# Show Me What’s Wrong: Impact of Explicit Alerts on Novice Supervisors of a Multi-Robot Monitoring System

Bahodi, Maria-Theresa; van Berkel, Niels; Skov, Mikael B.; Merritt, Timothy Robert

*Published in:*
TAS 2024 - Proceedings of the 2nd International Symposium on Trustworthy Autonomous Systems

*DOI (link to publication from Publisher):*
10.1145/3686038.3686069

*Publication date:*
2024

*Document Version*
Publisher's PDF, also known as Version of record

*Citation for published version (APA):*
Bahodi, M.-T., van Berkel, N., Skov, M. B., & Merritt, T. R. (2024). Show Me What’s Wrong: Impact of Explicit Alerts on Novice Supervisors of a Multi-Robot Monitoring System. In *TAS 2024 - Proceedings of the 2nd International Symposium on Trustworthy Autonomous Systems* Article 4 Association for Computing Machinery (ACM). https://doi.org/10.1145/3686038.3686069

# Show Me What's Wrong: Impact of Explicit Alerts on Novice Supervisors of a Multi-Robot Monitoring System

Maria-Theresa Bahodi
Aalborg University
Aalborg, Denmark
mtoh@cs.aau.dk

Niels van Berkel
Aalborg University
Aalborg, Denmark
nielsvanberkel@cs.aau.dk

Mikael B. Skov
Aalborg University
Aalborg, Denmark
dubois@cs.aau.dk

Timothy Robert Merritt
Aalborg University
Aalborg, Denmark
merritt@cs.aau.dk

## ABSTRACT

With the rise of autonomous multi-robot systems, the role of the robot operator shifts from controlling and observing a single robot to that of a supervisor overseeing multiple robots. Previous studies suggest that timely warnings of problematic events improve a user's ability to monitor multiple robots, however, research has not examined the influence alerts have on user monitoring behavior and their perceptions of the system. We present a 2x2 study design where we manipulate the operator's cognitive load through the number of simultaneous robots under their control and the level of support through the absence or presence of warnings. Our findings suggest that users are more likely to fixate on alerts when shown explicitly, and task difficulty influenced the user's willingness to allow the robots to act autonomously. Our research offers insights for advancing the design of autonomous multi-robot interfaces, emphasizing strategies to enhance the simultaneous monitoring of numerous robots.

## CCS CONCEPTS

• **Computing methodologies → Multi-agent systems**; • **Human-centered computing → Empirical studies in HCI**.

**ACM Reference Format:**
Maria-Theresa Bahodi, Niels van Berkel, Mikael B. Skov, and Timothy Robert Merritt. 2024. Show Me What's Wrong: Impact of Explicit Alerts on Novice Supervisors of a Multi-Robot Monitoring System. In *Second International Symposium on Trustworthy Autonomous Systems (TAS '24), September 16–18, 2024, Austin, TX, USA*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3686038.3686069

## 1 INTRODUCTION

The usage of unmanned vehicles (UVs) has become increasingly widespread, such as ground vehicles (UGVs) used for agriculture [9, 21] and as aerial vehicles (UAVs) or drones used for monitoring of critical infrastructure [43] and search and rescue [36, 37, 56]. While these robots can be used by an operator who manually controls the UV, the technology has rapidly improved to have autonomous capabilities [20] to the point where the research community has explored how to extend UV usage from a single robot-operator relationship to multi-robots [2, 8]. Studies have implied that autonomous systems that utilize remote communication between the robots and the operator will become highly autonomous to the extent that the operators will supervise the system rather than exercise fine-grained control [18].

As the technology matures, errors and failures are bound to happen. Especially in dynamic outdoor environments, failures can be tedious, requiring the operator to attend physically to resolve the issue. When robots are spread over large work areas, as in agriculture, operators spend a lot of time reaching a failed robot, which impacts their availability to focus on other tasks. Furthermore, delays in reaching a failed robot could lead to detrimental outcomes in the task or damage to the robot. An abundance of failures can also garner user mistrust and can harm the relationship between the vehicles and operators [28, 40], ultimately lowering the willingness to adopt these autonomous systems. Considering that the operator's role is evolving into more of a supervisor, this raises concerns about how many robots a person can monitor simultaneously, especially in the likelihood of failures and errors. Monitoring can involve observing a robot's state and notifications about its immediate and potential problems. Research suggests that the maximum number of robots a person can maintain is primarily dependent on the number of robots and the number of steps it takes for the user to remove the robot from the dangerous situation [22, 61]. Research has shown that timely warnings and alerts significantly enhance a user's ability to respond to situations while operating a vehicle. These alerts provide immediate visual feedback, alerting the user to focus fully on the vehicle and allowing them adequate time to respond to any emerging issues [10].

While evidence suggests that alerts can positively impact the user's performance, insufficient support for identifying, prioritizing, and handling failures can have a detrimental effect on attitudes toward trusting and adopting automation. There is an incentive to understand how alerts can improve monitoring performance and identify how the monitoring interface's design can negatively affect a user's strategy and interactions. In the study presented in this paper, we explore ***how participant behavior in monitoring***

*TAS '24, September 16–18, 2024, Austin, TX, USA*
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0989-0/24/09
https://doi.org/10.1145/3686038.3686069

***a group of robots changes when explicitly notified about robot misbehavior.*** To do this, we created a platform designed after existing multi-robot applications and gamified the experience by incorporating a point system to incentivize the participants to diligently monitor all the robots and correct problems as they arise. We conducted a study with non-experts gathering performance metrics and self-reported feedback and interviewed them to understand their beliefs and attitudes toward automation.

## 2 RELATED WORK

Our work builds on prior research on the design of autonomous systems. We outline relevant literature on the monitoring of multi-agent systems and related topics, the design of monitoring systems, and the impact of user trust on the usage of these systems.

### 2.1 Monitoring Multi-Robot Systems

Prior work has explored the monitoring capabilities of operators when it comes to engaging with multi-robot systems [17, 80]. Effective interface design for a multi-robot system supports situation awareness [30], both in routine work and when faced with instances of unexpected failures in the robots' automation [88]. An interface needs to provide enough information so that the user is not missing information to make an informed choice, but simultaneously not too much information so that the operator is overloaded and cannot find the appropriate information [69]. Here, visual saliency can help an operator to easily differentiate essential and urgent information from routine information [71]. While studies have discussed how to display information in a multi-robot system [2, 3] as well as how people react to faulty physical robots [27, 73], there is still an incentive to understand how these two research topics relate to one another, i.e. how to display information when robots are exhibiting unintended behavior. In recent years, research has been exploring models and techniques for fault detection and diagnosis [50, 51], and have taken a more technical perspective of how to mitigate and resolve faulty behavior with limited to no involvement of potential users in their studies.

*2.1.1 Supervisory Role in Automation System.* In the monitoring of semi-autonomous multi-robot systems, the user's role is moving towards becoming a supervisory role rather than continuous interaction with the robots [72]. Various challenges arise when supervising a group of robots, including operator fatigue and boredom [24, 38] and situation awareness, cognitive load, and trust in the automation [66]. While there have not been many studies that directly explore how a person continuously monitors active robots, research has addressed how users monitor other groups of entities, such as in the field of air traffic control (ATC) [12, 31]. Especially in ATC research, workload and situation awareness (SA) are prevalent topics, as the air traffic controller (ATCO) needs to be aware of the current space to ensure safety among the aircraft [32, 59, 63]. Research has indicated that there is a difference in performance when exposed to different levels of task loads [47]. However, Friedrich et al. [32] found in their study that when using probing techniques to measure SA, there was, in fact, no such difference. That said, participants in this study did not find a probing technique appropriate to measure SA since ATCOs have a more holistic view of SA, and a probing technique tries to break down SA into specific questions of specific events.

*2.1.2 Fan-Out Model.* Olsen and Wood developed an equation to predict the maximum number of robots a single operator can manage simultaneously. This equation is known as the *fan-out* model [22, 61]. In this model, the neglect time is the amount of time a robot can be neglected before it needs attention, while the interaction time is the time it takes to interact with a robot before it no longer needs attention. Cummings et al. [25] further developed the equation such that wait time (WT) was taken into account. They defined three WT sources: the queue of problems to be resolved before moving to another problem, the time needed to regain situation awareness when completing one task and moving to another, and the time needed to take corrective action and resolve a problem. By incorporating WT, they found that the number of unmanned vehicles an operator could manage dropped by 36%–67%, highlighting the importance of accounting for the limitations of human decision-making. Later, Breslow et al. [10] adapted the fan-out model to predict the likelihood of a supervisor failing to prevent damage to an unmanned vehicle and using the prediction to alert the supervisor appropriately. As a result of their study, they also highlighted the importance of timely feedback – stressing that an alert that appears too late will lose its usefulness. The research involving the fan-out model displays how performance increases when alerts are displayed in a timely manner, and they provide detailed descriptions of the interfaces used. However, their results did not focus on the designs and how to improve them to support the users, and a gap persists in how the different display approaches affect a person's interaction strategies while monitoring a group of robots.

