Evaluating Robotic Performative Autonomy in Collaborative Contexts Impacted by Latency

Rafael Sousa Silva, Cailyn Smith, Lara Bezerra, and Tom Williams

Abstract— **Maintaining Situational Awareness (SA) is critical in space exploration contexts, yet made particularly difficult due to the presence of communication latency. In order to increase human SA without inducing cognitive overload, researchers have proposed *Performative Autonomy* (PA), in which robots intentionally interact at a lower level of autonomy than they are capable of. While researchers have demonstrated positive impacts of PA on team performance even under high latency, previous work on PA has not examined how the benefits of PA might be mediated by latency. In this work, we thus evaluate the impact of latency and PA on trust, SA, and human perceptions of robot intelligence and autonomy. Our results suggest that lower performed autonomy leads to increased cognitive load, especially when robot communication happens frequently and latency is present. In addition, we observe no effect of the PA strategies used within our experimental paradigm on SA, and instead find evidence that operating under high latency leads to negative perceptions of robots regardless of choice of PA strategy.**

I. INTRODUCTION AND MOTIVATION

Robots deployed in space exploration contexts, whether autonomous, semi-autonomous, or teleoperated, can have significant impact on the success of exploration missions [1]. Free-flying Astrobee robots, for instance, can assist with maintenance and repetitive tasks, redirecting human crewmates' attention to more important tasks [2]. Other types of robots, such as rovers, are critical to long-term space exploration contexts that are dangerous or infeasible for human astronauts [3]. Critically, humans teaming with robots in space exploration contexts, as well as other high-stake domains where safety is paramount (e.g., urban search and rescue [4], disaster response [5], industry [6], surgery [7]) tend to experience high levels of cognitive load, which may lead to reduced Situational Awareness (SA). Moreover, it is critical for astronauts and ground control operators to maintain high SA about their robot teammates' ongoing tasks, as the loss of human SA can lead to fatal consequences [8], [9].

To minimize cognitive load while boosting human efficiency, researchers have proposed the use of *adaptive automation* to dynamically adjust robot autonomy levels to account for cognitive load [10], [11]. Yet while adaptive automation can help decreasing human cognitive load, it may also lead to the undesired loss of SA [12]. As such, robots also need to mitigate the loss of SA by bringing human attention to task-relevant stimuli while preserving their cognitive load. To meet this need, Roy et al. [13] introduced *Performative Autonomy (PA)*, in which autonomous robots deliberately choose to "perform" a lower level of autonomy than they are truly capable of, by asking questions they do not truly need the answer to (or engaging in other communicative activities that increase SA). The authors showed that robots that leverage PA as a communication strategy may increase human teammates' SA by bringing their attention to important task elements without raising cognitive load. In addition to SA, natural language strategies like PA can help calibrate other important factors to human-robot collaboration, such as trust [14], [15]. Yet, while natural language interaction is expected to be increasingly important as an enabling modality for collaborative space tasks [16], the potential for natural language-based human-robot interactions in space, especially between remote interactants, may be limited by the communication latency that characterizes space contexts.

In previous work, Sousa Silva et al. [17] further demonstrated positive impacts of PA on human perceptions of robots as teammates even in high latency scenarios. However, they did not examine whether increased latency might attenuate – or accentuate – the benefits of PA. In addition, prior work has not yet studied whether changes in PA actually lead to noticeable changes in *perceived* robot autonomy. This is important because robots that are perceived to be less autonomous may also be perceived as less intelligent [18], and lower levels of autonomy and intelligence may in turn lead to lower levels of human-robot trust [19]. In this work, we thus experimentally assess the effects of PA and latency on SA, Cognitive Load, Perceived Autonomy, Perceived Intelligence, and Trust through a human-robot collaborative task affected by latency.

