Patterns of Effective Human-Agent Teams

Kazuhiko Momose
kmomose@my.fit.edu
Human-Centered Design, L3Harris Institute for Assured Information
Florida Institute of Technology
Melbourne, FL, USA

Troy R Weekes
tweekes@fit.edu
Human-Centered Design, L3Harris Institute for Assured Information
Florida Institute of Technology
Melbourne, FL, USA

Rahul Mehta
rmehta2022@my.fit.edu
Computer Science, L3Harris Institute for Assured Information
Florida Institute of Technology
Melbourne, FL, USA

Cameron Wright
cameron2018@my.fit.edu
Computer Science, L3Harris Institute for Assured Information
Florida Institute of Technology
Melbourne, FL, USA

Josias Moukpe
jmoukpe2016@my.fit.edu
Computer Science, L3Harris Institute for Assured Information
Florida Institute of Technology
Melbourne, FL, USA

Thomas C Eskridge
teskridge@fit.edu
Computer Science, Human-Centered Design, L3Harris Institute for Assured Information
Florida Institute of Technology
Melbourne, FL, USA

ABSTRACT

Intelligent systems are increasingly interacting with people, both in their daily lives and through their use in safety critical systems. Current research is focused on how to use intelligent systems in a collaborative way as a teammate, rather than a tool. This requires a better understanding of what behaviors enable effective human-agent teams. This paper reports an experiment where a human player collaborates with an agent to perform a maneuvering task while concurrently performing a memory task. The player must determine in which contexts the agent requires their input to achieve better combined game plus memory task accuracy scores. We hypothesized that high-performing teams would exhibit different patterns of control inputs when compared to low-performing teams and that these patterns of control would be made more evident with user interfaces that increased operator situation awareness. Preliminary results are inconclusive, but show different patterns of interaction between high- and low-performing teams.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Human-agent teamwork, patterns of interactions, user interface design

ACM Reference Format:
Kazuhiko Momose, Troy R Weekes, Rahul Mehta, Cameron Wright, Josias Moukpe, and Thomas C Eskridge. 2023. Patterns of Effective Human-Agent Teams. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3544549.3585608

1 INTRODUCTION

There are an increasing number of sophisticated automation systems used in contexts ranging from our daily lives to safety critical systems. Although such sophisticated machines are expected to enrich lives, there have been some undesired incidents that result from the interaction of automation with the automation user (e.g., [22, 24]). The Human-Computer Interaction (HCI) and other research communities have advocated a focus on how users could collaborate with such sophisticated automation systems as a teammate [5, 9, 19].

Current automation systems are leveraging both handoff and composed input collaboration with user to overcome both performance and regulatory hurdles [7, 17]. The inputs from the user are integrated with those from automation so that the handover of control authority can be made quickly or respect user guidance.

In human-human team contexts, Shared Cooperative Activity [1] describes ways in which teammates cooperate to achieve a shared goal. In particular, the *cooperatively neutral* approach describes the case where the teammates have the capability to perform the task on their own but may choose to work together. In this study, a test environment is used to model a shared, cooperatively neutral activity between a user and an intelligent automation system to determine what enables effective teamwork and how interaction between teammates can be improved.

2 BACKGROUND AND RELATED WORK

With the advancement of machine capabilities, including Artificial Intelligence (AI)-powered systems, the research community has shifted the focus from automation to autonomy [21]. Whereas automation executes well-defined tasks and requires humans to play a supervisory role, key features of autonomy include self-governance, adaptability, and learning, generating the expectation that autonomous agents work together with humans as teammates [21, 23, 30]. As a means of facilitating teamwork between humans and agents, Johnson et al. [18] introduced coactive design, identifying interdependence between human and agents and emphasizing mutual observability, predictability, and directability (OPD). Cila [5] identified design considerations for effective human-autonomy

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585608

teaming, such as intelligibility, task delegation, and when to release or retain agent autonomy.

The simple mechanics and controllability of microworld domains enable researchers to focus on key aspects of teamwork, and empirical studies using them have produced insights into the dynamics of human-agent teamwork (e.g., [2, 4, 6, 25]). The research community has also applied knowledge from the literature of human-human teams to human-agent teamwork, including social exchange theory [3, 4], and implicit communication [2, 20] to name a few. For instance, Chiou et al. [3] conducted a human-robot teaming study employing the notion of primacy effect, indicating that more team communication does not necessarily lead to better team performance, and the information exchange strategies (e.g., which entity pushes or pulls information) could contribute to improving human-robot teamwork. Understanding such patterns of effective teams will enable the definition of best practices of User Interface (UI) and interaction designs for enhancing team performance.

3 EXPERIMENT

3.1 Objective

This experiment was designed to provide a baseline for investigating effective teams in a compositional control setting, where all team members have the same action channels, and the effect on the system is dependent on a combination of inputs from all entities. The lunar lander game was chosen as a microworld for teamwork experimentation because it is cooperatively neutral, where the human and agent can, but are not *required*, to work together to successfully land the lunar lander. The game provides the basic infrastructure necessary to investigate a wide range of teamwork issues, focusing in this experiment on identifying how the patterns of interactions between human and agent differ between high- and low-performing teams. In this experiment, we asked three key research questions in this study:

- RQ1 (Agent Capability): Is there a distinction between teams with a more capable agent and team with a less capable agent?
- RQ2 (UI Design): Does UI design change team interactions and augment teamwork?
- RQ3 (Patterns of Interactions): Do high-performing teams have different patterns of interactions from low-performing teams?

3.2 Participants

A power analysis was carried out using G*Power (version 3.1.9.7) [10], suggesting a required sample size of 211 (effect size: 0.25, α: 0.05, power: 0.80, number of groups: 6, number of measurements: 2). With a 10% margin, we aimed to achieve the sample size of 240 and assign 40 participants to each group. Participants were recruited via a convenience sampling, the university's information forum, and announcements in classes.

3.3 Experiment Setting

The browser-based game (Figure 1)[1] was employed to investigate human-agent teamwork in a compositional control setting. Participants were asked to conduct a moon lander maneuvering task in concert with an agent while conducting a memory task concurrently.

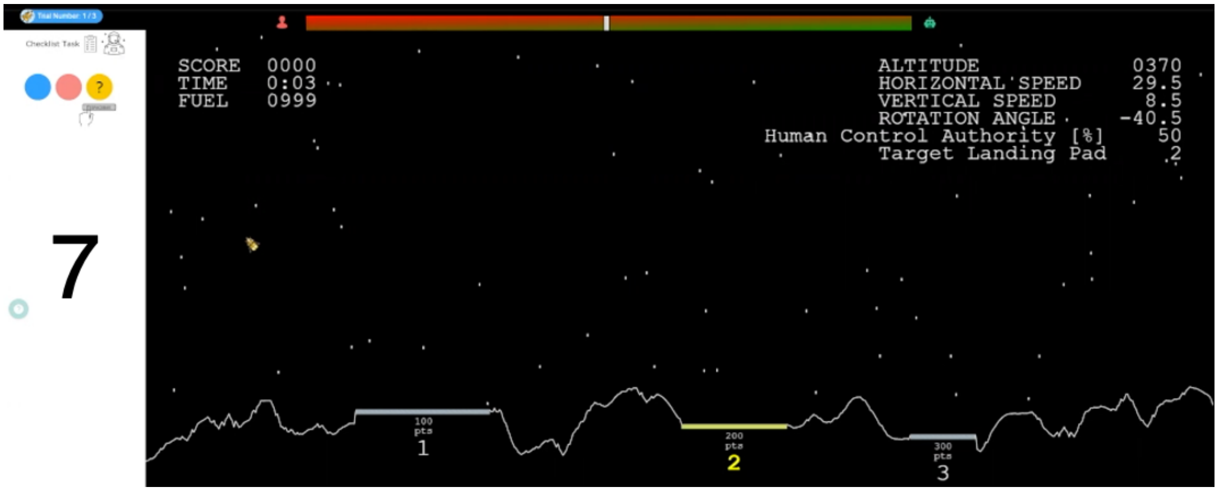
3.3.1 Moon Lander Maneuvering Task.