### 2.2 Interface Design for Monitoring Systems

The literature highlights various ways of designing interfaces to support the monitoring of multi-robot systems. A common component is a view of the environment such as a geographical map [19] or a digital representation of the traversable space [5, 80] – an often-seen functionality is to see all robots related to the system and their current position on a map. Another commonly incorporated element is an information panel, often placed beside the map, which can provide more information on the status of individual robots [26, 70, 71]. Such a menu allows for information to be displayed without cluttering the map, and more oversized items such as video feeds can be displayed [17].

Other studies have sought to establish user needs concerning the interface design of multi-robot monitoring systems. For example, Agrawal et al. [2] conducted participatory design studies with firefighters to develop a multi-drone system for search and rescue (SAR) purposes. They presented the challenges related to drone usage in SAR missions and provided design recommendations to address these challenges through interface modifications, for example, by giving the user the ability to see the previous and planned path of the drone. Rule and Forlizzi [71] had previously laid out design recommendations for providing appropriate salience of information such that the operator would not be cognitively overloaded. Here, they provided an interface separated into cells where each cell was meant for different levels of awareness and salience. Using contextual inquiry with specialized robot operators, they found multiple scenarios that needed to be addressed with different salience, such as being able to differentiate important information from urgent ones.

Appropriate salience is also a prevalent topic in notification design in monitoring systems. Earlier research has sought to understand how notification and visual indicators can be designed to grab the operator's attention without distracting them from their main task [55, 82]. Imbert et al. [42] conducted a study where they created the **LABY microworld** to explore the issues of notification on larger screens in an air traffic control scenario. Here, they found that notification that involved movement with a box-animation alert would allow the operator to notice it without losing sight of their primary monitoring task. While the existing work thoroughly examines how domain experts interact with these systems, only a few examine how novices or non-domain experts interact with a monitoring system. With the fast adoption of the technology in diverse areas, it is not unlikely that multi-robot systems will be used by people that are not necessarily experienced with the technology or work domain. Hence, we want to expand the research topic and explore similar design approaches to existing work are received and perceived differently when the end-users are novices.

## 2.3 User Perception of Autonomous Agents

Automation is a key aspect of multi-robot systems, as it is typically unfeasible for a single operator to effectively control an arbitrary number of robots [38]. However, prior work has identified differences in how willing users are to adapt to automation [39]. Critically, some users might be too eager to utilize the autonomous behavior of systems without considering the value and potential downsides of this approach [13]. In contrast, others can be overly critical of the capabilities of the system and thereby miss out on its potential benefits [77].

*2.3.1 Lack of Trust.* There are various reasons why the potential users of a system are critical of automation and tend to be skeptical of the results and recommendations it provides. Research has indicated that the experience of false positives, such as false alarms, decreases a person's trust in a system [6, 11, 54, 76, 85]. Gupta et al. [34] looked at how early and late alarms of upcoming skids affected a user's trust in automation while driving and found that a higher rate of early alarms led to lower levels of confidence compared to late alarms. The authors hypothesized that users perceived the early alarms as the system not being able to determine what a dangerous situation was accurately. Failure in autonomy also affects how a person perceives the capabilities of the automation and how they interact with the system [53, 73]. The work of Moralez et al. [60] explored how willing a person was to help a robot if they were shown robotic failures prior to interacting with the system. In their study, they found that if participants were not exposed to any failures, 81.3% of them were willing to help the robot, but when they had previously seen the robot perform two major errors, only 59.4% assisted it. Gradual exposure to failures also resulted in not wanting to help at the end of the experiment, with some reporting that it felt as if the robot did it on purpose. The studies point towards the fact that prolonged exposure to failure affects a person's willingness to help, which has also been shown to be present in human-swarm research [4]. However, recent research has pointed out that there is a need to understand if self-correction also affects how people perceive the robots' abilities [16].

*2.3.2 Complacency & Automation Bias.* User complacency and automation bias can similarly be issues in the adoption of autonomous systems. Complacency originally referred to insufficient pilot monitoring in aircraft subsystems. With the increase in automation, this concept expanded and became known as *automation-related* complacency [57, 64]. This complacency can lead to accidents as people do not carefully monitor the system in question. The term is more often used for alerting systems rather than systems in which automation acts as decision support [35, 65]. On the other hand, automation bias is often used in the context of decision aid, in which the automation provides wrong advice and the user fails to notice the mistake, resulting in them following an incorrect recommendation [78, 85]. In human-swarm interaction (HSI) research, there has been some indication that complacency is also present in these systems [86]. When a lack of trust could be a result of a faulty system, overtrust would be present when the system showed high reliability [41].

*2.3.3 Time-Critical Decision-Making.* In dynamic environments, time often plays a key role in the decision-making process as there is a lack of time to react to the situation and make a well-thought-out decision [75]. Time pressure will likely affect a person as the stressor of running out of time leads to the fear of failing to complete a task [62, 89]. The research community has extensively explored how stress affects user interaction and how decision support systems (DSS) can help users overcome errors in these situations [1, 67, 74]. DSSs provide the user with the information that could support their decision for a task. This, however, again runs the risk of introducing automation bias [23]. Prior work further shows that DSSs usefulness is reliant on the task and person [84] and implies that less decision time also leads to more usage of the recommendation of the decision support system [68, 81]. Though, Cao et al. [15] suggest in their work that the participants are in fact more likely to follow the automation of the support system when they are given more time.

# 3 METHOD

We conducted a within-subjects study across four conditions. Participants were asked to monitor a group of simulated robots and strive to have them follow their planned path as much as possible. At the end of each monitoring task, participants answered questionnaires to measure cognitive load and their perceptions of the UVs. After the final task, we conducted a semi-structured interview to gather their thoughts and opinions about the experience.

## 3.1 Study Design

We designed the study around monitoring multiple robots that can show problematic behavior throughout their monitoring task and simplified the capabilities of the robots. In the monitoring task, the robots could only follow a pre-planned route, which they could divert as this is a problem that can occur in the real-world. Whenever a robot diverts from its planned route, we define its state as being in a *degraded* state. If the robot does not find its way back to its original path, there is a risk of it entering a *failed* state, i.e., it has fully stopped.

*3.1.1 Conditions.* We manipulated two independent variables to explore the influence of explicit alerts in multi-robot systems. The variables in question are:

(1) **Cognitive load (Number of robots)**: Each condition presented a fixed number of robots - three robots for the low cognitive load conditions and six robots for the conditions with high cognitive load. We based this on the original **fan-out model** [61] that suggests an operator can reasonably monitor up to three robots, but above this number, the operator would find the monitoring task very challenging. We chose to increase the number of robots yet keep the overall number of events the same; thus, in both conditions, every 5 s, another robot would enter a degraded state. This is similar to the study by Breslow et al. [10] where every 4 s, a threat was randomly selected to appear in a new position, causing damage to any UAV in that position. By holding the number of events constant in their study, they avoided the wait time queue caused when multiple events occur simultaneously [25]. We follow this approach in our study design. The increase in the number of robots on the map was intended to increase the workload for the participants, requiring them to scrutinize more locations to check for degraded robots. This choice meant that in the high load condition, the number of errors overall remained constant, yet the number of errors per robot declined.

(2) **Level of Support (Displaying warnings)**: To see if explicit warnings affected the ability to perceive and react to robots in a degraded state, we designed a binary condition of alert salience with warnings displayed in the robot list on the side in addition to the visible deviation on the map (high support) and the (low support) condition in which no warnings were displayed in the robot list in addition to the deviations shown on the map.

We conducted a 2x2 study with the variables, resulting in four conditions the participant needed to complete. We used a balanced Latin square to mitigate ordering bias.

*3.1.2 Platform.* The study platform was implemented with React and TypeScript and was developed to be used on a PC with a screen size of at least 14". While we were inspired by several interfaces from multi-robot research [2, 38], we were particularly inspired by the map interface in the study by Breslow et al. [10], in which participants monitored simulated robots by observing their position on a map and interacted with them through a side menu beside the map. In our study platform, robots move around on a map following snake-like patterns (see Figure 1) typical for many coverage tasks such as drone mapping missions and agriculture robots. Icons indicating *Normal, Warning, or Error* states were displayed in the robot list. Apart from seeing the position of the robots and any deviations from their planned path on the map, the icons were the only other means to assess the state of each robot. All participants interacted with the same robots and paths specific to the different study conditions.

*3.1.3 Robot Behavior.* We programmed the UVs to move slowly and usually follow the planned path displayed on the map. To introduce realistic tasks, the system randomly selected a robot to enter a degraded state every 5 seconds, meaning it would begin to deviate from its planned path. To make the degraded state noticeable, the angle of the deviation was always $15^\circ$ from the planned path. After a robot entered this degraded state for 6 seconds, it would either enter the *Error* state and stop completely or redirect itself back to its planned path and resolve the degraded state. Alternatively, participants could intervene and manually redirect a robot in the degraded state back to the planned path.