We expect robots' performance of lower levels of autonomy to benefit human SA. This is because robots ask questions that direct attention to important task aspects when lower levels of autonomy are performed [13]. However, the introduction of such questions makes human input necessary for robot activity to resume. As such, we believe that human perceptions of robot autonomy and intelligence will be lower as lower levels of autonomy are performed [20]. These lowered perceptions of robot capabilities may in turn lead to reduced human-robot trust, as robots may be regarded as less capable of completing their tasks without human input [21]. Furthermore, we believe that higher levels of latency will have a negative impact on SA due to increased delays in human-robot communication. Yet, we do not expect to observe latency effects on perceived robot autonomy, perceived robot intelligence, or trust ratings, as we do not expect robots to be blamed for longer experienced communication delays. Finally, we do not believe that cognitive load will be affected

MIRRORLab, Colorado School of Mines, Golden, CO 80401, USA. rsousasilva@mines.edu, ccsmith1@mines.edu, lbezerra@mines.edu, twilliams@mines.edu

by PA nor latency, since previous work has observed stable cognitive load rates when PA is in use, even in high-latency contexts [13], [17]. Based on these intuitions, we propose three key hypotheses:

H1: When robots perform lower levels of autonomy, human SA will improve. However, Perceived Autonomy, Perceived Intelligence, and Trust ratings will be lower. Cognitive load will not be affected.

H2: As Latency increases, human SA will deteriorate. Cognitive load and ratings for Perceived Autonomy, Perceived Intelligence, and Trust will not be affected.

H3: There will be no observed interaction effects between Latency and PA Strategy.

II. METHODOLOGY

A. Experimental Design

To evaluate our hypotheses, we conducted a human-subjects study that was approved by an Institutional Review Board. In this study, participants were tasked with completing a collaborative resource management game simulating an interaction that could happen in a space context. A Misty II robot *collected* four different types of resources that could then be *spent* by participants. Participants' goal was to spend resources to clear 20 "stations" randomly distributed across a 10x10 grid as fast as possible. Each station required either 60 resources of one color or 30 resources of two colors to be cleared. Participants were able to increase their SA about collected resource amounts by observing a panel that displayed estimates of how many resources of each color they had available. Similarly, they were able increase their SA about the robot's status by checking a separate monitor with a live video feed of the robot. The robot was surrounded by four colored indicators that represented the game resources and always collected resources of the color it faced. To coordinate which type of resource the robot should collect, participants needed to wait for a robot message, i.e., participants were not able to task the robot with collecting a specific resource type, unless expressly asked by the robot. As in [17], robot messages were sent to participants every 50 seconds through concurrent text and audio. Communication latency was designed to account for (a) the time needed for the robot to send a message to the participant, (b) the time needed for the participant's input to be relayed back to the robot, and (c) the time needed for the robot to process human input and switch between resources. Due to space constraints, we refer readers to [17] for additional game details.

Our experiment followed a Greco-Latin Square design involving three latency levels and three PA levels. The three latency levels were:

- *Low Latency* – 0 seconds to represent immediate, uninterrupted collaboration;
- *Medium Latency* – 15 seconds to represent the round-trip communication delays of lunar missions [22];
- *High Latency* – 30 seconds to represent collaboration in missions slightly beyond lunar range.

The three PA Strategies were those used by Roy et al. [13]:

- *Low PA* – Request selection between multiple options (e.g., "I was collecting orange resources. Should I keep collecting orange resources, or switch to pink resources, or to red resources?");
- *Medium PA* – Request confirmation of a single option (e.g., "I was collecting red resources. Can I now switch to collecting blue resources?");
- *High PA* – Propose and select a single option without opportunity for veto (e.g., "I was collecting pink resources. I am now going to collect orange resources.").

Participants played three games, experiencing all levels of latency and PA strategies, as the Greco-Latin Square design handled ordering effects for both factors simultaneously.

B. Measures

After each game, participants were given a survey that measured their ratings of the robot.

1) **Situational Awareness** was measured in game through questions inspired by Level 2 comprehension queries from the Situational Awareness Global Assessment Technique (SAGAT) [23]. Periodically the information about how many resources were available was hidden, and participants were asked to report which resource level was highest from memory.
2) **Cognitive Workload** was measured through the NASA Task Load Index (TLX) scale [24].
3) **Perceived Autonomy** was measured through the preliminary version of the Robot Autonomy Perception Scale (RAPS) described in [25].
4) **Trust** was measured through the Multi-Dimensional Measure of Trust (MDMT) Capacity scale, using both the Reliable and Capable subscales [26].
5) **Perceived Intelligence** was measured through the Godspeed IV scale [27].