3.3.1 Moon Lander Maneuvering Task. The participants were asked to safely land the moon lander on their choice of three landing pads working in concert with an agent teammate. The player used the up arrow key to engage a thruster and change the lander speed and altitude, and the right & left arrow keys to rotate the lander (see Appendix A.1). The initial amount of fuel was set at 1000 in each trial, and the fuel was consumed only when the thruster was engaged. The agent teammate was also able to engage the thruster and rotate the lander. There was a control authority indicator on the top center of the screen showing the control authority ratio between the player and the agent, which was inspired by an interaction strategy called "Horse-Mode" or H-Mode [12]. We included the indicator in the game screen to investigate when and to which extent the player explicitly grants or retrieves control authority. The ratio was used to determine how to consolidate the inputs from the player and agent (see Appendix A.2). In the beginning of each trial, the player selected one of three target landing pads by pressing the number 1, 2, or 3 key so that the agent teammate could assist with the maneuvering task. The size and location of each pad was computed based on an increasing Index of Difficulty (*ID*), which corresponded to the Fitts' Law paradigm [11] (see Appendix A.3), and provided 100, 200, or 300 points upon successful landing.

3.3.2 Cognitive Load. While the team was working on the moon lander maneuvering task, the player was asked to simultaneously perform a memory task that simulates common secondary flight tasks, such as checklist management. We employed the n-back task to divert the player's attention from the maneuvering task, requiring the player to allocate cognitive resources to both tasks. In the n-back task, a sequence of digits is presented one-by-one on the left side of the screen, and the participants are asked to hit the space-bar when they see a target stimulus. We employed a non-audio version of the n-back task with 500 ms of the stimulus presentation and 2500 ms of the blank period. Additionally, we presented a target stimulus with a 25% chance to ensure that participants who did not conduct the n-back task would receive penalties. In the experiment, we computed the accuracy of the n-back task (*Accuracy*) by $\frac{H+CR}{H+M+FA+CR}$ where H, CR, M, and FA are Hits, Correction Rejections, Misses, and False Alarms respectively.

3.3.3 Game Performance. The participants were instructed that both the moon lander maneuvering task and the n-back task contributed to the overall game scores. This instruction was made to avoid a situation where the participants did not pay attention to the n-back task. In each trial, the team gained *GameScores* that were calculated by $Landing\ Points + (Remaining\ Fuel/Overlap\ Factor) \times Accuracy$ where $Overlap\ Factor$ determines the degree to which the teams can achieve the same game scores in different ways. We selected the $Overlap\ Factor$ value of 5 (see Appendix C). For

[1] The game was built on [26], and some of the icons used in the game were downloaded from https://icons8.com and https://imgbin.com/ [accessed on 03/13/2023]



```
SCORE  0000
TIME   0:03
FUEL   0999
                              ALTITUDE              0370
                              HORIZONTAL SPEED      29.5
                              VERTICAL SPEED         8.5
                              ROTATION ANGLE       -40.5
                    Human Control Authority [%]       50
                              Target Landing Pad       2
```

Figure 1: Moon lander game: a human player was asked to conduct the moon lander maneuvering task (right) in concert with an agent teammate while the player also performed a memory task (left).

a failed trial, the team did not receive any *Landing Points* and *Remaining Fuel* points, resulting in zero game scores.

3.4 Experiment Design

With the experiment setting, we employed a 2 (agent capability) × 3 (UI design) × 2 (n-back task difficulty) mixed design, where between-subject factors include agent types and UI designs, and the n-back task difficulty served as a within-subject factor.

3.4.1 Agent Capability. In this study, we assigned each participant to either the more or less capable agents whose actions were coded based on heuristics of the moon lander maneuvering task (see Appendix D). By tweaking some thresholds of final approach behaviors, the more and less capable agents were designed to exhibit approximately 80% and 50% chance of successful landings respectively.

3.4.2 UI Design. In the moon lander maneuvering task, the participants were assigned to one of the three UI design conditions: (i) baseline, (ii) trajectory projection, and (iii) trajectory projection with a vertical speed indicator (see Appendix E). The baseline UI was a text-only Heads-Up Display (HUD) at the top of the screen showing flight-related information. The trajectory projection displayed a predicted trajectory with a white arc in addition to the HUD display, allowing the player to anticipate the lander’s future position. The third UI design had the same features as the trajectory projection except for during the final approach phase. The third UI condition presented thrust recommendations based on the vertical speed in a visual fashion with a color-coded vertical speed scale. The design makes it easier to land successfully by focusing attention on the vertical speed in close proximity to the lander without having to read the text of the HUD at the top of the UI.

3.4.3 N-Back Difficulty. There were two levels of the n-back task (Appendix B): the 0-back task (easy) and the 2-back task (hard). In the 0-back task, the participants were told a target digit before each trial and needed to press the space-bar once the target digit appeared and before the next digit appeared. In the 2-back task, the participants were not told a target digit before each trial. Instead, they were asked to hit the space-bar once they saw the same digit presented in two previous steps, requiring them to memorize the previously presented digits.

3.5 Experiment Procedure

The experiment was carried out remotely using the browser-based moon lander game, and the length of the experiment was approximately 45 minutes. This study was reviewed and approved by the university’s Institutional Review Board (IRB Number 22-114). In the beginning of the experiment, the participants signed an informed consent form and filled out a demographic questionnaire (Appendix F.1). Then, a familiarization session began, where they watched short tutorial videos and conducted familiarization trials of the moon lander maneuvering task and the n-back task first individually and then concurrently. The participants were randomly assigned to one of the six groups (i.e., 2 agent capability levels for the 3 UI designs). During the familiarization session, all the participants interacted with a familiarization agent that exhibited better performance than the more and less capable heuristic agents, and were instructed that the familiarization agent was only used during the familiarization session. The familiarization session was followed by two blocks with 10 trials each of the moon landing task. In each block, the participants collaborated with the assigned agent on the moon maneuvering task while also performing the 0-back or 2-back task. The order of the n-back task difficulty was randomized. After each block, the participants filled out workload and agent interaction questionnaires. On completion of the two blocks, the participants signed out from the online experiment.

3.6 Measures and Data Analysis

As objective measures, we recorded the game scores, explicit control authority changes, and human-agent keystroke ratios. As subjective measures, we administered two questionnaires: NASA-Raw Task

Load Index (RTLX) (see Appendix F.2) [14, 15] and an agent interaction questionnaire (see Appendix F.3). NASA-RTLX was used to capture participants' perceived workload and administered after each block. The agent interaction questionnaire was also administered after each block, asking the participants about: (Q1) agent's predictability, (Q2) agent's performance, (Q3) whether they felt the agent needed their help, and (Q4) whether they felt their help was effective.

Using R (version 4.2.2) [27], a three-way repeated measures MANOVA test was carried out for the game scores, NASA-RTLX scores, and agent interaction ratings. We used the MANOVA.RM package [13] and reported the modified ANOVA type statistic (MATS). When appropriate, univariate follow-up analyses and post-hoc pairwise comparisons were performed applying the Aligned Rank Transform (ART) [29] and the ART-Contrasts (ART-C)[8]. The alpha level was set at 0.05. There were five main hypotheses as follows; H5 was investigated through an exploratory analysis:

- H1 (Agent Capability | Objective): Participants who interact with the more capable heuristic agent will exhibit higher game scores when compared to participants in the less capable heuristic agent group
- H2 (Agent Capability | Subjective): Participants will report the same level of agent's predictability while there will be a difference in perceived agent's performance level between the more and less capable heuristic agents
- H3 (UI Design | Performance): The trajectory projection with the vertical speed indicator will exhibit the best teams' game scores and be followed by the trajectory projection and the baseline condition
- H4 (UI Design | Workload): The trajectory projection with the vertical velocity indicator will require the least amount of workload and be followed by the trajectory projection and the baseline in terms of game scores
- H5 (Patterns of Interactions): High-performing teams will show different patterns of control authority changes and keystrokes from low-performing teams.