*3.1.4 Redirecting a Degraded Robot.* When a user wants to resolve a robot in a degraded state manually, they have to select the robot in the robot list and then click on a point on the robot's planned path. A teal path displays the trajectory from the robot's position to the selected point, and the robot then begins moving toward the point and rejoins the path. The type of redirection was made visible through the path color on the map – automatic redirections resulted in a path of the robot's assigned color, and manual redirections resulted in a teal path, as shown in Figure 3. We intended to make it noticeable to the participant through the map of the various moments in the task where the robot corrected itself and where participants intervened to redirect manually.

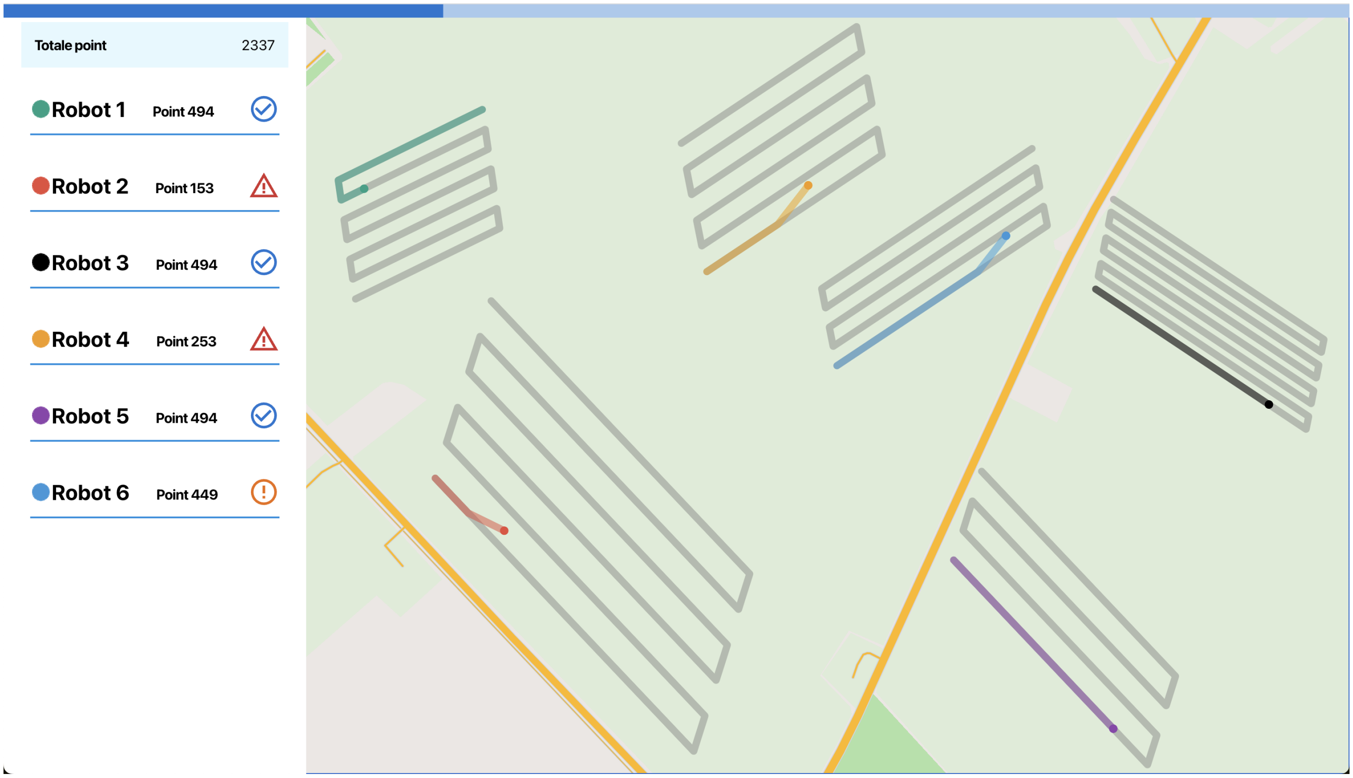
*3.1.5 Point System.* To measure the effectiveness of the participant actions in monitoring and correcting robots in degraded or error states, we incorporated a point system where they would gain 10 points every second a robot followed the planned path. If the robot entered a degraded state, it would earn only 5 points per second, and if the robot fully stopped (failed state), no points would be gained by that robot until the participant navigated it back to a point on the planned path. The point system was explained to the participants to incentivize them to be vigilant and work diligently during the study. The accumulated points were displayed during each task, as well as the points the individual robots had collected.

## 3.2 Participants

We recruited 22 participants (13 male and 9 female) who were primarily students from the university whose ages ranged from 19 to 28 (M = 23.68), and none had difficulty with differentiating colors. The number of participants was chosen based on the local standards for in-person within-subject experiments [14]. All participants were fluent in Danish, and 10 out of the 22 participants reported having some experience with autonomous robots, mostly home appliances such as robot vacuum cleaners and lawnmowers. The participants were compensated for their participation and were told that they could withdraw their data if they had any doubts regarding their participation.

## 3.3 Procedure

The study was conducted in a meeting room to ensure that the participants were not disturbed during the study. The study took around 30-40 minutes per participant, where at least 20-25 minutes were dedicated to using the study platform. We gave the participants a consent form, in which we informed them that they could withdraw their participation at any time. Afterward, we asked the



Totale point 2337

Robot 1 Point 494 ✓

Robot 2 Point 153 ⚠

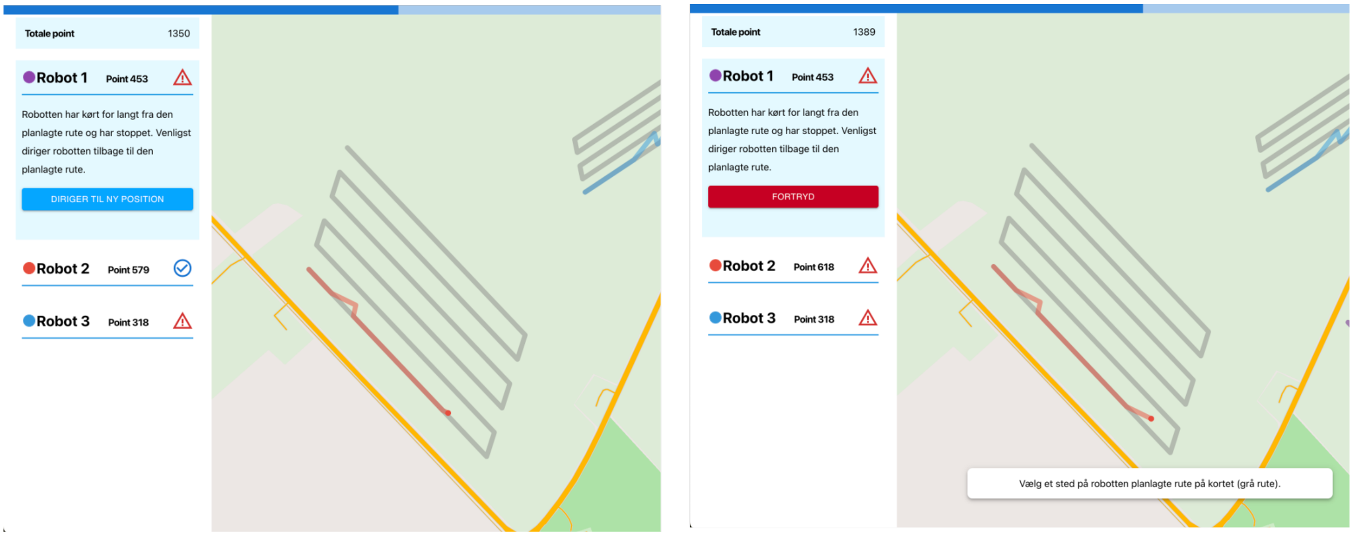
Robot 3 Point 494 ✓

Robot 4 Point 253 ⚠

Robot 5 Point 494 ✓

Robot 6 Point 449 !

**Figure 1: Screen capture of the study platform specifically for the condition with many robots and a high level of support. On the side menu, three different icons are shown representing the different states. The blue icon with a tick represents a robot acting normally, the orange icon with an exclamation point alerts that the robot is in a degraded state, and the red triangle is shown when the robot has fully stopped, i.e., it has entered a faulty state. In conditions with low support, the orange icon would not be displayed.**



Totale point 1350

Robot 1 Point 453 ⚠

Robotten har kørt for langt fra den planlagte rute og har stoppet. Venligst diriger robotten tilbage til den planlagte rute.

Robot 2 Point 579 ✓

Robot 3 Point 318 ⚠

Totale point 1389

Robot 1 Point 453 ⚠

Robotten har kørt for langt fra den planlagte rute og har stoppet. Venligst diriger robotten tilbage til den planlagte rute.

Robot 2 Point 618 ⚠

Robot 3 Point 318 ⚠

Vælg et sted på robotten planlagte rute på kortet (grå rute).