C. Procedure

Upon arrival, participants signed an informed consent form, surrendered electronic devices, and filled out a demographics survey. They were randomly assigned to one of our three experimental conditions and then taken to a room with two desks: one with a computer and one with a monitor displaying the Misty II robot. Participants were told that the robot was operating in a remote location. Participants were then walked through a tutorial to understand game mechanics. After answering participants' questions, the experimenter headed to a separate room and remotely started the first game. After the first game was finished, the experimenter guided participants back to the survey room, where they filled out the questionnaires and scales listed in the previous subsection based on their most recent game. While participants were completing the survey, the experimenter set up the next game. Upon survey completion, participants were guided back to the experiment room to play the second game. This procedure was repeated until participants were done with their third game. Finally, participants were debriefed and paid $15.

D. Participants

115 participants (47 F, 66 M, 2 NB) were recruited from the Colorado School of Mines academic body. Participant ages ranged from 18 to 46 ($M = 21.2, SD = 4.35$). We discarded 31 data points due to participants' failed attention checks, 6 data points due to robot malfunction, and 2 data points from participants who completed the experiment twice (the data from their first run was preserved). Data analysis was thus performed on the data from the resulting 76 participants (35 F, 40 M, 1 NB), averaging 25 participants per condition. All data is available online at `https://osf.io/5ja4s`.

E. Statistical Analysis

We conducted Bayesian Repeated Measures (RM) ANOVA tests with Inclusion Bayes Factor (BF_{10}) Analysis through the bayestestR [28] and BayesFactor [29] R packages. These BF_{10} values were calculated across matched models through model averaging [30], and indicate the relative strength of evidence for models including each candidate main effect or interaction effect, in terms of ability to explain gathered data. When a main or interaction effect could not be ruled out ($BF_{10} > 0.333$, i.e., evidence *against* inclusion (BF_{01}) no greater than 3:1), post hoc RM Bayesian t-tests were used to examine pairwise comparisons between conditions. Please refer to [17] for a detailed explanation as to why we use a Bayesian statistical framework.

III. EXPERIMENTAL RESULTS

In this section, we report the results of our statistical analysis. Table I shows the mean and standard deviation for each dependent variable within each condition. Table II reports the Inclusion Bayes Factors for the impact of Latency, PA strategy, and their interactions, on each dependent variable. Table III reports post hoc test results for Latency and PA strategy. Finally, Figure 1 displays results across conditions for each dependent variable.

A. Situational Awareness

An RM-ANOVA (see Table II) revealed very strong evidence against effects of Latency or PA strategy on SA, and anecdotal evidence against an interaction effect.

B. Cognitive Load

An RM-ANOVA (see Table II) revealed moderate evidence for an effect of PA Strategy and for an interaction effect of Latency and PA Strategy on cognitive load, and moderate evidence against an effect of Latency. Post hoc tests (see Table III) suggest that higher PA led to lower levels of cognitive load. Finally, these tests also show that while increased levels of Latency lead to higher cognitive load levels in the *Low PA* and *Medium PA* conditions, they actually lead to reduced cognitive load readings in the *High PA* condition.

C. Perceived Autonomy

An RM-ANOVA (see Table II) revealed extreme evidence for an impact of PA strategy on Perceived Autonomy, anecdotal evidence against an effect of Latency, and moderate evidence against an interaction effect. Post hoc tests (see Table III) revealed that higher PA led to higher Perceived Autonomy ratings. Moreover, results suggest that increased Latency might also lead to lower ratings.

D. Trust

An RM-ANOVA (see Table II) revealed moderate evidence for an effect of Latency on Trust in the robot, strong evidence against an effect of PA strategy, and anecdotal evidence against an interaction effect. Post hoc tests (see Table III) suggested that increased levels of Latency lead participants to rate robots as less trustworthy teammates. In addition, as Latency levels increased participants rated robots using the *Low PA* and *Medium PA* strategies as less trustworthy. Yet, increased Latency in the *High PA* condition led to higher trust ratings.

E. Perceived Intelligence

An RM-ANOVA (see Table II) revealed strong evidence of an effect of Latency on Perceived Intelligence, moderate evidence against an effect of PA strategy, and anecdotal evidence against an interaction effect. Post hoc tests (see Table III) revealed no difference in Perceived Intelligence ratings between robots operating under *Low* and *Medium Latency*. However, results suggest that robots operating in *High Latency* situations are generally perceived to be less intelligent. In addition, robots using the *Low PA* and *High PA* strategies were perceived to be more intelligent when operating under *Medium Latency* over *Low* and *High Latency*. On the other hand, robots using the *Medium PA* strategy were perceived to be less intelligent as Latency levels increased.