4 RESULTS

A total of 60 participants aged from 17 to 61 years (*M*: 26.5 and *SD*: 10.3) completed the experiment session, resulting in a power of 0.44, significantly less than the target 0.8. The summary of the demographic questionnaire appears in Appendix F.1. Figure 2 shows averaged trial game scores across all the conditions in one block. Figures 3 and 4 present NASA-RTLX scores and participants-reported Likert-scale ratings respectively. The MANOVA only detected a significant effect of the agent types ($MATS = 50.0$, $p < 0.01$). Post-hoc analyses indicated a significant difference in the game scores ($p < 0.01$), the ratings of Q1 ($p < 0.01$), Q2 ($p < 0.01$), and Q3 ($p < 0.01$) between the more and less capable agents. Therefore, the results supported H1, and H2 was partially supported (i.e., the participants reported different level of perceived agent's performance but also predictability). H3 and H4 were not confirmed as the MANOVA did not detect a significant difference in the UI types ($MATS = 19.8$, $p = 0.377$). As part of an exploratory data analysis, we computed the human key input ratio as follows: first, we normalized the length of each trial. Then, we computed the human key input ratio by dividing the number of human keystrokes by the total number of keystrokes from the player and the agent per one normalized time step. Figure 5 shows the profiles of human key input ratios exhibited by the top three and bottom three players in the course of a trial. Note that these profiles indicate that high-performing teams choose different patterns of interaction than low-performing teams when working with the same type of agent.

5 DISCUSSION

5.1 Agent Capability: H1 and H2

The significant difference in the agents' performance levels was confirmed via the game scores, supporting H1. In terms of H2, the participants reported different level of perceived performance (Q2) but also predictability (Q1). As both agents were designed in a rule-based, threshold-oriented manner, we expected the same level of predictability of agents' actions. As part of our exploratory analysis, we computed the number of human key inputs (see Appendix G), showing a trend that the players with the less capable agent tended to provide more key inputs. This may be because they considered the less capable agent less predictable and wanted to have more dominant control of the lander, which is supported by Q3 (i.e., the players with the less capable agent were more likely to feel the agent needed their help). We did not find a significant difference in the responses to Q4. The majority of the participants conducted the maneuvering task with the 50%-50% control ratio during the experiment, meaning that the magnitude of their inputs was reduced by half. This may have made it difficult for them to feel the impact of their input on the lander movement regardless of the agent capability.

5.2 UI Design: H3 and H4

The results did not suggest a significant difference between the three UI designs, contrary to H3 and H4. Interestingly, the trajectory projection with vertical speed indicator exhibited slightly lower average game scores in the 0-back task condition with the more capable agent when compared to the other two UI designs. This may be explained if the inclusion of the vertical speed indicator enabled the player to participate in the maneuvering task more actively, resulting in more conflict inputs between the player and the more capable agent. Also, we assumed that the participants were prone to over-trust the more capable agent with the baseline UI design, leading to higher scores. In further exploratory analysis, we computed the number of conflicts between the player and agent (see Appendix G), which does not appear to confirm the two assumptions. Therefore, it is still unclear as to the effects of the UI designs.

We did not detect any significant difference between the 0-back and 2-back tasks although subjective feedback indicated participants found the workload to be different. In this study, we attempted to measure human-agent teamwork employing the game scores computed based on the maneuvering task performance and n-back task accuracy, and the participants were instructed to maximize the game scores. Although we followed our prescribed analysis approach in this paper, the results seem to suggest that we should analyze the maneuvering task performance (i.e., landing points and

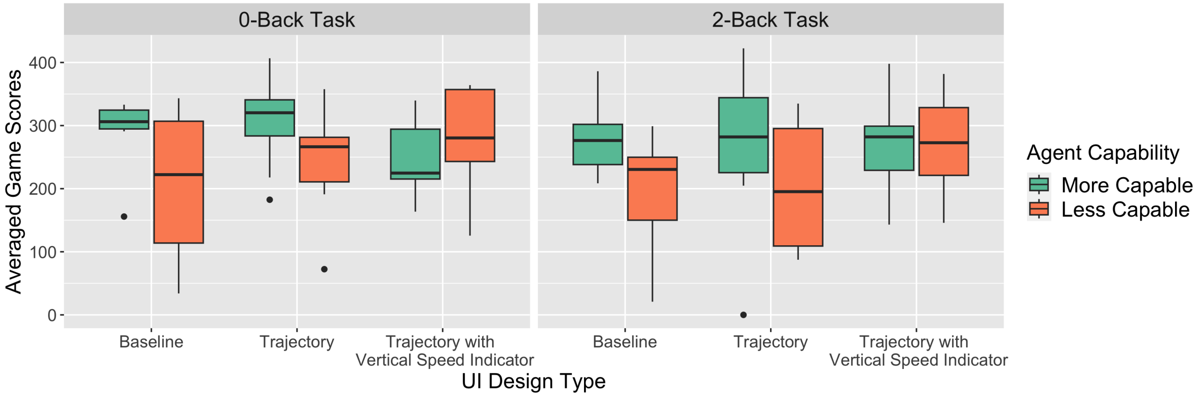


Figure 2: Averaged trial game scores in one block across all the conditions

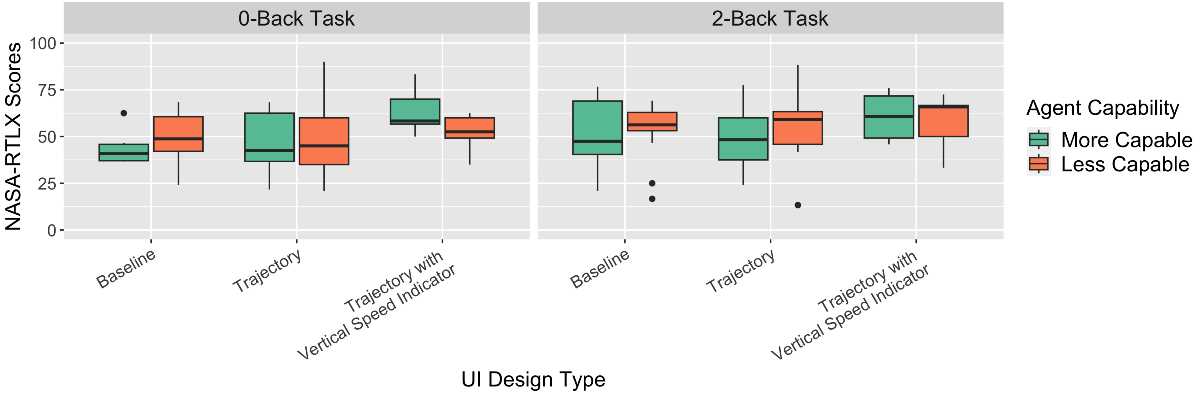


Figure 3: NASA-RTLX scores across all the conditions; no significant differences were detected in this experiment.

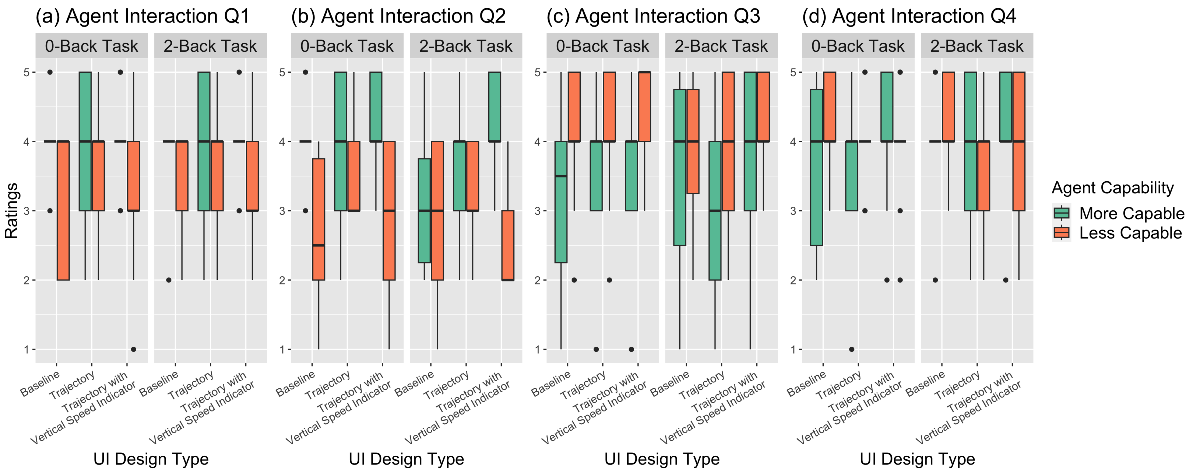


Figure 4: Responses to the agent interaction questionnaire on a Likert-scale (1: Strongly Disagree - 5: Strongly Agree); in this set of game: Q1 the agent’s actions were very predictable, Q2 the agent was performing the landing task very well, Q3 I felt that the agent needed my help for the lander control, and Q4 I felt that my help was effective.

