type of the robot's recommendation (*robot error type*, 3 levels). Each participant completes ten scenarios in their condition.

A. Decision-Making Architecture

To reduce overreliance on the robot's recommendation, we seek to calibrate the user's trust of the robot by extending the decision-making process. Our decision-making architecture enables users to refine their decision as they see new information sources, with the robot's recommendation only shown in the final stage of the process. Fig. 1 visualizes our architecture.

Our architecture forms a three-stage decision-making process. At each stage, users are shown an information source and are granted an opportunity to change their decision. The information sources are not independent – the *forecasted trajectory* and *robot's recommendation* are derived from the *world state information* – allowing users to change their decision as they see alternative views of the scenario data and the robot's recommendation.

The three stages of our architecture are:

- **S1** Users review the *world state information* and declare an initial decision, without having seen the other information sources. The user's decision indicates their *Initial Situational Awareness* (*Initial SA*).
- **S2** Users next review the *figures of merit* for the forecasted trajectory, which they are told will be used by the robot for its recommendation. Users are able to change their decision. The user's decision indicates their *Final Situational Awareness* (*Final SA*). The *world state information* is no longer shown to the user in this stage.
- **S3** Lastly, the *robot's recommendation* is displayed. Users are shown the *robot's recommendation* and are asked to make a final decision. This decision indicates the *mission performance* (w.r.t. the ground truth). No other information sources are shown to the user in this stage.

Importantly, our architecture can accept any black box recommender system for the *robot's recommendation*. **S3** can also be adapted to complex non-binary recommendations including explanations or additional figures of merit, without impacting the objective of introducing a *cognitive forcing function* and separating the user's initial judgement from the *robot's recommendation*. As a result, our methods can be adapted to other human-robot teaming domains.

B. World State Information & Forecasted Trajectory

Participants are shown two sets of figures for each scenario. As detailed in Fig. 1, our architecture includes a decision point after the participant is shown each information source. We adjust the amount of world state information presented to participants to control their situational awareness.

The first information source presented to participants is the world state information. In our domain, the world state information contains three high-level aspects of the scenario: the robot's entry angle, the number of GPS nodes covering the forecasted trajectory, and the cloud coverage along the forecasted trajectory. The aspects are fictitious, but plausible for the domain. The value of each world state aspect classifies the aspect as either *safe* or *risky*. Scenarios with at most one

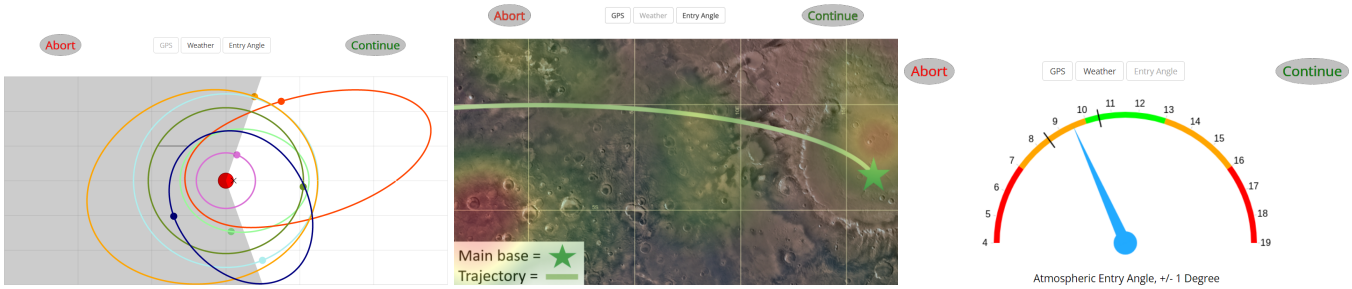

Fig. 2: World state information source shown to users in *Stage 1*. The rating buttons below each image are used in the *interaction* world state awareness condition, while the *observation* condition lacks those features. From left to right: GPS coverage, cloud coverage, entry angle.

risky aspect are considered *safe* for the robot to land. Fig. 2 shows the world state information shown to participants.

The second information source is six figures of merit of the robot's forecasted trajectory. The figures of merit describe the goodness of the trajectory along six different dimensions and are derived from the world state information, including the trajectory, acceleration, heat load, and landing confidence. Fig. 3 shows a forecasted trajectory shown to participants.