IV. DISCUSSION

A. Hypothesis One

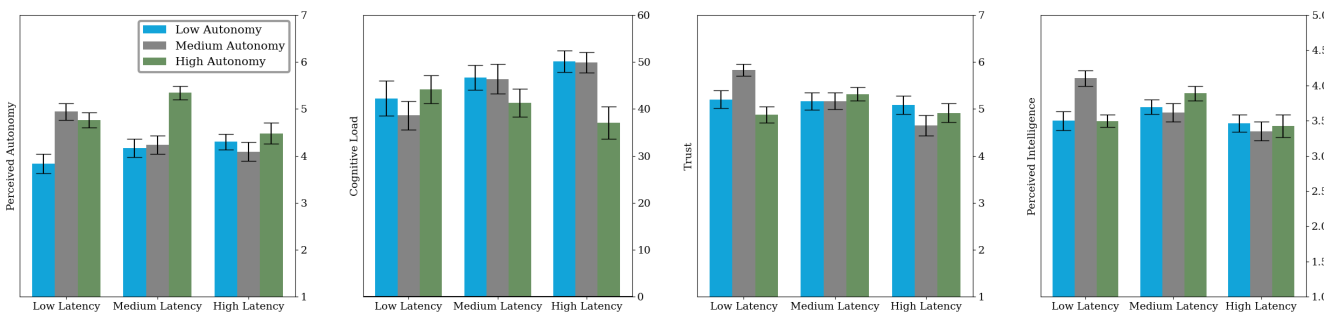
H1 stated that Perceived Intelligence, Perceived Autonomy, and Trust ratings would be lower as the robot's PA Strategy level decreased while SA would be improved and cognitive load would remain the same. This hypothesis was partially supported.

1) Perceived Autonomy: We confirmed that robots were indeed perceived to be less autonomous when they were performing lower PA strategies. This finding replicates [17]'s results for perceived dependency and reinforce [13]'s proposition that different levels of autonomy can be performed through different types of dialogue.

2) Situational Awareness: The experimental results from [13]'s work suggested that lower PA strategies were beneficial for human teammates, increasing their overall SA. However, the results obtained from our study suggest no observable differences in SA across conditions. We believe two key differences between our experiment and [13]'s may help explain our conflicting results. First, in [13]'s work robot communication was designed to happen every 100 seconds

	Perceived Autonomy		Situational Awareness		Cognitive Load		Trust		Perceived Intelligence	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Low Latency (LL)	4.513	1.191	0.879	0.099	41.707	21.499	5.306	1.070	3.700	0.730
Medium Latency (ML)	4.584	1.170	0.880	0.095	44.795	19.125	5.219	1.111	3.733	0.751
High Latency (HL)	4.291	1.292	0.878	0.100	45.728	17.446	4.886	1.354	3.413	0.921
Low PA (LPA)	4.100	1.241	0.876	0.097	46.374	19.055	5.153	1.259	3.553	0.800
Medium PA (MPA)	4.425	1.254	0.888	0.086	44.986	18.348	5.218	1.138	3.690	0.823
High PA (HPA)	4.864	1.157	0.872	0.110	40.870	20.666	5.039	1.138	3.602	0.780
LL + LPA	3.831	1.335	0.898	0.081	42.256	24.398	5.205	1.256	3.500	0.880
LL + MPA	4.944	1.182	0.855	0.095	38.653	20.655	5.832	0.838	4.104	0.751
LL + HPA	4.765	1.055	0.883	0.120	44.213	19.443	4.880	1.116	3.496	0.560
ML + LPA	4.165	1.306	0.886	0.087	46.707	17.768	5.166	1.259	3.696	0.705
ML + MPA	4.241	1.262	0.900	0.085	46.385	20.329	5.171	1.165	3.615	0.854
ML + HPA	5.347	0.941	0.853	0.113	41.293	19.277	5.319	0.909	3.888	0.695
HL + LPA	4.304	1.083	0.845	0.123	50.160	15.000	5.089	1.261	3.464	0.814
HL + MPA	4.091	1.318	0.909	0.079	49.920	14.059	4.650	1.411	3.352	0.863
HL + HPA	4.479	1.476	0.880	0.098	37.103	23.278	4.918	1.389	3.423	1.086

TABLE I: Mean and Standard Deviation values for each analysis group.