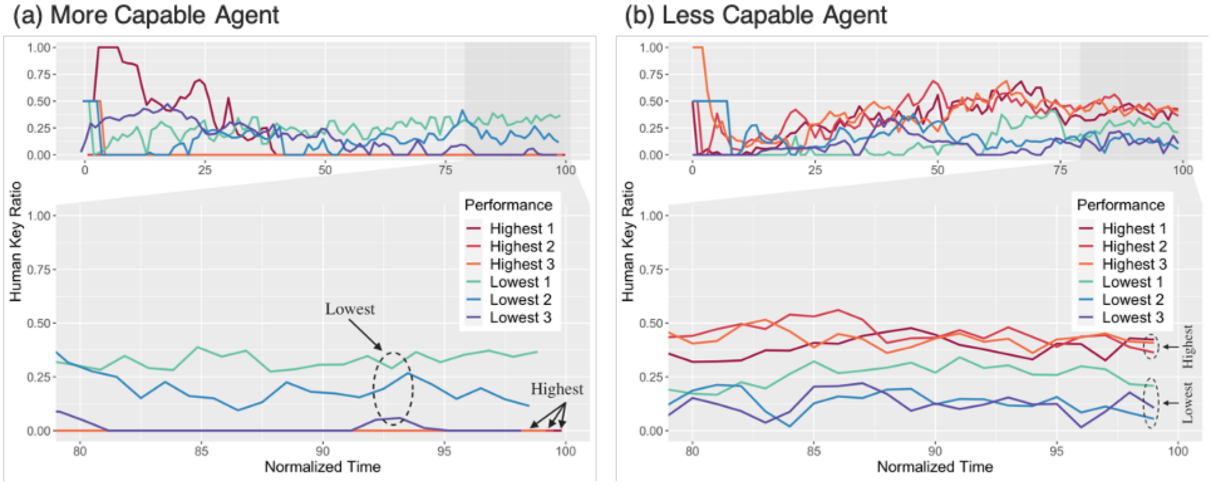


Figure 5: Normalized time vs. human key input ratio; we extracted the three highest scores and the three lowest scores across the two agent types. (a) when working with the more capable agent, fewer human key inputs during the second half of a trial led to higher game scores; note that we depicted each line by slightly shifting it in a x-axis direction for better readability (b) the interaction trend was flipped when working with the less capable agent, meaning more human key inputs result in higher scores.

remaining fuel) and the n-back task performance individually, or further increase the difficulty of the task.

5.3 Patterns of Interactions: H5

We expected different patterns of explicit control authority changes between high- and low-performing teams. However, we observed less frequent control authority changes than expected in the entire dataset; 246 trials exhibited at least one control authority change out of 1200 trials. Although we provided instructions as to how to change the control authority during the familiarization session and reminded the players of the corresponding keys between trials, the players might have been prone to forget to explicitly change the control authority levels due to their workload and ability to "work around" the explicit setting with increased or decreased keystrokes.

The lack of explicit changes to control authority via the UI indicator prompted exploratory analysis of the profiles of human key inputs measured by the ratio of human key inputs to agent key inputs. We arbitrarily chose to examine the three highest and lowest scoring players to compare interaction patterns. Figure 5 (a) shows that the high-performing players tended to provide few inputs in the second half of a trial when working with the more capable agent. In contrast, Figure 5 (b) indicates the opposite interaction pattern where more human key inputs led to higher game scores when working with the less capable agent. Figure 5 appears promising to further investigate H5 for our future work. With a solid understanding of patterns of effective human-agent interactions, we could focus on how to augment low-performing teams' performance by prompting them to exhibit the same interaction patterns of the effective human-agent teams. In future work, we will analyze the patterns of all players to determine the natural separations between high, low, and average scoring players and examine their implications for improving teamwork.

5.4 Future Work: More Sophisticated Teamwork Setting

The three UIs tested in this experiment were designed to merely display the game environment information (i.e., the lander physics); there were no bidirectional information exchanges about each entity's intent or reasoning between the player and the agent. Investigating a more sophisticated human-agent teamwork setting where teammates are more transparent, provide explanation for decisions, and adapt to coordinated behavior are promising directions for future research (see Appendix H). For instance, we could allow the agent to also select a target landing pad and examine how the team resolves a conflict situation where the player's selection differs. In this scenario, the agent could explain its rationale by conveying the relevant information supporting the decision (e.g. the remaining fuel and distance).

6 CONCLUSION

More autonomous agents are expected to work with humans as teammates, which poses a question as to how to formulate effective human-agent teams. We hypothesized that effective human-agent teams exhibit different patterns of interactions from low-performing teams, and UI designs would augment humans' situation awareness helping to achieve a better level of performance. In this preliminary attempt, we conducted an experiment using the moon lander game environment where the human player worked on the lander maneuvering task in concert with the agent while simultaneously performing the memory task. We examined the effects of the agent capability and the UI designs on the overall team performance in the different memory task difficulty levels. Although the results suggested that agent capability affected the overall game scores, we did not reveal any significant differences between the UI designs, which is not in line with our hypothesis or expectations. Yet, our exploratory data analysis provided us with some implications for understanding patterns of effective human-agent teams. We observed the different trends of human key input ratio between high- and low-performing teams in both more and less capable agent conditions.

Our future work continues to examine whether high-performing teams exhibit different patterns of interactions from low-performing teams. We believe that understanding such patterns could help us to establish design principles for enhancing human-agent teamwork. Built on the design considerations [5], we discussed potential future work. We believe that the experiment reported here identifies key issues necessary to understand and build effective human-agent teams.

ACKNOWLEDGMENTS

The authors would like to thank Mr. Daniel Kidwell, visiting scholar from DoD with expertise in cognitive systems engineering and visual design, for his valuable feedback on this study.

REFERENCES

[1] Michael E Bratman. 1992. Shared cooperative activity. *The philosophical review* 101, 2 (1992), 327–341.

[2] Abhizna Butchibabu, Christopher Sparano-Huiban, Liz Sonenberg, and Julie Shah. 2016. Implicit coordination strategies for effective team communication. *Human factors* 58, 4 (2016), 595–610.

[3] Erin K Chiou, Mustafa Demir, Verica Buchanan, Christopher C Corral, Mica R Endsley, Glenn J Lematta, Nancy J Cooke, and Nathan J McNeese. 2021. Towards Human–Robot Teaming: Tradeoffs of Explanation-Based Communication Strategies in a Virtual Search and Rescue Task. *International Journal of Social Robotics* (2021), 1–20.

[4] Erin K Chiou, John D Lee, and Tianshuo Su. 2019. Negotiated and reciprocal exchange structures in human-agent cooperation. *Computers in Human Behavior* 90 (2019), 288–297.

[5] Nazli Cila. 2022. Designing Human-Agent Collaborations: Commitment, responsiveness, and support. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[6] Sylvain Daronnat, Leif Azzopardi, Martin Halvey, and Mateusz Dubiel. 2020. Impact of agent reliability and predictability on trust in real time human-agent collaboration. In *Proceedings of the 8th International Conference on Human-Agent Interaction*. 131–139.

[7] Dániel András Drexler, Árpád Takács, Tamás Dániel Nagy, Péter Galambos, Imre J. Rudas, and Tamás Haidegger. 2019. Situation Awareness and System Trust Affecting Handover Processes in Self-Driving Cars up to Level 3 Autonomy. In *2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*. 000179–000184. https://doi.org/10.1109/IWOBI47054.2019.9114533

[8] Lisa A Elkin, Matthew Kay, James J Higgins, and Jacob O Wobbrock. 2021. An aligned rank transform procedure for multifactor contrast tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 754–768.

[9] Mica R Endsley. 2022. Supporting Human-AI Teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior* (2022), 107574.

[10] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[11] Paul M Fitts. 1951. Human engineering for an effective air-navigation and traffic-control system. (1951).

[12] Frank Flemisch, Johann Kelsch, Christan Löper, Anna Schieben, Julian Schindler, and Matthias Heesen. 2008. Cooperative control and active interfaces for vehicle assitsance and automation. (2008).