**Figure 2: On the left side menu, when a robot is selected, the item expands providing more detailed information about the robot's state a button that can be clicked to activate manual redirection.**

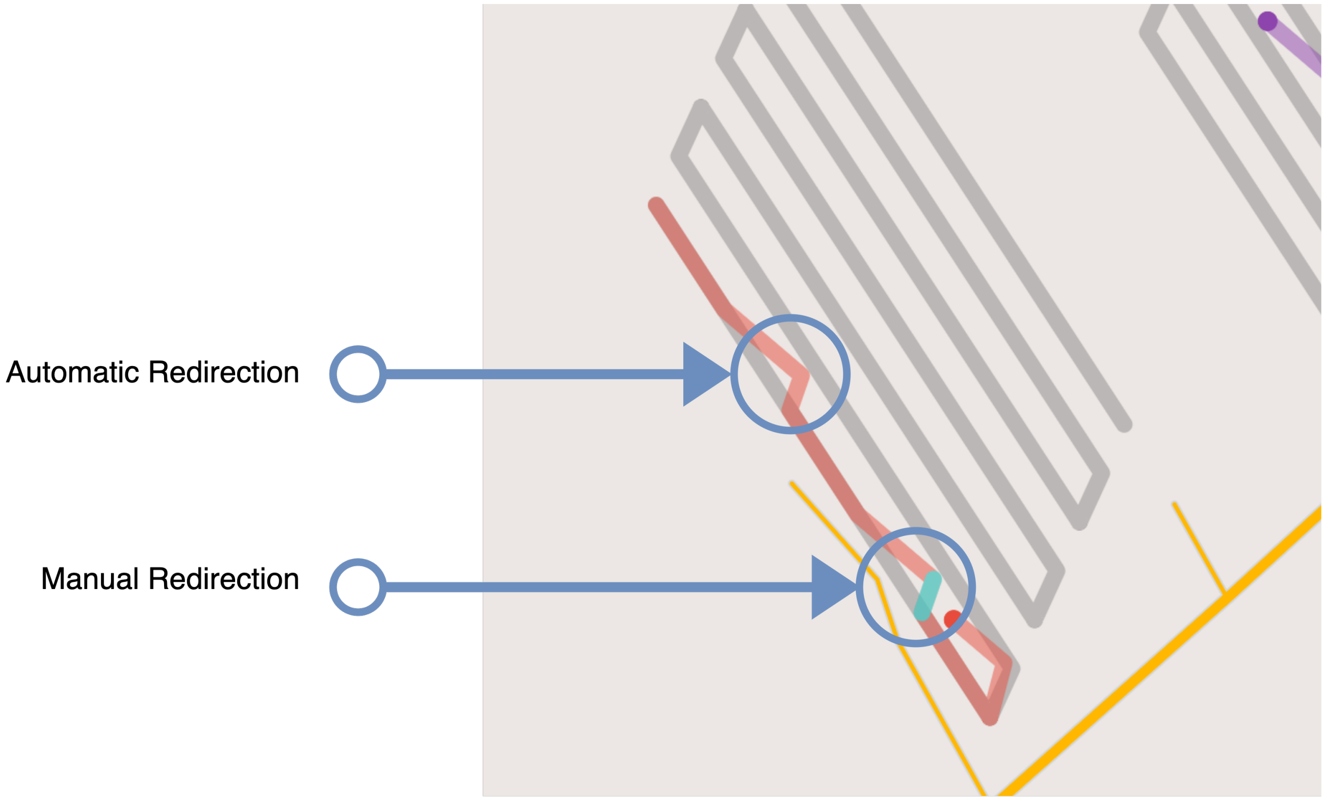


**Figure 3: Example of a robot diverting from its planned path and how it has been redirected. Automatic redirection: The robot has automatically resolved its degraded state, showing the actual path with the robot's color. Manual redirection: The participant has selected a point in the path the robot will move toward. The robot's path from the redirection is shown in teal.**

participants about their experience with autonomous robots to gather information about their opinions of these devices. We did not give much information about what the study was about other than that they would be instructed to monitor a group of robots doing some tasks, as we wanted the participants to complete the study without prior knowledge.

Participants were seated at a desk with a 14" MacBook Pro with a browser in full-screen mode presenting the system. The system prompted them to fill in their demographic information, followed by a tutorial showing how the system worked and details about the upcoming tasks in the study. The tutorial displayed a map of two robots following planned paths as shown in Figure 1 with an addition of a guided tour of the interface. They were instructed to monitor the robots and ensure that they stayed on their planned path as closely as possible to collect the highest amount of points. We decided to introduce warnings in the tutorial as we wanted to ensure that they understood how they could identify when a robot was following the planned path and the ways to determine when it has deviated from the path by either seeing that the robot has deviated from the path on the map (see Figure 3) or by the warnings provided in the side menu (see Figure 1). While the tutorial showed the warnings, participants received additional instruction before each condition that would explicitly state how the participants could determine if a robot was in a degraded state by indicating that the participant could look for warnings and the map or only the map without warnings.

Each condition took 2.5 minutes to complete and was followed by a questionnaire. The questionnaires were a combination of Klepsch et al.'s naïve rating questionnaire for reporting cognitive load [52] and three questions inspired by the work of Gilad et al. [33] about how well people perceived an AI system given its warmth and competence. The questionnaire was translated into Danish from the original languages (German and English). While we instructed the participants to answer as they understood the statements, we did not refrain from clarifying them if the participants made it clear that they did not understand the statement.

After completing all four conditions and questionnaires, we conducted a semi-structured interview about their experience with the different conditions and how their approaches to completing the tasks differed from condition to condition.

### 3.4 Data Analysis

Objective measures were gathered to support analysis of the task performance (see Table 1). Using the processed data, we checked the data's normality, and in cases where the data was not normally distributed, we made use of non-parametric tests. If the data showed any significant difference between the different conditions, we conducted a post-hoc test using the Wilcoxon test with Bonferroni correction to determine where the significance lay in the data. For the questionnaire, we used the Aligned Rank Test (ART) [87] to test for main effects and interactions between conditions. For the free-form responses given both in the free text at the end of each questionnaire and the post-interviews, we used Nvivo to conduct a thematic analysis of frequent topics and themes the participants mentioned.

# 4 RESULTS

Nearly all participants completed the study and attempted to resolve robots in degraded states. However, one participant chose not to resolve robots in degraded states and preferred to watch and not intervene in the robot tasks. We removed this participant's data from the final dataset, yielding the remaining dataset, which consists of 21 participants.

## 4.1 Performance

We first assess participant performance across metrics such as the number of errors generated by the robots, the number of times the robots went off-track, total points, and the percentage of automatic and manual resolves. We performed the Shapiro-Wilk test to test for normal distribution and found that none of the data points for each metric were normally distributed. We decided to use the Friedman test to find significant differences in the conditions. If the result showed significance, we performed a Wilcoxon test with Bonferroni correction to find which specific conditions differed significantly.

*4.1.1 Errors.* A Friedman test shows that the number of robots entering into a failed state, i.e., went into a full-stop, was significantly affected by our manipulation of cognitive load and support ($\chi^2(3) = 16.695, p =< .001$). We, therefore, conducted a post-hoc analysis using Wilcoxon signed-rank tests with Bonferroni correction. The results of our post-hoc tests are shown in Table 2. We visualize our results as split out by the two predictors in Figure 4.A. The median values for conditions with a high level of support ($median = 0.0$) were lower than those with a high cognitive load and low level of support ($median = 2.0$). There were significant differences between the conditions high cognitive load/high support and high cognitive load/low support ($W = 21.5, p = .016$) and high cognitive load/low support and low cognitive load/high support ($W = 8.0, p = .010$).

*4.1.2 Off-track.* The number of robots going off-track, thereby entering a degraded state, showed significant differences between the conditions ($\chi^2(3) = 8.346, p = .039$). However, after conducting the post-hoc test, the differences between the conditions were insignificant. This suggests that the behaviors of the robots across the conditions were consistent, which was expected because the behaviors of all robots followed the same logic and had the same likelihood of entering a degraded state.

*4.1.3 Automatic Resolves.* The proportion of robots automatically resolving their degraded state showed a statistically significant difference depending on the study conditions ($\chi^2(3) = 38.957, p =< .001$). The condition with high cognitive load and low support had a much higher median value ($median = 18.182$) compared to the other conditions ($median = 0.0$), as shown in Figure 4.C. After performing the Wilcoxon test with Bonferroni correction, the results shown in Table 3 indicate that the condition with high cognitive load and low support were significantly different from the remaining three conditions: high cognitive load/high support ($\chi^2(3) = 6.5, p = .001$), low cognitive load/low support($\chi^2(3) = 1.0, p = .001$) and low cognitive load/high support ($\chi^2(3) = 0.0, p =< .001$).

*4.1.4 Manual resolves.* Performing the Friedman test, we found that there was no significant difference in the number of manual resolves of robots in a degraded state ($\chi^2(3) = 5.283, p = .152$). This indicates that the participants were consistent across conditions in their decision to manually navigate a robot back on its planned path when it had gone off-track.

## 4.2 Timing

We look at the metrics related to measured time, such as how long a robot was in a degraded state or how long it took for participants to manually direct a robot back to its planned path when it entered a degraded or failed state. The same statistical analysis was performed on the timing metrics as for the performance metrics.