(a) Perceived Autonomy (b) Cognitive load (c) Trust (d) Perceived Intelligence

Fig. 1: Effects of Latency and PA Strategy on key dependent variables. Error bars represent 95% confidence interval.

	Latency	PA Strategy	Latency * PA Strategy
Perceived Autonomy	0.744	**5.34e+05**	*0.250*
Situational Awareness	*0.046*	*0.082*	0.354
Cognitive Workload	*0.110*	**5.32**	**4.14**
Trust	**4.69**	*0.098*	0.776
Perceived Intelligence	**20.43**	*0.150*	0.938

TABLE II: RM-ANOVA Inclusion Bayes Factors. Results with conclusively positive evidence are bolded; Results with conclusively negative evidence are grayed out and italicized.

whereas in our study it happened every 50 seconds. The more frequent rate of communication might justify why SA levels were high in all of our experimental conditions, as the robot was raising human awareness about its current status more often. Second, while in [13]'s experimental task participants had to navigate different windows that contained different sets of relevant information, in our experiment all important information was available in the same environment. Having the robot video feed and the panel displaying current resource amounts always available might have helped participants maintain high SA about the task in all of our conditions.

3) Cognitive Load: As in [13]'s and [17]'s work, we expected that the use of PA strategies would not affect participants' workload levels across conditions. However, participants felt more overwhelmed when collaborating with robots performing lower levels of autonomy. This may be explained by the fact that participants needed to perform additional tasks when interacting with robots operating under *Medium* and *Low PA*. Specifically, they needed to either answer the robot's Yes/No question in the *Medium PA* condition and needed to arbitrate between several options in the *Low PA* condition. In addition, the increased rate of communication of our experimental context might have contributed to participants' increased cognitive load, as they had to answer robot questions much more frequently. While in [13]'s work cognitive load was measured through a single Likert scale (which may not be enough to correctly assess human factors [31]), we believe the less frequent rate of robot communication contributed towards more stable cognitive load levels among participants. Furthermore, while [17]'s experiment had the same rate of communication as our exper-

	Latency			PA Strategy		
	Low Latency (LL) v. Medium Latency (ML)	LL v. High Latency (HL)	ML v. HL	Low PA (LPA) v. Medium PA (MPA)	LPA v. High PA (HPA)	MPA v. HPA
Perceived Autonomy	*0.201*	0.411	1.84	**3.10**	**2.86e+05**	**75.03**
Situational Awareness	*0.178*	*0.170*	*0.172*	*0.248*	*0.180*	*0.312*
Cognitive Load	*0.272*	0.354	*0.195*	*0.312*	**6.11**	2.35
Trust	*0.214*	**5.70**	**3.33**	*0.185*	*0.233*	0.412
Perceived Intelligence	*0.186*	**4.69**	**40.58**	0.357	*0.194*	*0.277*

TABLE III: Inclusion Bayes Factors (BF_{10}) for Latency and PA Strategy. Results with conclusively positive evidence are bolded; Results with conclusively negative evidence are grayed out and italicized.

iment, we believe that it might not have detected increased cognitive load levels when lower levels of autonomy were performed because they did not factor in the weights of the NASA TLX scale components into their analysis.

4) Trust: As the robot performed lower levels of autonomy, we expected to see a negative response in participants' ratings for trustworthiness. We believed that participants would see the questions asked by the robots under the *Low* and *Medium PA* conditions as indicators that the robot would not be able to perform its tasks without human input. However, there were no observed effects of PA strategy on trust, and participants provided positive ratings for trustworthiness across all three conditions. This suggests that even though participants perceived the robot as having different levels of autonomy, they trusted the robot to complete its tasks.

5) Perceived Intelligence: Finally, we expected participants to provide lower Perceived Intelligence ratings as the robot performed lower levels of autonomy. Similarly to our reasoning for trust, we believed that participants would perceive the robots that asked questions to be less intelligent, as they required human input to carry on with their next task. Yet, our results present no effects of PA strategy on Perceived Intelligence ratings. Overall, participants provided positive ratings for Perceived Intelligence, independently of which PA strategy was used by the robot.