[13] Sarah Friedrich, Frank Konietschke, and Markus Pauly. 2019. Resampling-based analysis of multivariate data and repeated measures designs with the R package MANOVA. RM. *R J.* 11, 2 (2019), 380.

[14] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[15] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[16] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[17] SAE International. 2021. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles J3016_202104*.

[18] Matthew Johnson, Jeffrey M Bradshaw, Paul J Feltovich, Catholijn M Jonker, M Birna Van Riemsdijk, and Maarten Sierhuis. 2014. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction* 3, 1 (2014), 43–69.

[19] Matthew Johnson and Alonso Vera. 2019. No AI is an island: the case for teaming intelligence. *AI magazine* 40, 1 (2019), 16–28.

[20] Claire Liang, Julia Proft, Erik Andersen, and Ross A Knepper. 2019. Implicit communication of actionable information in human-ai teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[21] Joseph B Lyons, Katia Sycara, Michael Lewis, and August Capiola. 2021. Human–autonomy teaming: Definitions, debates, and directions. *Frontiers in Psychology* (2021), 1932.

[22] Carl Macrae. 2021. Learning from the Failure of Autonomous and Intelligent Systems: Accidents, Safety, and Sociotechnical Sources of Risk. *Risk Analysis* 42 (2021).

[23] Nathan J McNeese, Mustafa Demir, Nancy J Cooke, and Christopher Myers. 2018. Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human factors* 60, 2 (2018), 262–273.

[24] Faiz Siddiqui, Rachel Lerman, and Jeremy B. Merrill. 2022. Telsas running Autopilot involved in 273 crashes reported since last year. *The Washington Post* (2022). https://www.washingtonpost.com/technology/2022/06/15/tesla-autopilot-crashes/

[25] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with humans without human data. *Advances in Neural Information Processing Systems* 34 (2021), 14502–14515.

[26] tblazevic. 2022. *Moonlander*. https://github.com/tblazevic/moonlander

[27] R Core Team et al. 2013. R: A language and environment for statistical computing. (2013).

[28] Nikolai von Janczewski, Jennifer Wittmann, Arnd Engeln, Martin Baumann, and Lutz Krauß. 2021. A meta-analysis of the n-back task while driving and its effects on cognitive workload. *Transportation research part F: traffic psychology and behaviour* 76 (2021), 269–285.

[29] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[30] Wei Xu. 2020. From automation to autonomy and autonomous vehicles: Challenges and opportunities for human-computer interaction. *Interactions* 28, 1 (2020), 48–53.

A MOON LANDER MANEUVERING TASK

A.1 Moon Lander Control

The successful landing criteria included: (i) the horizontal speed was between -1.0 and 1.0, (ii) the vertical speed was between 0.0 and 5.0, and (iii) the lander's rotation angle was between -5.0 and 5.0. The information about the speed and angle was presented via the HUD (i.e., the baseline UI design feature). The up arrow key was used to engage the thruster, which changed the lander's speed. The left and right arrow keys were used to rotate the lander in counterclockwise and clockwise directions respectively (Figure 6). By holding down the arrow key, the team was able to provide the inputs continuously. Additionally, the team was able to change the speed and lander's angle by pressing the up arrow key and the left or right arrow key simultaneously.

A.2 Control Authority Ratio

The control authority ratio was displayed via the horizontal bar with a color gradation on the upper center of the screen in addition to the HUD (Figure 6). Whereas the left side color was always red (i.e., the human icon side), the right side color was changed based on the agent's type: green for the familiarization agent, blue for the more capable agent, and yellow for the less capable agent. The gray rectangle showed the control authority ratio between the human player and the agent. In the beginning of each trial, the gray box was set on a 50-50% configuration (i.e., the middle of the bar), where the human player and the agent had the same control authority. With the 50-50% configuration, a full thrust or rotational input was achieved if both hit the same key simultaneously (i.e., agreement). If the human player hit the left key, and the agent hit the right key (or vice versa), both inputs canceled each other out (i.e., conflict). If only one hit the key, the key input became half of the full thrust or rotational input. The human player was able to change the control authority ratio by 25% by hitting the V or N key. The V key allowed the human player to have more dominant control of the lander and the gray rectangle moved toward the left edge. If the gray rectangle reached the left edge, the human player had full control of the lander, meaning that the agent's inputs did not affect the lander's movement at all (i.e., fully manual). In contrast, the N key increased the agent's control authority ratio by 25%. If the gray rectangle reached the right edge, the agent had full control of the lander, and the human player's inputs had no influences on the lander's movement (i.e., fully automated). Figure 6 shows an example of how to consolidate rotational inputs from the player and the agent in the case of 25% of the human control authority and 75% of agent control authority. In this example, the rotational inputs from the player and the agent were conflicting each other. However, due to the more dominant agent's control authority level, the lander rotated in a clockwise direction with 50% of the magnitude of the full rotational speed.

A.3 Landing Task Difficulty

In the beginning of each trial, the human player was asked to select one of the three landing pads. Each pad had different landing points depending on the Index of Difficulty (ID), which was inspired by the Fitts' Law paradigm [11], that is $ID = log_2\left(1 + \frac{D}{W}\right)$ where D was measured from the initial lander position to the center of each landing pad, and W was given by the width of each landing pad. The initial position of the lander and three landing pads were randomly determined, generating three distinct ID levels of 1.7, 2.7, and 3.7 (i.e., easy, medium, and hard), and each offered 100, 200, and 300 points respectively. Once the player hit the 1, 2, or 3 key to determine the target landing pad, the target landing pad was highlighted with yellow color (Figure 6) . The participants were instructed that they had to select a target landing pad so that the agent could perform the maneuvering task. We made the text information about the selected landing pad yellow in the HUD until the player hit the 1, 2, or 3 key, reminding them of the required action.

B N-BACK TASK

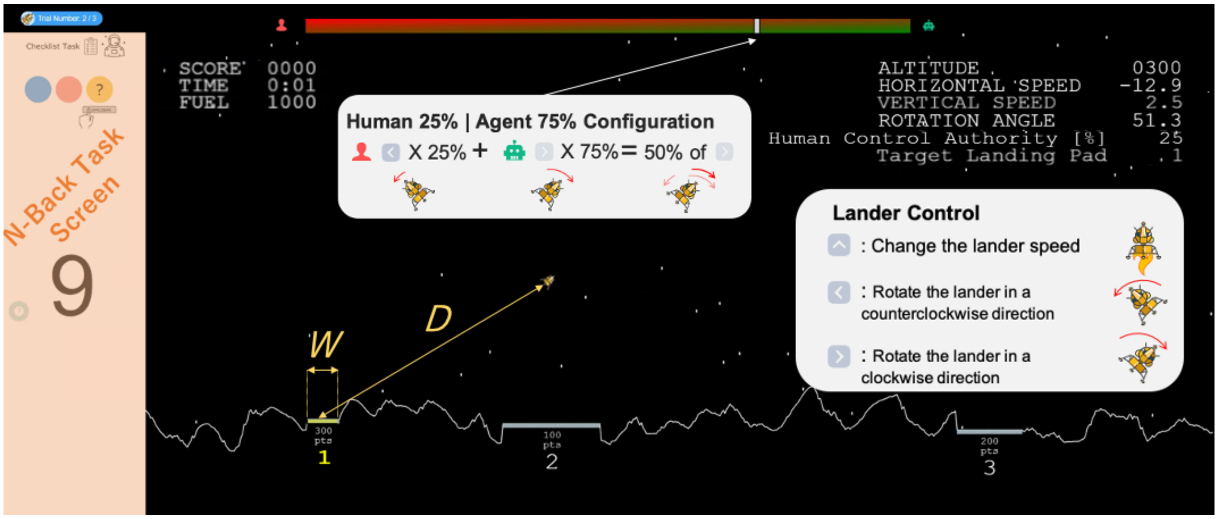
The n-back task is widely used as a cognitive secondary task in driving simulator studies [28]. We employed the n-back task to increase the player's cognitive load and divide attention with the maneuvering task. To associate the n-back task with the moon landing task context, we instructed the participants that the n-back task was an analog to a flight checklist task to simulate a more realistic moon landing scenario where astronauts communicate with a co-pilot or mission control. Figure 7 shows examples of the 0-back and 2-back tasks with the four possible outcomes. We made some modifications for this study while acknowledging the n-back task standard. We employed a non-audio version of the n-back task with 500 ms of the stimulus presentation and 2500 ms of the blank period. Due to the nature of the landing task, the length of each trial varies across trials as well as participants, meaning that each participant received a different number of stimuli in each trial. The n-back task accuracy was computed by $\frac{H+CR}{H+M+FA+CR}$ where H, CR, M, and FA are Hits, Correction Rejections, Misses, and False Alarms respectively. During our pilot study, we randomly showed a sequence of digits; however, we observed a situation where no space-bar inputs still led to the 100% accuracy (i.e., no target digits were presented, and all were correct rejections), making it difficult to distinguish whether the player was dedicated to the n-back task. Therefore, we decided to present a target stimulus (i.e., a space-bar input is required) with a 25% chance to ensure that participants who do not conduct the n-back at all can receive penalties for not paying attention to the n-back task.