*4.2.1 Error time.* The Friedman test results showed that there was a significant difference across conditions for how long the robots were in a failed state ($\chi^2(3) = 10.641, p = .014$). The results from the Wilcoxon test with Bonferroni adjustments shown in Table 5 indicate that only the conditions with high cognitive load/high support and high cognitive load/low support are significantly different ($W = 3.0, p = .028$). Figure 5.A also shows that the accumulative time robots were in a failed state was higher for some participants when there was low support ($median = .400$) compared to the opposite ($median = 0.0$). However, the median values do not imply such a big difference, and the significance could be a result of individual participant skills rather than being a consistent effect.

*4.2.2 Off-track time.* From the Friedman test results, there was a statistically significant difference between the conditions ($\chi^2(3) = 17.743, p =< .001$) and the post-hoc test reveals that the conditions with high cognitive load/low support were significantly different from the remaining three conditions as shown in Table 6. From Figure 5.B, this difference is seemingly due to the robots being longer off track in total compared to the remaining conditions, indicating that it either took longer for the participants to notice that a robot had entered a degraded state, but it could also be because the participant navigated the robot to a point that was far away from their position at the time, thereby leading to the robot being outside its planned path for a longer duration.

*4.2.3 Reaction to Error.* A significant difference was found for the metric representing how long it took for a participant to move a robot back to its planned path after it had entered a failed state ($\chi^2(3) = 9.769, p = .021$). However, the post-hoc test revealed that after performing the Wilcoxon test with Bonferroni corrections, there was no significant difference between the different conditions. This implies that the amount of time the robots stayed in a failed state did not differ much between conditions.

*4.2.4 Reaction to Off-Track.* We discovered a significant difference in how long it took for participants to resolve a robot that had gone into a degraded state. The Wilcoxon-test with Bonferroni correction displayed that the differences could be found for the

| **Performance Metric** | **Description** |
|---|---|
| Number of errors | Number of times the robots went into the failed state (full stop). |
| Number of off-track | Number of times the robots went into a degraded state (off-track). |
| Percentage of auto-resolves | The number of times the robot auto-resolved its degraded state out of the total amount of times the robot went out of a degraded state. |
| Percentage of manual resolves | The number of times the participant resolved a robot's degraded state out of the total amount of times the robot went out of a degraded state. |
| Total points | The total number of points gathered in the session |
| Time error | The accumulative number of seconds the robots were in a failed state. |
| Time off-track | The accumulative number of seconds the robots were in a degraded state. |
| Time to react to error | The average amount of seconds it took until the participants resolved a robot's failed state. |
| Time to react to off-track | The average amount of seconds it took until the participants resolved a robot's degraded state. |

**Table 1: Captured metrics in the processed data.**

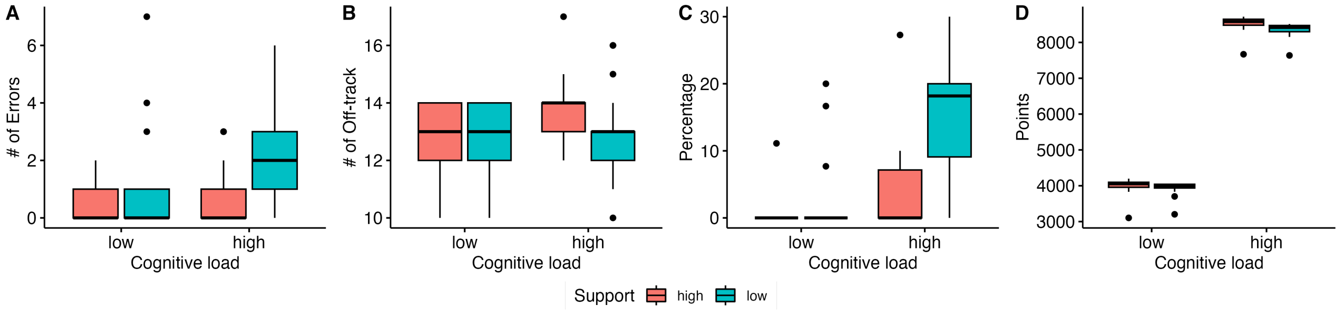


**Figure 4: Boxplots of the performance metric data that showed significance in the Friedman test. The boxplots shown are the following: A. Number of Errors, B. Number of Off-tracks, C. Percentage of Auto-Resolves, and D. Total Points.**

| **Comparison** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Cogn. load** | **Support** | | **Cogn. load** | **Support** | **Mean Dif.** | **SE** | ***W*** | ***p*-value** |
| High load | High support | - | High load | Low support | -1.524 | .440 | 21.5 | **.016** |
| | | - | Low load | High support | 0.0 | .320 | 22.0 | 1.0 |
| | | - | | Low support | -0.476 | .496 | 19.0 | 1.0 |
| | Low Support | - | | High support | 1.524 | .406 | 8.0 | **.010** |
| | | - | | Low support | 1.048 | .537 | 34.5 | .264 |
| Low load | High support | - | | Low support | -0.476 | .456 | 20.5 | 1.0 |

**Table 2: Post-hoc analysis for Number of Errors using Wilcoxon test with Bonferroni adjustments. P-values displayed after performing Bonferroni correction and values where p < .05 are highlighted with bold.**

condition with high cognitive load/low support and the remaining three conditions, but also between the two conditions with low cognitive load ($W = 34.0, p = .020$). As shown in Figure 5.D, we can see that this difference was due to the response time to a degraded robot being longer for conditions with low support.

## 4.3 User Perception

We assess the self-reported feedback the participants provided in the questionnaires that were given after every condition, as well as their responses in the post-interview after completing the study.

*4.3.1 Cognitive load.* We report the significant effects of the cognitive load and level of support uncovered when using the ARTool [87]

| Comparison | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Cogn. load** | **Support** | | **Cogn. load** | **Support** | **Mean Dif.** | **SE** | ***W*** | ***p*-value** |
| High load | High support | - | High load | Low support | -12.964 | 2.621 | 6.5 | **.001** |
| | | - | Low load | High support | 2.756 | 1.903 | 0.0 | .163 |
| | | - | | Low support | 1.173 | 1.934 | 11.0 | 1.0 |
| | Low Support | - | | High support | 15.720 | 2.096 | 0.0 | **<.001** |
| | | - | | Low support | 14.137 | 2.296 | 1.0 | **.001** |
| Low load | High support | - | | Low support | -1.583 | 1.303 | 2.0 | 1.0 |

**Table 3: Post-hoc analysis for Percentage of Auto-Resolves using Wilcoxon test with Bonferroni adjustments. P-values displayed after performing Bonferroni correction and values where p < .05 are highlighted with bold.**

| Comparison | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Cogn. load** | **Support** | | **Cogn. load** | **Support** | **Mean Dif.** | **SE** | ***W*** | ***p*-value** |
| High load | High support | - | High load | Low support | 171.857 | 85.215 | 10.0 | **<.001** |
| | | - | Low load | High support | 4526.857 | 94.774 | 0.0 | **<.001** |
| | | - | | Low support | 4587.476 | 85.864 | 0.0 | **<.001** |
| | Low Support | - | | High support | 4355.000 | 88.162 | 0.0 | **<.001** |
| | | - | | Low support | 4415.619 | 77.889 | 0.0 | **<.001** |
| Low load | High support | - | | Low support | 60.619 | 85.960 | 61.0 | .357 |

**Table 4: Post-hoc analysis for Total points using Wilcoxon test with Bonferroni adjustments. P-values displayed after performing Bonferroni correction and values where p < .05 are highlighted with bold.**

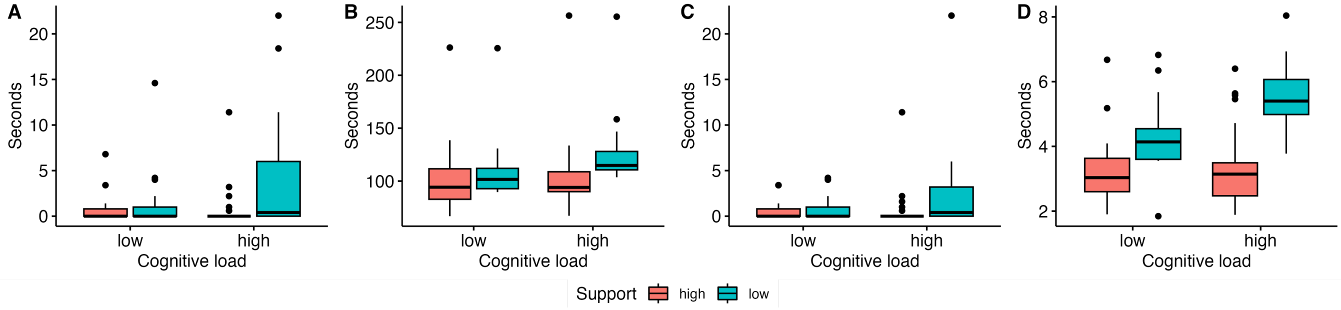


**Figure 5: Boxplots of the timing metrics are displayed in the following order: A. Time error, B. Time off-track, C. Time to react to error and D. Time to react off-track.**

on our collected data. We only report the results for the statements where a significant main effect or interaction was shown in the participant responses. See Figure 6 for a table of the aggregated responses.