B. Hypothesis Two

H2 stated that increased Latency levels would lead to worse SA, but would not affect any of the other measures. This hypothesis was partially supported.

1) Situational Awareness: We expected participants' SA levels to decrease as latency increased, as participants would get distracted with other tasks while waiting for the robot to complete its previous assignment. However, we did not observe any effects of latency on SA, and participants maintained high SA levels across all conditions. This may have been observed because latency did not affect the rate with which SA questions were displayed, requiring participants to maintain the same level of attention throughout each game, independently of how much latency was experienced.

2) Cognitive Load: As expected, there were no effects of latency on participants' cognitive load. We believe this happened for the same reason listed above for SA: the rate with which participants were performing tasks did not change according to latency. While games with higher latency naturally took longer to be completed, participants performed the same tasks at the same rate across all conditions. Higher latency affected the number of resources participants had in storage for each color, but did not change the game mechanics or the rate with which the robot would communicate with them.

3) Perceived Autonomy: Because the results observed from our analysis were anecdotal and inconclusive, more data needs to be collected to arrive at conclusive results for Perceived Autonomy in terms of latency. However, our analysis indicates that, overall, latency had no effect on participants' perceptions of the robot's levels of autonomy. This trend agrees with [17]'s results for perceived dependency.

4) Trust: Before each game, participants were reminded that any experienced latency was not the robot's fault but rather a product of the robot's remote location. As such, we did not expect any differences in participants' ratings for trust based only on latency. Yet, results show that participants placed less trust in robots operating under higher latency levels. We believe participants may have judged such robots to be less consistent and/or reliable in terms of task completion time, leading to stronger doubts about the robot's ability to quickly accomplish its tasks.

5) Perceived Intelligence: For the same reason listed above for trust, we believed participants would not change their ratings of Perceived Intelligence based only on experienced latency. However, results suggest that participants perceived the robots to be less intelligent when operating under higher levels of latency. This is interesting, as latency only affected the frequency of communication between the participant and the robot, but did not change the way in which the robot was communicating. It could be the case that participants expected the robot to work its way around the latency, although it was a factor beyond its control.

C. Hypothesis Three

H3 stated that we would not observe interaction effects on any of our measures. This hypothesis was partially supported.

1) Cognitive Load: The observed interaction effects for cognitive load indicate that while increased levels of latency led to higher cognitive load in the *Low PA* and *Medium PA* conditions, they actually led to reduced cognitive load levels in the *High PA* condition. When the robot was operating under the *High PA* strategy participants only had to listen to what the robot was going to do next. Instead, in the *Low PA* and *Medium PA* conditions, they were presented with prompts that displayed the robot's question and provided them with the available answer options. Dealing with these questions may have been overwhelming as latency levels increased. On the other hand, the lack of prompts combined with higher latency in the *High PA* condition may have given participants more time to focus on other aspects of the game, reducing their overall cognitive load.

2) Trust: The overall results of our Bayesian analysis present anecdotal evidence against an interaction effect on trust, meaning that we cannot rule out such an effect. The bar plots in Figure 1c show that increased latency levels in the *Low PA* and the *Medium PA* conditions led participants to rate robots as less trustworthy. However, participants rated robots operating under the *High PA* strategy as more trustworthy as latency increased, especially in the *Medium Latency* condition. Given that robots operating under *High PA* always collected the least available resource color, participants might have felt that the robot could still be trusted in situations where latency was present. In addition, the amount of trust that participants placed on the robot in *High Latency* conditions was higher than that of *Low Latency* scenarios, but seems to have diminished compared to the *Medium Latency* condition. Thus, although participants seemed to believe the robot was capable of completing its tasks under high latency conditions, they seemed to become more skeptical over time.

3) Perceived Intelligence: Similarly to trust, our results present only anecdotal evidence against an interaction effect on Perceived Intelligence. The bar plots in Figure 1d show that robots using the *Low PA* and *High PA* strategies were perceived to be more intelligent when operating under *Medium Latency*. On the other hand, robots using the *Medium PA* strategy were perceived to be less intelligent as latency increased. That is, robots that asked Yes/No questions were perceived to be less intelligent than robots that asked for human arbitration and robots that stated their next action. Participants may have interpreted the Yes/No questions as a reflection of the robot's inability to select and confirm a plan of action on its own. The use of statements and the requests for arbitration may have been interpreted as intelligent approaches to collaboration. While the former strategy prioritized collection from the colored pool that had lower numbers, the latter provided participants with more control over what their robot teammate's actions, mitigating the negative impacts of latency in collaboration.