C OVERALL GAME SCORES

In each trial, the team gained $GameScores$ that were calculated by $Landing\ Points + (Remaining\ Fuel/Overlap\ Factor)$. After exploring different values, We selected the $Overlap\ \ Factor$ value of 5. Figure 8 shows how $Landing\ Points$, $Remaining\ Fuel$, and $Accuracy$ contributed to $Game\ Scores$ with $Overlap\ Factor$ of 5.

D HEURISTIC AGENT

In the beginning of the experiment, the participants were told a hypothetical scenario where they were invited into an evaluation session of an AI agent that collaborates with astronauts on a moon landing task. They were instructed to evaluate one of the two AI agents that were developed by space companies and provide their feedback on their AI collaboration experience. For this experiment,



```
SCORE   0000
TIME    0:01
FUEL    1000
```

```
           ALTITUDE                0300
           HORIZONTAL SPEED       -12.9
           VERTICAL SPEED           2.5
           ROTATION ANGLE          51.3
Human Control Authority [%]          25
           Target Landing Pad         1
```

Human 25% | Agent 75% Configuration

X 25% + X 75% = 50% of

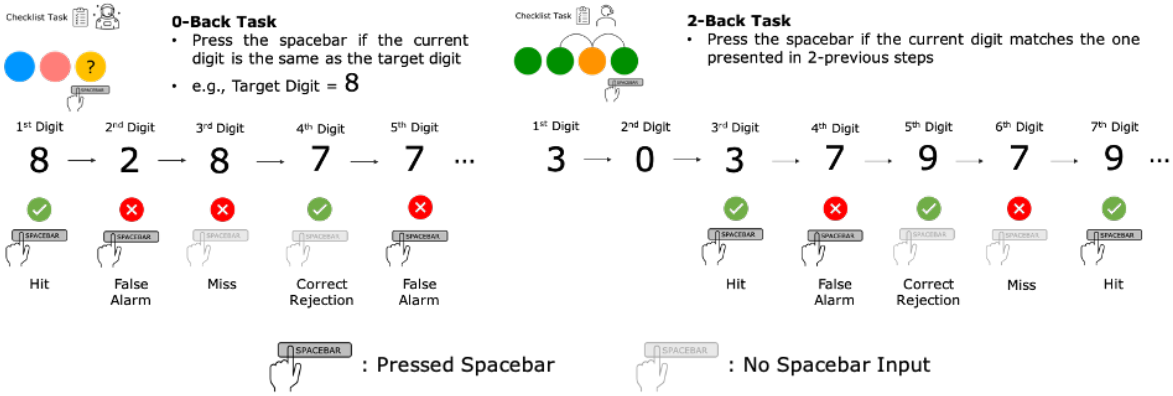
Lander Control

: Change the lander speed

: Rotate the lander in a counterclockwise direction

: Rotate the lander in a clockwise direction

Figure 6: The up, left, and right arrow keys were used for the moon lander maneuvering. The V and N keys were used to move the gray rectangle left (i.e., the player's control authority became more dominant) or right (i.e., the agent's control authority became more dominant) respectively. The player was able to reset the control authority ratio at the 50-50% configuration by hitting the B key. To compute the *ID* values, *D* was measured from the initial lander position to the center of each landing pad. It should be noted that the snapshot shows TIME = 0:01 (i.e., the lander moved for 1 [s]), and therefore, technically, *D* shown here is not the same as what was used for the *ID* value calculation.



0-Back Task
- Press the spacebar if the current digit is the same as the target digit
- e.g., Target Digit = 8

2-Back Task
- Press the spacebar if the current digit matches the one presented in 2-previous steps

: Pressed Spacebar

: No Spacebar Input

Figure 7: Examples of the four possible outcomes in the 0-back and 2-back tasks

three heuristic agents were developed: (i) the familiarization, (ii) the more capable, and (iii) the less capable heuristic agents. The familiarization agent was used only during the familiarization session and designed to exhibit the best landing performance. There are three steps that the heuristic agents execute. First, the heuristic agents aim to minimize the distance between the center of a selected landing pad and the end of the projected trajectory. Next, the heuristic agents focus on maintaining the vertical speed at a certain level so that the vertical speed does not become too fast. Then, the heuristic agents initiate the final approach maneuvering and reduce the lander speed as well as manipulate the lander angle to ensure a successful landing. We tweaked the thresholds pertaining to the final approach phase to design the more and less capable heuristic agents. The more capable agent was designed to initiate the final approach maneuvering earlier than the less capable agent. Figure 9 shows the landing performance levels of each type of the heuristic agents.

E UI DESIGN

Three UI designs were tested in the experiment: (i) the baseline, (ii) the trajectory projection, and (iii) the trajectory projection with the vertical speed indicator.

E.1 Baseline

The baseline condition displayed the HUD information, including: Score, Time, Remaining Fuel, Lander Speed & Rotation Angle, Human Control Authority, and Selected Landing Pad Number (Figure 10a). In this study, the player was required to select a target landing

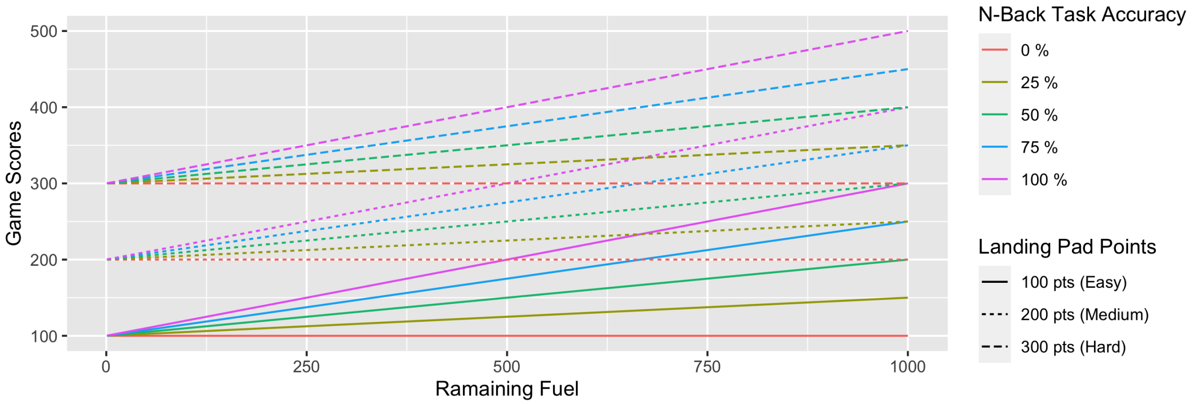


Figure 8: Possible game scores that the team could gain with *Overlap Factor* value of 5.