We found that for the first statement, a significant main effect of the cognitive load $(F(1,60) = 20.156, p =< .001)$ and level of support $(F(1,60) = 21.635, p =< .001$, which indicates that there was a measurable impact on the participants' answers given the increased cognitive load and support. The same was apparent for the second statement, where the cognitive load $(F(1,60) = 15.300, p =< .001)$ and level of support $(F(1,60) = 13.465, p =< .001)$ also showed significant main effects. As shown in Figure 6, when the cognitive load is higher, participants tended to agree more with the statement, especially in the condition where they were exposed to high cognitive load and low support. The opposite was the case when there was a higher level of support, where the participants would mostly disagree. There was, however, no significant interaction between the two effects for the first $(F(1,60) = .957, p = .332)$ and second statement $(F(1,60) = .023, p =< .001)$, indicating that cognitive load and level of support do not significantly affect one another.

We found that for statement five, the level of support had a significant main effect on the participants' responses on how exhausting it was to find important information in the task $(F(1,60) = 14.662, p =< .001)$. This indicates that the participants found it less

| Comparison | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Cogn. load** | **Support** | | **Cogn. load** | **Support** | **Mean Dif.** | **SE** | ***W*** | ***p*-value** |
| High load | High support | - | High load | Low support | -3.295 | 1.718 | 3.0 | **.0284** |
| | | - | Low load | High support | 0.210 | .631 | 17.0 | 1.0 |
| | | - | | Low support | -0.533 | .949 | 20.0 | 1.0 |
| | Low Support | - | | High support | 3.505 | 1.581 | 9.0 | .065 |
| | | - | | Low support | 2.762 | 1.664 | 21.0 | .521 |
| Low load | High support | - | | Low support | -0.743 | .848 | 18.0 | 1.0 |

**Table 5: Post-hoc analysis for Time error using Wilcoxon test with Bonferroni adjustments. P-values displayed after performing Bonferroni correction and values where p < .05 are highlighted with bold.**

| Comparison | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Cogn. load** | **Support** | | **Cogn. load** | **Support** | **Mean Dif.** | **SE** | ***W*** | ***p*-value** |
| High load | High support | - | High load | Low support | -19.686 | 14.689 | 23.5 | **.004** |
| | | - | Low load | High support | 5.276 | 15.274 | 73.0 | .882 |
| | | - | | Low support | -3.400 | 14.231 | 83.0 | 1.0 |
| | Low Support | - | | High support | 24.962 | 13.882 | 15.0 | **<.001** |
| | | - | | Low support | 16.286 | 12.761 | 29.0 | **.010** |
| Low load | High support | - | | Low support | -8.676 | 12.904 | 66.0 | .533 |

**Table 6: Post-hoc analysis for Time off-track using Wilcoxon test with Bonferroni adjustments. P-values displayed after performing Bonferroni correction and values where p < .05 are highlighted with bold.**

| Comparison | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Cogn. load** | **Support** | | **Cogn. load** | **Support** | **Mean Dif.** | **SE** | ***W*** | ***p*-value** |
| High load | High support | - | High load | Low support | -2.112 | .403 | 9.0 | **<.001** |
| | | - | Low load | High support | 0.177 | .478 | 90.0 | 1.0 |
| | | - | | Low support | -0.870 | .393 | 52.0 | .158 |
| | Low Support | - | | High support | 2.289 | .387 | 0.0 | **<.001** |
| | | - | | Low support | 1.242 | .363 | 21.0 | **.003** |
| Low load | High support | - | | Low support | -1.047 | .369 | 34.0 | **.020** |

**Table 7: Post-hoc analysis for Time to react to off-track using Wilcoxon test with Bonferroni adjustments. P-values displayed after performing Bonferroni correction and values where p < .05 are highlighted with bold.**

challenging to find the relevant information in the task when an explicit warning was displayed.

*4.3.2 Perception of autonomy.* When it comes to the participants’ perception of the robots’ autonomous behavior, we found that cognitive load has a significant main effect on their responses for statements 8 ($F(1, 60) = 9.403, p = .003$) and 9 ($F(1, 60) = 9.575, p = .003$). From their responses, which are visually represented in Figure 7, participants agreed more to the statements when the cognitive load was high compared to when it was lower. This indicates that for the conditions with a higher number of robots, the participants perceived the robots to have more capable automation.

*4.3.3 Free-Text and Post-Interview Responses.* Qualitative data were collected through free-text responses written at the end of each questionnaire and contextual post-interviews. We cover the main themes of their responses and beliefs about maintaining multiple robots, their perception of autonomy, manual intervention to resolve degraded robots, and how they used the side menu. Finally, we collect ideas for improving the interface. The responses were given in Danish, and the quotes mentioned have been translated into English to be as close as possible to their original meaning.

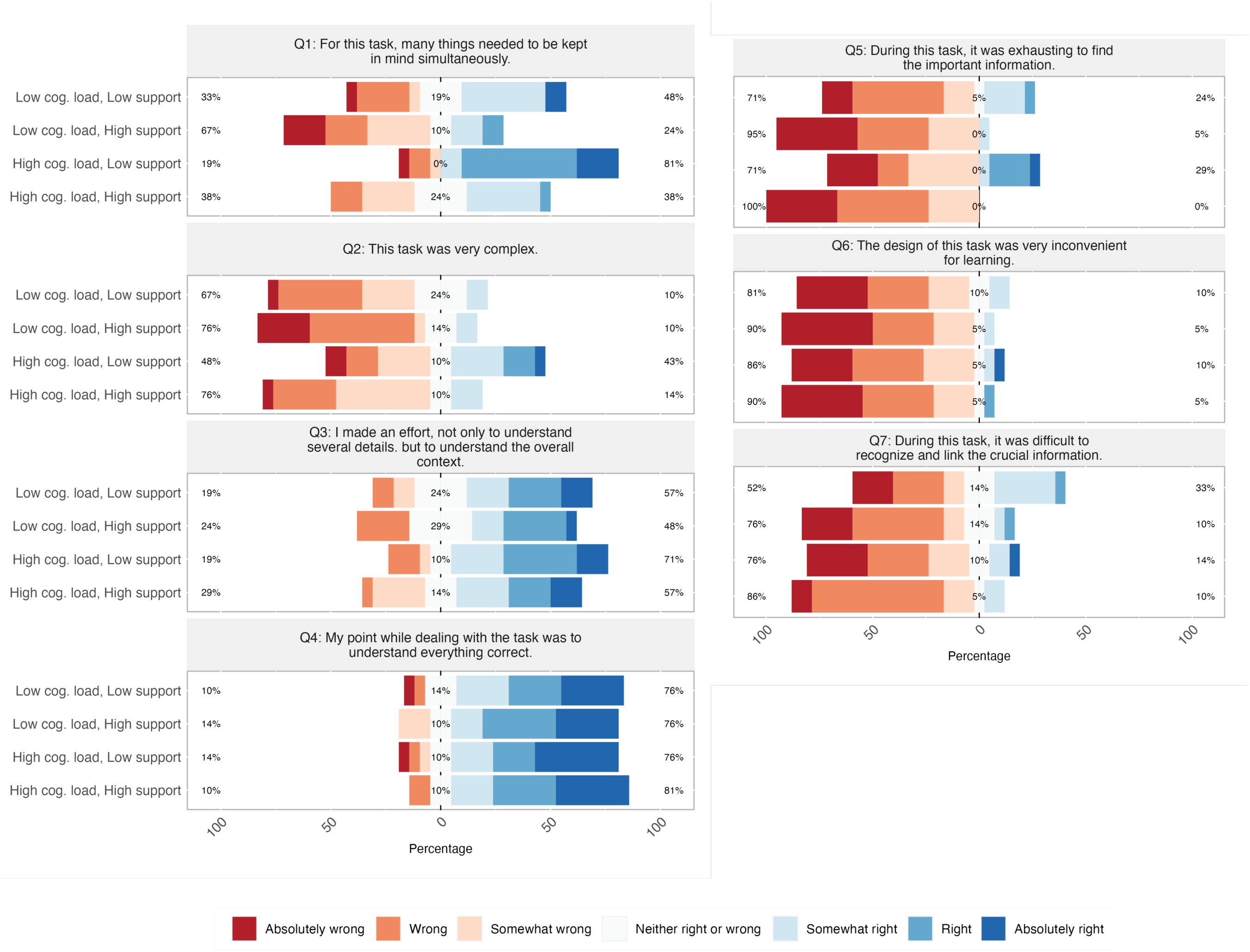


**Figure 6: Responses to the Klepsch et al.'s naïve questionnaire for cognitive load [52].**

*Maintaining multiple robots.* We asked how manageable it was to monitor a few robots compared to many robots; 66.66% of participants reported that they did not find any of the conditions unmanageable, and the remaining participants who found that the task could be difficult reported that it was not necessarily the number of robots that was the problem but how the task was presented. One participant said *"I did not think it was unmanageable. Sometimes it just took a little longer until I noticed, but I still had an overview of them"* (P4). Another said *"I don't think it really mattered that much if there were more or fewer robots. I think it was more significant whether you got warnings or not"* (P18). However, the participants also mentioned that while the tasks were not unmanageable, it was easier to maintain fewer robots than many.