While the robot always has access to the world state information to make its recommendation, participants are assigned to one of three levels of world state information:

1) **Absent**: Participants have access to the forecasted trajectory figures of merit, but not the world state information. While accurate decisions are possible without the world state information, there are fewer decision points in our architecture and the human-robot team has a lower shared situational awareness.
2) **Observation**: Participants have access to the forecasted trajectory figures of merit *and* the world state information. Both information sets are presented as images for the user to review. As the information is presented passively, the user's situational awareness depends on their effort towards interpreting the figures.
3) **Interaction**: Participants have access to the robot's forecasted trajectory figures of merit *and* the world state information. Participants are required to judge each *world state information* aspect as *"Good"*, *"Maybe"*, or *"Bad"*. Additionally, participants answer a multiple-choice follow-up question such as "How are the weather conditions near the landing zone?" to gauge their situational awareness. We anticipate that participant situational awareness will improve from interacting with the information sources, compared to only observing them.

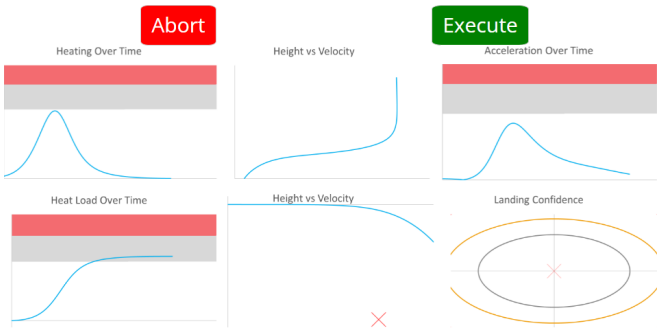

Fig. 3: Forecasted trajectory shown to users in *Stage 2*. The figures of merit are derived from the world state information source. Users are trained to interpret each figure (e.g., *"Heating over Time"* should not be in the red zone), and are told that the figures inform the robot's recommendation.

We hypothesize that a participant's world state information condition will impact their situational awareness – and therefore their mission performance – across Type I and Type II errors. We define *mission performance* as the accuracy between the participant's decision and the *ground truth* evaluation across the ten scenarios. We hypothesize:

H1 Participants in the *absent* condition will have a *lower* mission performance than participants in the *observation* condition or the *interaction* condition.

H2 Participants in the *interaction* condition will have a *higher* mission performance than participants in the *observation* condition.

C. Robot Error Types & Robot Accuracy

We are interested in how the robot's error type affects the participant's confidence and mission performance. Since the decision in our domain is a binary "Land" or "Abort", the robot can fail via the *false positive* case (recommending "Land" when the approach is high-risk) or the *false negative* case (recommending "Abort" when the approach is safe).

In *false positive* cases where the robot erroneously recommends a landing, we anticipate users will recognize risky world state aspects and confidently disagree with the robot.

However, in *false negative* cases we anticipate participants will second-guess themselves and try to reason for why the robot advised to abort. We anticipate that *mission performance* will be higher from the additional consideration.

To explore the error type attribute, participants are assigned to a *true* condition (where the robot is always correct), a *false positive* condition, or a *false negative* condition. In our study, the robot's recommendations have three *true positive* scenarios, three *true negative* scenarios, and four scenarios

of the participant's error type condition. Therefore the robot has a 60% accuracy in the two error types.

We hypothesize:

H3 The mission performance of participants in the *false negative* condition will be greater than the mission performance of participants in the *false positive* condition.

H4 Participants in the *false positive* condition will report higher rational trust than participants in the *false negative* condition.

III. Metrics

To evaluate how the *robot's recommendation* error type affects participants, we define several evaluation metrics.

Prior to the experiment, we record the participant's *dispositional trust*. We use the i-THAu trust assessment's [16] dispositional trust component.

During the experiment, we use participant decisions to determine several metrics of interest:

- *Initial SA*: The participant's initial understanding of the world, after viewing the *world state information* (**S1**). A participant's *Initial SA* is the percent accuracy between the participant's initial decision (*land* or *abort*) and the ground truth evaluation across the ten scenarios. Participants in the *absent* condition do not see *world state information*, so are not assessed an *Initial SA*.
- *Final SA*: The participant's final understanding of the world, after viewing the *figures of merit* of the forecasted trajectory information source (**S2**). Similar to *Initial SA*, *Final SA* is measured by the percent accuracy between the participant's decision at **S2** and the ground truth evaluation across the ten scenarios.
- *Initial Decision Agreement*: The percent agreement between the participant's decision at **S2** and the *robot's recommendation* (which may be incorrect). At this point the user has not yet seen the *robot's recommendation*. Since the *robot's recommendation* has a 60% accuracy in the Type I and Type II error types, an optimal user will have an *Initial Decision Agreement* of 60%.
- *Sway*: The participant's change of decision to the *robot's recommendation*. We measure a participant's *sway* as the number of scenarios that a participant changes their decision to the *robot's recommendation* between **S2** and **S3**. High *sway* can indicate overtrust in the robot, while low *sway* can indicate undertrust.
- *Mission Performance*: The percent agreement between the participant's final decision at **S3** and the ground truth evaluation across the ten scenarios, or the percentage of correct final decisions. A mission performance greater than 60% indicates that the participant disagreed with the *robot's recommendation* in favor of the correct decision. We use mission performance to help indicate a calibrated trust in the robot.