V. CONCLUSION

In this study, we investigated the impacts that Performative Autonomy and Latency have on important factors for collaborative, remote human-robot interaction scenarios. This investigation yielded three main insights. First, robots that use Performative Autonomy to send frequent status updates can boost overall human Situational Awareness. However, if communication is too frequent it might increase human cognitive load to undesirable levels. Second, robots that perform lower levels of autonomy were perceived as less autonomous. Yet, that did not lead participants to regard these robots as less intelligent or less trustworthy. Finally, robots operating under high latency were regarded as less trustworthy and less intelligent, even when they were not directly responsible for the latency itself.

Overall, our results suggest that lower performed autonomy leads to increased cognitive load, especially when robot communication happens frequently and latency is present. In addition, we observe no effect of Performative Autonomy on Situational Awareness, and instead find evidence that operating under high latency leads to negative perceptions of robots regardless of choice of Performative Autonomy strategy. While these findings allow us to better understand the benefits and drawbacks of using Performative Autonomy in collaborative human-robot interaction contexts affected by latency, such as space exploration, there are a few limitations regarding our study that should be addressed by future work.

A. Limitations and Future Work

In this work, our results (in contrast to [13]'s) showed that our use of PA strategies had no impact on SA and actually increased human levels of cognitive load. We believe this may have been the case because in our experimental context robot communication happened much more frequently than in [13]'s. Furthermore, while in [13]'s setup participants needed to switch views between two windows with different information, in our setup all necessary information could be accessed simultaneously. Future work should investigate how these differences in communication frequency, task complexity, and use of PA strategies impact SA, cognitive load, and other important human factors.

In addition, we assumed that the robot was aware of the ideal actions to take, and this information informed the questions that were asked to the human. In order to obtain a complete assessment of PA, future work should investigate scenarios in which the robot may not know the ideal course of action ahead of time. This could lead to scenarios in which humans must supervise robot activity in a much more active way [32]. It could also lead to situations in which the questions asked by the robot are informed by suboptimal plans that could lead to worse task performance and cognitive load levels. On the other hand, it could increase SA if it encouraged human teammates to think about why the proposed plan was suboptimal.

Finally, we cannot rule out the possibility of learning effects between the three games played by participants. For example, it is possible that participants learned to place greater attention on the resource amounts available in anticipation of SA questions. As such, future work should conduct similar experiments in a between-subjects fashion to address participants' views of robots that are based on a single interaction. These experiments should also vary the types of SA questions that are asked to prevent participants from predicting their contents and their timings.

ACKNOWLEDGEMENTS

We would like to thank Michelle Lieng for assisting with experiments and data analysis. This work was funded in part by NASA award 80NSSC25K7246.

References

[1] R. Bogue, "Robots for space exploration," *Industrial Robot: An International Journal*, vol. 39, no. 4, pp. 323–328, 2012.

[2] M. Bualat, J. Barlow, T. Fong, C. Provencher, and T. Smith, "Astrobee: Developing a free-flying robot for the international space station," in *AIAA SPACE 2015 Conference and Exposition*, 2015.

[3] A. Ellery, *Planetary rovers: robotic exploration of the solar system*. Springer, 2015.

[4] T. R. Colin, N. J. Smets, T. Mioch, and M. A. Neerincx, "Real time modeling of the cognitive load of an urban search and rescue robot operator," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, pp. 874–879.

[5] L. Orlov Savko, Z. Qian, G. Gremillion, C. Neubauer, J. Canady, and V. Unhelkar, "Rw4t dataset: Data of human-robot behavior and cognitive states in simulated disaster response tasks," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 924–928.

[6] A. Rajavenkatanarayanan, H. R. Nambiappan, M. Kyrarini, and F. Makedon, "Towards a real-time cognitive load assessment system for industrial human-robot cooperation," in *2020 29th IEEE international conference ON robot AND human interactive communication (Ro-Man)*. IEEE, 2020, pp. 698–705.