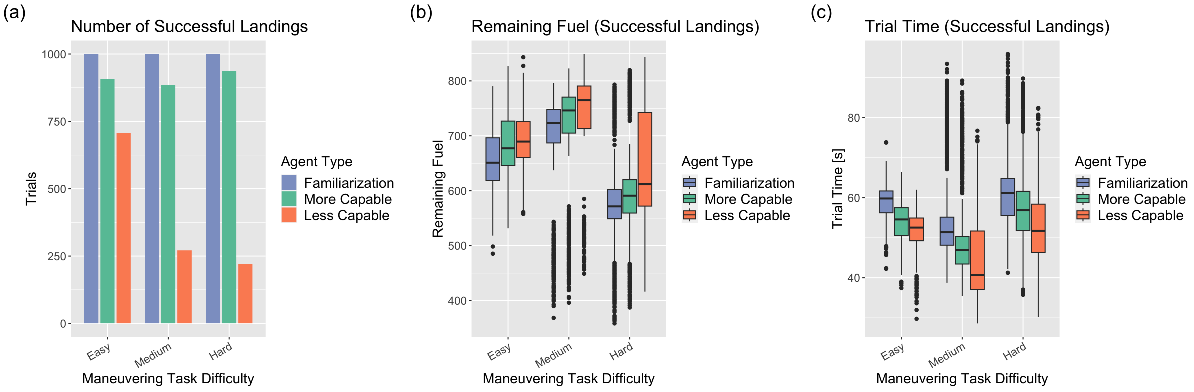


Figure 9: Performance comparison between the three heuristic agents; (a) the familiarization agent lands successfully all the time. Whereas the more capable agent exhibits approximately 90% chance of successful landing across the three difficulty levels, the less capable agent shows roughly 20% chance of successful landings for the medium and hard difficulty levels. (b) & (c) the familiarization and more capable agents initiate the final landing procedure earlier to land safely with some vertical speed buffer. In contrast, the less capable agent tends to land more aggressively and initiate the final landing procedure later than the other two agents. The trial time shows that the less capable agent exhibits a shorter length of a trial when it lands successfully, indicating the low frequency of thrust activities.

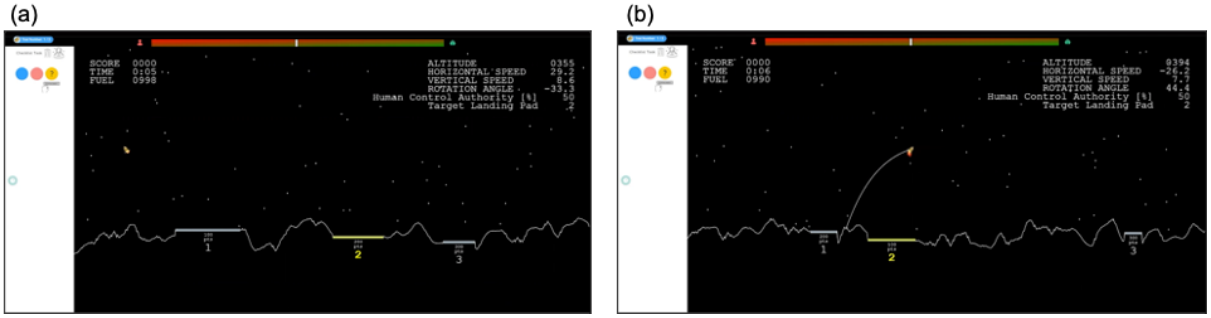
pad in the beginning of each landing attempt; otherwise, the agent teammate could not conduct the maneuvering task. During our pilot study, we observed situations where the player was prone to forget to select a target landing pad. Therefore, we decided to highlight the text line of the target landing pad information with a yellow color until the player hits number 1, 2, or 3 key so that the player could notice the need for selecting the target landing pad. In addition to the HUD information, the baseline condition displayed the control authority indicator on the upper center of the screen. The screen zoomed in when the lander approached the surface.

E.2 Trajectory Projection

With the HUD information and the control authority indicator, the trajectory projection condition displayed a predicted lander trajectory with a white arc (Figure 10b). The white arc starts from the lander and ends at the point where the predicted trajectory meets the terrain.

E.3 Trajectory Projection with Vertical Speed Indicator

The third UI condition provided the player with the same UI features as the trajectory projection condition, and the only difference was the provision of the vertical speed indicator that was represented with the color box. The needle (i.e., gray horizontal bar) is vertically moving in the box, and the color of the box is changing based on the vertical speed. The indicator corresponds to four actionable recommendations: (i) Disengage, (ii) Keep, (iii) Engage, and (iv) Thrust. The disengage recommendation indicates that the lander is ascending, and therefore the player needs to provide no inputs (Figure 11a). The keep recommendation is a green state where the player can ensure a soft landing with a buffer (Figure 11b). The



```
SCORE  0000                                 ALTITUDE                 0355
TIME   0:05                                 HORIZONTAL SPEED         29.2
FUEL   0998                                 VERTICAL SPEED            8.6
                                            ROTATION ANGLE          -33.3
                                Human Control Authority [%]            50
                                            Target Landing Pad          2
```

```
SCORE  0000                                 ALTITUDE                 0394
TIME   0:06                                 HORIZONTAL SPEED        -26.2
FUEL   0990                                 VERTICAL SPEED            7.7
                                            ROTATION ANGLE           44.4
                                Human Control Authority [%]            50
                                            Target Landing Pad          2
```

Figure 10: (a) the baseline UI design contained the HUD information, (b) with the trajectory projection UI design, the white arc informed the player of where the lander was heading to.

engage recommendation indicates that although the player can land successfully, some thruster inputs are encouraged to ensure a successful landing; the color of the box turned yellow with this state (Figure 11c). The thrust recommendation shows a red color, requiring the player to reduce the vertical speed (Figure 11d). The borderline between the engage and thrust recommendation zones was designed to align with the top of the landing pad, making the player’s task easier; the player just needed to ensure that the gray horizontal bar is above the surface of the landing pad (Figure 11e). Figure 12 shows an example trial using the trajectory projection with the vertical speed indicator.

F QUESTIONNAIRES

F.1 Demographic Questionnaire

The demographic questionnaire was administered after the participants signed up for the experiment by submitting the informed consent. The demographic questionnaire asked the participants about the following aspects: (i) age, (ii) handedness, (iii) computer experience, and (iv) game experience. The participants were prompted to type their ages for the first question item and select right, left, or both for the handedness question. The computer experience question item offered options: (CE1) more than 70 hours/week, (CE2) between 50-70 hours/week, (CE3) between 30-50 hours/week, (CE4) between 10-30 hours/week, and (CE5) less than 10 hours/week. Likewise, the participants were asked to indicate their game experience by selecting one from the following options: (GE1) Esports (e.g., Esports club, competition), (GE2) greater than 20 hours/week, (GE3) 10-20 hours/week, (GE4) less than 10 hours/week, and (GE5) not a gamer. Table 1 shows the summary of the responses to the demographic questionnaire across the six groups.

F.2 NASA-RTLX

We administered the NASA-RTLX to shorten the length of the workload assessment process. We adapted the wording of the mobile version of the NASA-TLX (https://humansystems.arc.nasa.gov/groups/tlx/tlxapp.php [accessed on 03/13/2023]) and slightly modified by providing some examples; for instance, we included “e.g., thinking, deciding, calculating, remembering, looking, searching, etc.” in the mental demand question. The anchor appeared once the respondent clicked on a scale. The NASA-RTLX questionnaire was administered after each block.

F.3 Agent Interaction Questionnaire

The agent interaction questionnaire consisted of four question items. The first two questions were designed to ask the participants about their perceived agent predictability and capability, which were adapted from a trust scale suggested by [16]. We expected them to report the same level of their perceived agent predictability regardless of the heuristic agent types due to the rule-based and threshold-oriented implementation. However, we expected that the participants in the more capable heuristic agent group would provide higher ratings when compared to those who interacted with the less capable heuristic agent. The third and fourth question items were set to investigate the relationship between their ratings and behavioral measures captured by the keystrokes and the control authority changes.

G SUPPLEMENTARY RESULTS

Figure 13 shows the number of human key inputs and conflicts between the player and agent.

H HUMAN-AGENT TEAMWORK CONSIDERATIONS FOR FUTURE WORK

Table 2 presents a set of considerations and examples to further investigate human-agent teamwork using the moon lander game environment, which is built on the design considerations proposed by [5] and OPD [18].