*Perception of autonomy.* To complement the Likert-type questions about the perception of autonomy, we also asked the participants to discuss the robots' capabilities. Though we did not ask the participants to directly compare across conditions, surprisingly, one participant mentioned a difference in robot abilities. P4 stated *"I personally think I do better with more robots than with fewer robots. Perhaps it is because the frequency with which they drove wrong was greater with the fewer robots"*. As explained previously, we kept the number of alerts constant throughout conditions, which meant that the individual robots were less likely to go into a degraded state on average in the high cognitive load conditions.

The study platform was designed to randomly select a robot to enter a degraded state, which sometimes meant that a robot would go off-track and stop near a corner. In these cases, a robot could stop very close to the planned path of the next row. Some participants noted that this seemed like odd behavior because it was so close to rejoining the path. One participant noted *"[a robot] can go insanely far and still straight up, but it can also come out a little bit [. . . ], There was one [. . . ] where it had gone all the way over here. The dot was almost on the line, but it was stopped"*(P22). Another side effect

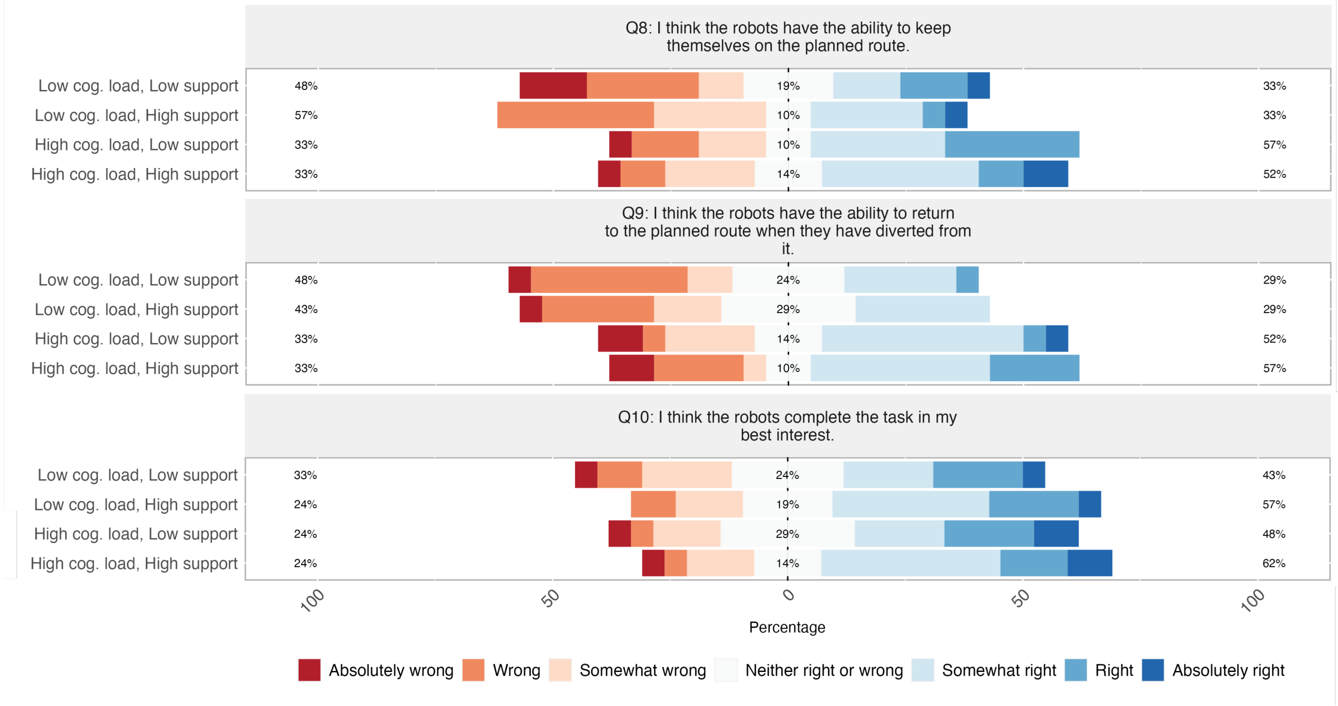


**Figure 7: Responses to the 7-point Likert scale on the perception of the robots' autonomous behaviors.**

of the randomly occurring degraded states was that some robots entered a degraded state more than once during a session. Some participants noticed this and remarked that seeing a robot fail more than once led them to distrust that robot and to be more cautious of its behavior compared to the others. One participant mentioned that *"[...] if I could, I would have stopped it"*(P13). Upon further discussion, the participant explained that stopping the robot would have resulted in fewer points, but they did not want to waste more attention and effort on a problematic robot.

*Resolving degraded state manually.* 80.95% of the participants expressed that they tried to resolve degraded robots manually rather than wait for the robots to possibly resolve by themselves. Three of those participants explained that they wanted to get as many points as possible, so they directed the robots back to the path as soon as they saw the degraded robot. The uncertainty of whether a robot would resolve itself and some robots going to a degraded state immediately after resolving it added to the participants' preferences of resolving the issues themselves. One participant suggested a visual indicator that acts as *"[...] a kind of prediction of how big is the chance that this will be resolved"* (P22) so that the user could evaluate the likelihood of that if it was a higher percentage of it being resolved automatically so they could focus their attention elsewhere.

Two participants responded that they were more lenient in allowing the robots to resolve themselves, specifically when exposed to many robots. One participant noted *"[...] there were times where it took longer before I saw that the [robots] that diverted from their path, but they sometimes went back themselves, which made me trust them more"* (P16). Another participant said *"[...] I directed them immediately, but I trusted that they would maybe do it themselves [...]"* (P15). Other participants also noted that they did not necessarily allow the robots to redirect themselves in the conditions with many robots, but because they were not fast enough to manually redirect them, the robots managed to do it automatically.

*Focus on the Whole vs the Parts .* More than half of the participants reported that when exposed to a high level of support (warning messages provided in the menu), they used the map mainly as a tool to navigate robots back on their planned path rather than as a means to monitor the robots' current state. Participants explained their strategies for how they carried out their tasks, claiming, *"I did not really look at the map other than when I had to direct the [robots] back"* (P16) and *"at first [...] I kept an eye on the robots themselves because I thought it was the easiest. And then it dawned on me very quickly that it was much easier to just watch for the warning signs"* (P21). During the low support condition (no warning messages in the menu), some participants reported shifting their focus to the map rather than the side menu. One such participant stated *"When there were few robots, then I would primarily use the map [...] because it was easier to manage [...]* (P11). However, one participant reported that they lost sight of the other robots because when they saw a warning, they would immediately attend to it, and when they would return to the side menu, they would find a new robot that had gone off-track.

Although most preferred using the side menu when warnings were displayed, six participants preferred to rely on the map. For one participant, the menu was also a source of frustration where they mentioned *"[...] it frustrated me that I had to go to the side menu because I felt that the side menu was redundant in a lot of ways"* (P3). Two participants expressed mixed opinions about the side menu and claimed they did not need the dedicated side menu to display warnings but then also stated that if the task involved a larger map, then some robots might not be visible so that the menu could be more helpful.

*Suggested interface improvements.* The system was designed to be simple, with few UI elements and basic means of interacting with the robots. We asked the participants if they had ideas for improving the interface to support them in the task. Ideas focused on increasing the options for selecting and interacting with the side menu and map elements. Five participants mentioned that the explicit warnings could also be designed to appear directly on the map instead of only appearing on the side menu to reduce having to look back and forth from the menu to the map. However, some reasoned that the side menu would still be useful if the task involved more robots. Four participants also noted that too many steps were needed to resolve a robot's degraded state. They suggested that an improvement could be moving the robots by clicking on their icon directly on the map and dragging them to the desired target location. One participant mentioned that instead of manually selecting a point on the planned path for redirecting a robot, further interface automation could be helpful and suggested that a button could be clicked that automatically selects a point on the path and reduces having to look between the map and side menu.

# 5 DISCUSSION

## 5.1 Maintaining Overview

An essential element in this study was incorporating the side menu, which displays the robots' state and, in two conditions, explicitly signaled to the participant that one or multiple robots had gone into a degraded or failed state. From the participants' responses, we found that the majority would focus on the side menu in the conditions where they would be notified if a robot had entered a degraded state. While participants were faster and found it easier to detect when the robots entered a degraded state when only looking at the side menu, they also expressed a loss of overview of the remaining robots. Participants did not seek to understand the robots' relation to their space on the map; instead, they focused only on the moment-to-moment events. This behavior can be problematic, especially when relating it to situation awareness design theory [45] since the user no longer considers the situation as a whole. Situational awareness has been a prevalent topic in monitoring systems as mentioned in section 2, where it is favored for a person to *perceive* the elements in the system, such they can *comprehend* their meaning and *project* what implications they have in the future [30]. With the current layout of the system, our result suggests that the user is incentivized only to pay attention when alerts appear rather than grasp the whole situation, which can negatively impact their situational awareness—something that should be avoided in a multi-robot system.