Post experiment, participants complete a final set of questionnaires to evaluate their workload and trust in the system:

- *Task Workload*: The task workload of the participant's condition, conducted via a NASA TLX survey [8].
- *Rational Trust*: The participant's trust in the robot after the experiment. We measure rational trust through the i-THAu trust assessment as a questionnaire.

From our metrics we can present insights into how participant decision-making is affected by the error type of black box recommender systems in human-robot teams.

IV. User Study Design

We implemented our study as a web-based application. 135 participants (Ages 19-63, 87 Female) were recruited through an online recruitment platform (Prolific). Our exclusion criteria included individuals who were under 18, located outside of the USA, not proficient in English, and/or did not have normal or corrected-to-normal vision. Participants were assigned randomly to one of the nine study conditions, with 15 participants per condition, and the order of scenarios shown to participants was counterbalanced. The experiment run time averaged 15-30 minutes, and participants were compensated $\$2.50 - \$5.00$. The study was IRB-approved.

V. Analysis & Results

We divide our analysis into analyses of objective metrics and subjective metrics. We meet all assumptions of and conduct ANOVAs as pre-hoc tests and evaluate the pairwise significance between key metrics (*Initial SA*, *Final SA*, *mission performance*, *sway*, *workload*, & *rational trust*) and our independent variables (*world state awareness* & *robot error type*).

A. World State Awareness vs. Situational Awareness

To verify our use of *world state awareness* to control participant situational awareness, we first assess the effect of the three levels of *world state awareness* on *Initial SA*, *Final SA*, and *mission performance* with a fixed "true" *robot error type* (the recommendation is always correct). A one-way ANOVA determined that world state awareness was only statistically significant for *Final SA* ($F(2,87) = 10.41$, $p = 0.0001$). Post-hoc pairwise t-tests found that the mean value of world state awareness was significantly different between *absent* and *observation* ($p = 6.2e\text{-}05$) and *absent* and *interaction* ($p = 0.00048$). There was no statistically significant difference between *observation* and *interaction* ($p = 0.56008$).

Participants in the *absent* world state awareness condition averaged a *Final SA* of 79.3% (SD=11.9%), indicating that the *forecasted trajectory* information source provided enough information for participants to conclude correct decisions.

Per the post-hoc pairwise comparisons, we found no improvement to *Final SA* or *Mission Performance* between the *observation* and *interaction* conditions. This indicates that requiring interactions with the world state information source did not improve user situational awareness beyond the *observation* condition.

Additionally, no differences were found in the *mission performance* between all three *world state awareness* conditions. These results align with expectations that the world state information improves participant situational awareness,

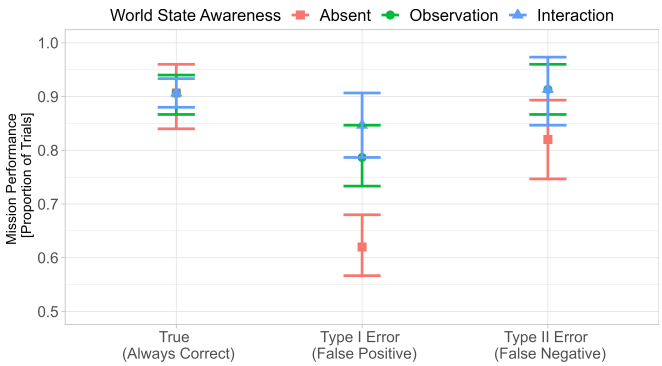

Fig. 4: Plot of the *mission performance* of each *world state awareness* and *robot error type* independent variable. Error bars represent standard error.

however does not affect participants selecting the correct decision after seeing the robot's recommendation.

From our findings we conclude that *world state awareness* can be used to control for two levels of participant situational awareness.

B. World State Awareness vs. Mission Performance

The next two sections assess how *world state awareness* affects *mission performance*, and how *mission performance* is affected by the *robot's error type* across the levels of *world state awareness*. The outcomes are visualized by Fig. 4.