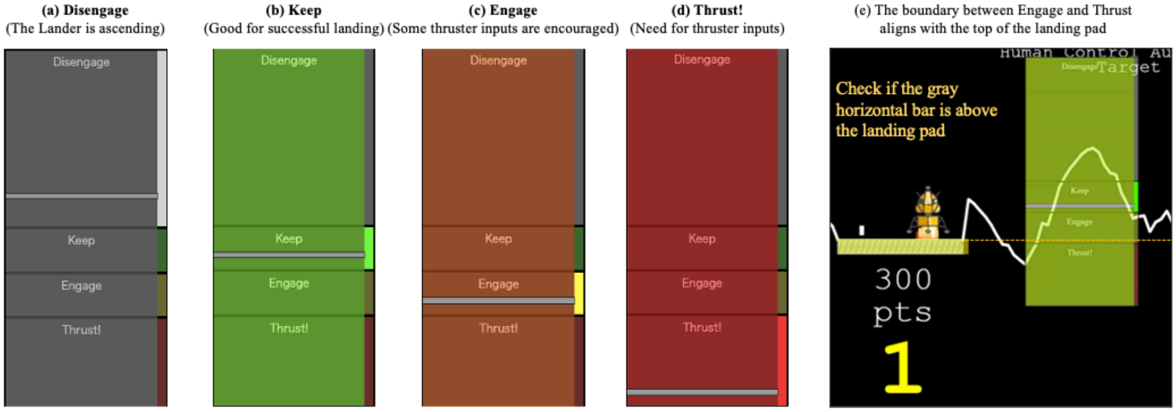
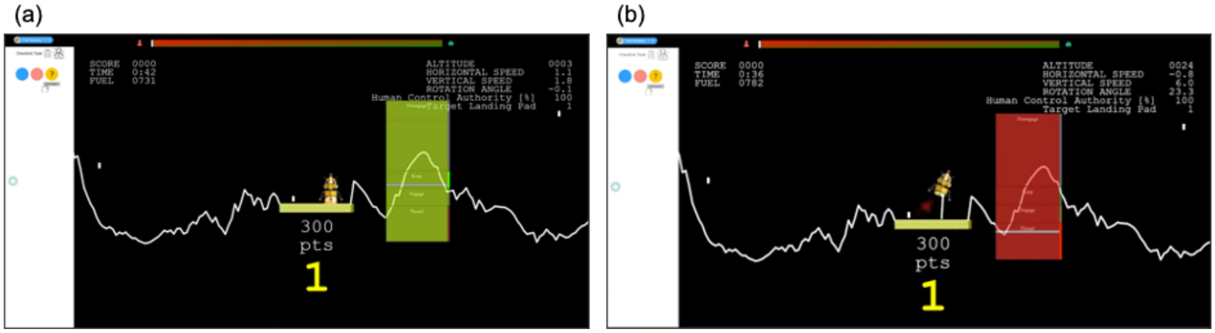

(a) Disengage
(The Lander is ascending)
(b) Keep
(Good for successful landing)
(c) Engage
(Some thruster inputs are encouraged)
(d) Thrust!
(Need for thruster inputs)
(e) The boundary between Engage and Thrust
aligns with the top of the landing pad
Disengage
Keep
Engage
Thrust!
Check if the gray
horizontal bar is above
the landing pad
Human Control Au
Target
300
pts
1

Figure 11: The vertical speed indicator was designed to convey the information about the vertical speed in a more visual manner. The gray horizontal bar indicates the current vertical speed, and the color of the entire box is changed depending on the vertical speed.



(a)

SCORE 0000
TIME 0:42
FUEL 0731

ALTITUDE 0003
HORIZONTAL SPEED 1.1
VERTICAL SPEED 1.8
ROTATION ANGLE -0.1
Human Control Authority [%] 100
Target Landing Pad 1

(b)

SCORE 0000
TIME 0:36
FUEL 0782

ALTITUDE 0024
HORIZONTAL SPEED -0.8
VERTICAL SPEED 6.0
ROTATION ANGLE 23.3
Human Control Authority [%] 100
Target Landing Pad 1

Figure 12: The vertical speed indicator appeared when the camera zoomed in. (a) the keep recommendation indicates that the vertical speed is good for a successful landing, (b) the player is required to reduce the vertical speed when the indicator's color turns red.

Table 1: Summary of responses to demographic questionnaire across six Groups (abbreviations are as follows: MC: More Capable, LC: Less Capable, B: Baseline, T: Trajectory Projection, T+V: Trajectory Projection with Vertical Speed Indicator, CE: Computer Experience, and GE: Game Experience. The standard deviation is presented with parentheses.

Agent	UI	Players	Age (years)	Left	Right	Both	CE1	CE2	CE3	CE4	CE5	GE1	GE2	GE3	GE4	GE5
MC	B	6	27.0 (10.7)	0	6	0	1	2	1	2	0	1	1	0	3	1
	T	13	26.2 (8.30)	1	11	1	3	3	3	4	0	1	2	2	2	6
	T+V	9	28.1 (10.9)	1	6	2	2	3	3	0	1	0	2	1	4	2
LC	B	14	27.8 (11.7)	1	13	0	1	2	5	4	2	0	4	4	4	2
	T	9	24.3 (9.01)	0	9	0	1	1	4	3	0	1	1	1	3	3
	T+V	9	25.1 (10.6)	0	9	0	3	0	5	1	0	0	1	4	1	3

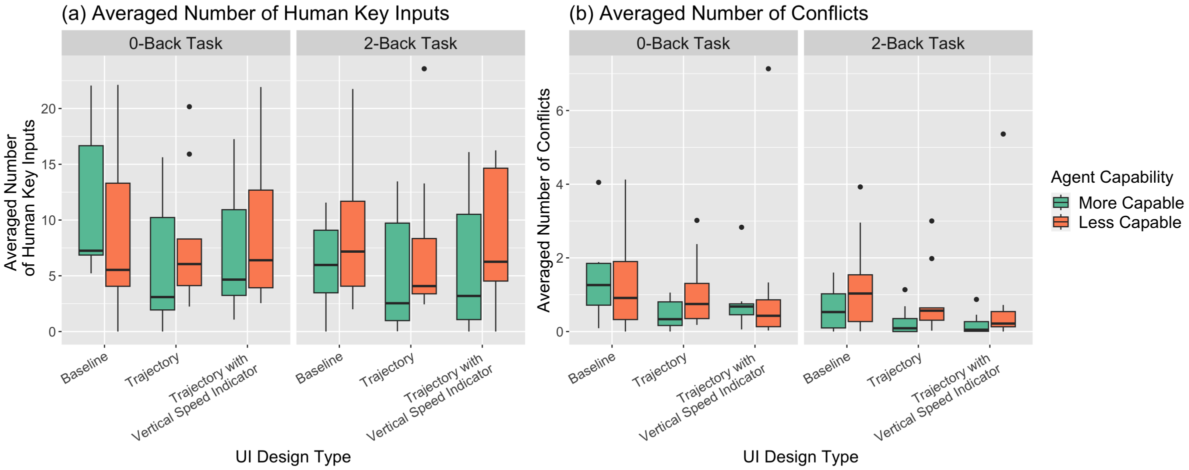


Figure 13: (a) Averaged number of human key inputs in a trial and (b) averaged number of conflicts across all the conditions.

Table 2: Potential avenues for further investigation of more sophisticated human-agent teamwork using the moon lander game in relation to the design considerations for Human-Agent Collaboration [5] and Coactive Design [18]. The design considerations with * are presented on [5, p.5].

Key factors	Design considerations for human-agent teamwork	Examples for future work with moon lander game setting
Observability	- Are the agent's intentions and behaviors observable to users?	- The agent could indicate its own self-reported confidence level during the maneuvering task.
Predictability	- Are the agent's actions predictable to users?	- The agent could convey its intent explicitly and/or implicitly.
Directability	- *What task is the agent to perform? - *What level of autonomy is appropriate for this agent? - How to nudge teammates to take a specific action to improve team effectiveness?	- The agent could be designed to select a target landing pad and change the control authority level. - The agent could suggest that the player should change the control authority level.
Interpretability	- *How to explain the intent and behaviors of agent? - How to enhance agent's transparency keeping observability, predictability, and directability in mind?	- The agent could convey its intent, reasoning, and future state via UIs. - The team could employ explicit and implicit communication.
Teamwork with Agents	- *How to establish a common ground between human and agent? - How to facilitate processes of resolving a conflict situation and reaching an agreement?	- The team consists of different types of members who have different attitudes (e.g., risk-seeking or -averse team members) - The player and the agent select different target landing pads, requiring them to resolve the conflict. - The team detects cracks on a landing pad that prevents the team from landing successfully, requiring the team to reset their target landing pad).
Resilience	- *When and how an agent can offer help to humans? - *What are the most effective means for an agent to request for help?	- The agent could offer or request help based on its self-reported confidence level. - The agent could over- or under-estimate its own capability and the player could over- or under-trust the agent.