[7] K. Guru, S. Shafiei, A. Khan, A. Hussein, M. Sharif, and E. Esfahani, "Understanding cognitive performance during robot-assisted surgery," *Urology*, 2015.

[8] P. Thuot, "International space station (iss) eva suit water intrusion, high visibility close call," NASA, Tech. Rep. IRIS Case Number: S-2013-199-00005, 2013.

[9] D. Cowan and T. Bond, "Coolant leak from iss external active thermal control system (eatcs)–an examination of most probable," in *International Conference on Environmental Systems*, 2022.

[10] M. Dorneich, W. Rogers, S. Whitlow, and R. DeMers, "Human performance risks and benefits of adaptive systems on the flight deck," *The International Journal of Aviation Psychology*, 2016.

[11] M. Dorneich, S. Whitlow, S. Mathan, P. M. Ververs, D. Erdogmus, A. Adami, M. Pavel, and T. Lan, "Supporting real-time cognitive state classification on a mobile individual," *Journal of Cognitive Engineering and Decision Making*, 2007.

[12] L. Onnasch, C. D. Wickens, H. Li, and D. Manzey, "Human performance consequences of stages and levels of automation: An integrated meta-analysis," *Human factors*, vol. 56, no. 3, pp. 476–488, 2014.

[13] S. Roy, T. Smith, B. Coltin, and T. Williams, "I need your help... or do i? maintaining situation awareness through performative autonomy," in *Proceedings of the 2023 ACM/IEEE international conference on human-robot interaction*, 2023, pp. 122–131.

[14] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human factors*, vol. 53, no. 5, pp. 517–527, 2011.

[15] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[16] T. Fong and I. Nourbakhsh, "Interaction challenges in human-robot space exploration," *interactions*, vol. 12, no. 2, pp. 42–45, 2005.

[17] R. Sousa Silva, M. Lieng, E. Muly, and T. Williams, "Worth the wait: Understanding how the benefits of performative autonomy depend on communication latency," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 126–133.

[18] J. Choi, Y. Kim, and S. Kwak, "The autonomy levels and the human intervention levels of robots: The impact of robot types in human-robot interaction," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 2014.

[19] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld, and H. Yanco, "Effects of changing reliability on trust of robot systems," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 73–80.

[20] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000.

[21] G. Matthews, J. Lin, A. R. Panganiban, and M. D. Long, "Individual differences in trust in autonomous robots: Implications for transparency," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 3, pp. 234–244, 2019.

[22] M. Parisi, T. Panontin, S.-C. Wu, K. McTigue, and A. Vera, "Effects of communication delay on human spaceflight missions," in *14th International Conference on Applied Human Factors and Ergonomics (AHFE)*, 2023.

[23] M. R. Endsley, "Situation awareness global assessment technique (sagat)," in *Proceedings of the IEEE 1988 national aerospace and electronics conference*. IEEE, 1988, pp. 789–795.

[24] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[25] R. Sousa Silva, C. Smith, L. Bezerra, and T. Williams, "Toward raps: the robot autonomy perception scale," *arXiv e-prints*, pp. arXiv–2407, 2024.

[26] B. F. Malle and D. Ullman, "A multidimensional conception and measure of human-robot trust," in *Trust in human-robot interaction*. Elsevier, 2021, pp. 3–25.

[27] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International journal of social robotics*, vol. 1, pp. 71–81, 2009.

[28] R. Morey, J. Rouder, T. Jamil, and M. R. Morey, "Package 'bayesfactor'," 2015.

[29] D. Makowski, M. Ben-Shachar, and D. Lüdecke, "bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework," 2019.

[30] M. Hinne, Q. F. Gronau, D. van den Bergh, and E.-J. Wagenmakers, "A conceptual introduction to bayesian model averaging," *Advances in Methods and Practices in Psychological Science*, vol. 3, no. 2, pp. 200–215, 2020.

[31] M. L. Schrum, M. Johnson, M. Ghuy, and M. C. Gombolay, "Four years in review: Statistical practices of likert scales in human-robot interaction studies," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 43–52.

[32] C. Smith, T. Mott, and T. Williams, "Robot, take the joystick: Understanding space robotics experts' views on autonomy," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2024, pp. 2182–2188.