A two-way ANOVA (IVs: *world state awareness*, *robot's error type*, DV: *mission performance*) reveals that there was a statistically significant interaction between the effects of *world state awareness* and *robot's error type* ($F(4, 124) = 4.190$, $p = 0.0032$). Simple main effects analyses showed that both *world state awareness* ($p = 0.0002$) and *robot's error type* ($p < 0.0001$) had statistically significant effects on *mission performance*.

Through pairwise t-tests, we found that *world state awareness* significantly affects *mission performance* across Type I and Type II *robot error types* ($p < 0.01$). Participants in the *absent* level had a significantly lower *mission performance* compared to participants in the *observation* and *interaction* levels ($p < 0.01$). We therefore affirm hypothesis **H1**.

However, participants in the *true* error type (robot's recommendation is always correct) saw no performance difference between *world state awareness* levels. This can be attributed to participants recognizing the infallibility of the robot.

The *mission performance* between the *interaction* level and the *observation* level did not differ between conditions. While we found significance in the Type I error level, we are hesitant to claim the difference is meaningful. We therefore refute hypothesis **H2**.

C. Mission Performance vs. Robot Error Type

The robot's error type has a significant effect on a participant's *mission performance*, with participants performing significantly higher in the Type II error type (the robot recommends aborting when it is safe) than in the Type I error type (the robot recommends landing when it is too risky). The finding holds across all *world state awareness* levels ($p < 0.01$ for all). We therefore affirm hypothesis **H3**.

We suspect the performance difference relates to the working memory required for each condition. In the *false positive* condition, participants can filter out information as "safe" except for risky figures of merit. However, in the *false negative* condition participants may be less confident that enough "risky" figures of merit are present, and may be swayed by the robot's incorrect recommendation.

No significant difference is found between *world state awareness* levels in the *true* error type. However, *mission performance* is already close to 100%. This indicates that participants in the *absent* condition were able to obtain the correct decision through their evaluation of the *forecasted trajectory* figures of merit and the robot's decision.

D. Sway vs. Robot Error Types

We next assess the impact of the *robot error type* on swaying participants to the robot's decision after seeing the *robot's recommendation*. To evaluate sway, we looked at the four scenarios unique to each error type condition, i.e., the scenarios where the robot made an incorrect decision.

A two-way ANOVA revealed that there was not a statistically significant interaction between the effects of *world state awareness* and *robot's error type* ($F(4, 124) = 0.96866$, $p = 0.4273$) for this metric. Simple main effects analysis showed that *robot's error type* ($p = 0.0181$) had a statistically significant effect on whether the participant was swayed by the robot's recommendation, while *world state awareness* did not ($p = 0.0818$).

From a post-hoc pairwise t-test, we found participants in the Type I error type (*false positive*) swayed significantly more than participants in the Type II error type ($p < 0.01$ for all world state awareness levels). This aligns with our *mission performance* findings, indicating that participants were more likely to erroneously trust the robot's recommendation in the Type I error type.

E. Rational Trust vs. Robot Error Type

We subjectively measured two factors of the user's experience: *workload*, and *rational trust* via post-experiment questionnaires. We found no significant difference in *workload* across our nine study conditions.

A pairwise t-test (pooled across *world state awareness* conditions) found that participants reported lower *rational trust* in the Type II error levels (*false negative*) compared to the Type I ($p = 0.01$) or *true* ($p < 0.01$) error levels. Participants in the *true* error level also reported significantly higher *rational trust* than participants in the Type I error level ($p = 0.02$). We thus affirm hypothesis **H4**.

Our findings show a trade-off between *rational trust* and *mission performance* – while participants were more likely to override the erroneous robot recommendation in the Type II error condition, they did so with a reduced trust in the system. Alternatively, participants in the Type I error condition were more likely to be swayed by the recommendation while ascribing greater rational trust in the system.

VI. Discussion

In this paper we demonstrate that the error type of a robot's recommendation matters. In practice, recommender system designers adjust Type I and Type II error by changing the confidence threshold for a positive decision – a higher threshold reduces Type I error (*false positive*) and increases Type II error (*false negative*).

Our experiment found that *mission performance* is significantly more reactive to *false negative* recommendations (Type II error) than to *false positive* recommendations (Type I error). However, while *false negative* recommendations led to greater *mission performance*, they resulted in a reduced user *rational trust* in the system. The findings were consistent across two levels of user situational awareness, as controlled by our *world state awareness* independent variable.

Our findings suggest that users outperform in *false negative* situations, indicating that it is preferable for system designers to prioritize *false negative* errors when *mission performance* is the driving factor. Due to the generic nature of our study design, we anticipate that our findings will apply to other virtual decision-making domains. However, an important distinction should be made between virtual and physical human-robot interactions – the physical presence of a robot may be more convincing than a virtual prompt.

Therefore we recommend that system designers choosing between reducing Type I or Type II error consider the two error types independently, and evaluate the trade-off between *mission performance* and *rational trust*.

In addition, our analysis of *world state awareness* found that requiring users to interact with the *world state information* did not affect their *mission performance* across Type I and Type II error levels. We recommend that system designers be mindful of the extent to which their systems actively engage with users. While enhancing user situational awareness can improve *mission performance* and user resilience to robot errors (as shown in the mission performance difference between the *absent* and *observation* levels), we found that additional interaction did not significantly enhance situational awareness. Therefore, system designers should consider the trade-off between the fluency of their human-robot interface and the benefits to mission outcomes.

Our work presents opportunities for future work. While our domain used a binary recommendation (*Land*, *Abort*), methods from the explainable AI community can present the reasoning behind the recommendation to improve the user experience and *mission performance*. Additionally, we are interested in how our findings hold across other human-robot teaming domains, in particular domains with physical robots. Future work can adapt our methods to novel environments and situational contexts.

References

[1] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, Connor T Esterwood, X Jessie Yang, Lionel P Robert Jr, and Dawn M Tilbury. Comparing the effects of false alarms and misses on humans' trust in (semi) autonomous vehicles. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 113–115, 2020.

[2] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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):1–21, 2021.

[3] Devleena Das, Been Kim, and Sonia Chernova. Subgoal-based explanations for unreliable intelligent decision support systems. *arXiv preprint arXiv:2201.04204*, 2022.

[4] Jinyue Feng, Chantal Shaib, and Frank Rudzicz. Explainable clinical decision support from text. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1478–1489, 2020.

[5] Gavan J Fitzsimons and Donald R Lehmann. Reactance to recommendations: When unsolicited advice yields contrary responses. *Marketing Science*, 23(1):82–94, 2004.

[6] Felix Gervits, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz. Toward genuine robot teammates: Improving human-robot team performance using robot shared mental models. In *Aamas*, pages 429–437, 2020.

[7] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[8] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[9] Michail Karakikes and Dimitris Nathanael. The effect of cognitive workload on decision authority assignment in human–robot collaboration. *Cognition, Technology & Work*, pages 1–13, 2022.

[10] Gregory E. Kersten and Stan Szpakowicz. Decision making and decision aiding. defining the process, its representations, and support. *Group Decision and Negotiation*, pages 237–261, 1994.

[11] Jack Kolb, Harish Ravichandar, and Sonia Chernova. Leveraging cognitive states in human-robot teaming. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 792–799. IEEE, 2022.

[12] Yixiang Lim, Nichakorn Pongsakornsathien, Alessandro Gardi, Roberto Sabatini, Trevor Kistan, Neta Ezer, and Daniel J Bursch. Adaptive human-robot interactions for multiple unmanned aerial vehicles. *Robotics*, 10(1):12, 2021.

[13] Yizhi Liu, Mahmoud Habibnezhad, and Houtan Jebelli. Brainwave-driven human-robot collaboration in construction. *Automation in Construction*, 124:103556, 2021.

[14] Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for human-ai collaboration. *Plos one*, 15(2):e0229132, 2020.

[15] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–15, 2019.

[16] Yosef Razin. Interdependent trust for humans and automation survey. Available at `https://sites.gatech.edu/feigh-lab/publications/` (2022).

[17] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 101–108. IEEE, 2016.

[18] H.A. Simon and C.I. Barnard. *Administrative Behavior: A Study of Decision-making Processes in Administrative Organization*. Free Press paperback. Free Press, 1976.

[19] Aaquib Tabrez and Bradley Hayes. Improving human-robot interaction through explainable reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 751–753. IEEE, 2019.

[20] Alan R Wagner, Jason Borenstein, and Ayanna Howard. Overtrust in the robotic age. *Communications of the ACM*, 61(9):22–24, 2018.

[21] Sarah E Walsh, William Sealy, and Karen M Feigh. Optimal experimental design methods for acquiring and restricting information to improve decision making. In *Advances in Neuroergonomics and Cognitive Engineering: Proceedings of the AHFE 2020 Virtual Conferences on Neuroergonomics and Cognitive Engineering, and Industrial Cognitive Ergonomics and Engineering Psychology, July 16-20, 2020, USA*, pages 290–297. Springer, 2021.

[22] Holly A Yanco and Jill Drury. Classifying human-robot interaction: an updated taxonomy. In *2004 IEEE international conference on systems, man and cybernetics (IEEE Cat. No. 04CH37583)*, volume 3, pages 2841–2846. IEEE, 2004.