**Recommendation 1:** Multi-robot interfaces need to balance explicit alerts that support the monitoring task yet encourage users to keep an overview of the whole situation rather than only parts of it. Designers should consider providing information in a manner that incentivizes maintaining a holistic perspective and prevents hyperfixation on specific interface elements. Commercial pilots undergo training to develop good scanning behaviors across the instrument panel and horizon to improve situation awareness [79]. Designers should consider techniques for encouraging novices to adopt effective visual scanning practices.

## 5.2 Interaction Strategies

From both participants' performance data and the qualitative responses, we observed a consensus that a high level of support, i.e., being shown explicit warnings, was preferred and positively affected participant performances, specifically when given many robots to maintain. While participants reported that they did not personally feel like any of the tasks were unmanageable, our data shows that they performed worse when confronted with a high cognitive load and low support, both in terms of the number of errors and total points. This difference could also be due to how the participants approached the tasks. If a warning were explicitly shown on the side menu, they would primarily look there and only use the map to navigate the robots back when needed. From the reported feedback, we found that even when participants started out trying to maintain a broader overview of the robots by systematically scanning the map, when they saw an approach they found easier, they would quickly change their strategy to whatever they felt would help produce more points.

Some participants suggested potential solutions to the issue of continuously looking back and forth from the map to the side menu. Instead of separating the two elements from one another, warnings should also be displayed on the map itself, for example, by highlighting the faulty robot. If the user is already monitoring the map, they can detect robots in degraded and faulty states without relying on the side menu to gain this information. While research has already explored ways to notify users through the map [42, 49], they have primarily focused on how domain experts would interact with a multi-robot system. Our work involved non-expert users, and from the limited time and knowledge we gave them, they quickly found a monitoring strategy that would help them accomplish their tasks while gaining the most points. This simple interface provided a clear perspective of which interface elements had a direct influence on participant behavior. It displayed that when presented with alerts, even when starting without explicit alerts, they quickly saw how their task could be less tasking if changing strategies.

Many interface designs in multi-robot research utilize a similar design approach, with the map and another interface component to provide an area to provide more information about the robots' states [17, 26, 70, 71]. Though our study interface is straightforward and therefore limits our findings' applicability to real-world applications, our results indicate that multi-robot research should consider designs that do not visually separate information from one another as it increases the number of steps to complete the monitoring task.

**Recommendation 2:** As autonomous systems become more readily available for the general audience, designers must ensure that potential users monitor responsibly without fixating on achieving an immediate task. Our study displays that novices quickly find strategies that lessen the task load by removing the number of steps it takes to help a robot. Designers need to thoroughly consider how adequate monitoring behavior is not distinctly more demanding than inadequate monitoring behavior.

### 5.3 Distrust in Automation

Capiola et al. [16] presented in their work that there was a need to understand if self-correction affected the participant's perception of the autonomous capabilities of the robots. We addressed part of this by allowing the robots to resolve their degraded state after a few seconds. From the responses, we found that the majority of participants were primarily distrustful of the robot's autonomous behavior. Instead, they would rather manually resolve robots in a degraded state than allow the robots to do it by themselves. The distrust came from the uncertainty of when the automation would happen and whether the outcome would align with their expectations. This sentiment has been highlighted in previous work, where the experience with automation that frequently fails impacts a user's trust in its capabilities [58, 76]. However, interestingly, some participants reported that in conditions with higher cognitive load, it seemed as if the robots were more capable of resolving themselves. Some mentioned that they were more trusting of the robots in these scenarios, which is also reflected in the questionnaire responses (see Figure 7). This could also add to the performance data indicating that more auto-resolves happened at high cognitive load than when fewer robots were displayed. However, the increase could also be due to increased reaction time to an off-track event. Therefore, some auto-resolve was not done voluntarily but rather because the participants were not fast enough to manually correct it themselves.

From the results, we can see that there is an indication that when the task is sufficiently simple, participants would rather trust their own ability when possible, as also indicated by previous work [54]. Still, as our results show, when the complexity or cognitive requirements of a task increase, participants are more lenient in allowing autonomous behavior to take over or assist in the task. Our results also suggest that even if the task is seemingly simple, the amount of time it takes to resolve a degraded state also significantly influences how autonomously a robot is allowed to act.

A big part of why the participants did not allow the robots to redirect themselves autonomously was the unreliability of when they would be successful. One participant mentioned that displaying the probability of the robot self-resolving a degraded state could make the capabilities of the robot more transparent and provide enough information for the user to determine whether they want to take over themselves. We have seen many studies of **explainable AI (XAI)**, where the aim is to make it more clear to the user why and how an AI has provided the answer it has. Particularly, we have seen many examples of how to display probability and certainty for the detection of objects [7, 29] and the functionality of the AI [44, 48, 83]. Given the rise of XAI research, incorporating the research space into monitoring autonomous robots could provide novel insights that allow users to make more informed choices.

**Recommendation 3:** As automation is necessary for a system where multiple agents work simultaneously, designers need to consider how to visualize and explain the robots' capabilities and intentions to the user. This transparency will help them delegate tasks such that in a stressful and overwhelming situation, the user is aware of which robot can complete a task themselves, and the user can focus on more complex and problematic areas.

### 5.4 Limitations and Future Work

There are various limitations of our work related to sample size and diversity, duration of use, and generalizability that should be noted. Our study involved data from relatively few participants ($N = 22$), primarily students from a European university. While gaining insights from a larger number of participants and views from other contexts would be useful, the number of participants is in line with community standards for experiments [14]. We acknowledge that a wider participant pool from other cultures might yield more diverse attitudes and beliefs about automation, which we will consider in future studies. The study focused on short monitoring tasks that were a few minutes in duration. While the insights from this study provide initial impressions of how users respond to explicit alerts, studies of longer duration are important next steps to understanding the strategies emerging as users transition from novice users to experts. The generalizability of our results may be limited, considering that participants utilized a simulation of robots with a reduced set of interactive features instead of controlling actual fleets of robots. In future work, more realistic studies involving higher complexity and real-world consequences will be helpful and necessary as companies increasingly bring autonomous robots into various new work contexts (e.g., agriculture [46]). Exploring how XAI and other intelligent systems, such as conversational agents [70], can be used in multi-robot systems is also encouraged, given the uncertainness of automation presented in the results. Additional directions for future work include investigating the design considerations in this paper and how they improve the effectiveness of presenting alerts and robot status for monitoring groups of robots, as well as enabling the operator to understand the evolving situation and take the appropriate action when required.

## 6 CONCLUSION

We examined how the explicit display of alerts can affect a person's performance and perception in a multi-robot monitoring task. We asked participants to monitor the status of robots that were actively engaged in a coverage task in which they should follow a pre-determined path. The robots could go off-track at any time, and potentially redirect themselves back to the path. Participants were asked to monitor and intervene by redirecting robots if they wandered off track or stopped completely. The results of the within-subject design showed that having a higher level of support would motivate the users to only look at the robots' status in a robot list rather than observe their behavior on the displayed map. They would also perform better when monitoring fewer robots that displayed explicit warnings prior to entering a failed state and faster reaction time to robots going off-track. However, incentivizing users to focus on alert messages and ignore the map could lead to a loss of overview and degraded situation awareness. Participants also reported that as the presented task was fairly simple, they did not trust the robot to redirect itself and would consequently try and manually redirect them when needed. However, as the task complexity became more difficult, they became more reliant on automation. Instead of using the common design of a side menu for alert information, future research could examine how to design and integrate status information directly into the map or other ways to avoid fixation habits when monitoring multi-robot systems and other complex environments. This work becomes increasingly important considering the rise of human-AI collaboration in various contexts.

## ACKNOWLEDGMENTS

This work is supported by the Innovation Fund Denmark for the project DIREC (9142-00001B).

## REFERENCES

[1] Monica Adya and Gloria Phillips-Wren. 2019. Stressed decision makers and use of decision aids: a literature review and conceptual model. *Information Technology & People* 33, 2 (Jan. 2019), 710–754. https://doi.org/10.1108/ITP-04-2019-0194 Publisher: Emerald Publishing Limited.

[2] Ankit Agrawal, Sophia J. Abraham, Benjamin Burger, Chichi Christine, Luke Fraser, John M. Hoeksema, Sarah Hwang, Elizabeth Travnik, Shreya Kumar, Walter Scheirer, Jane Cleland-Huang, Michael Vierhauser, Ryan Bauer, and Steve Cox. 2020. The Next Generation of Human-Drone Partnerships: Co-Designing an Emergency Response System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376825

[3] Sarah Al-Hussaini, Jason M. Gregory, Yuxiang Guan, and Satyandra K. Gupta. 2020. Generating Alerts to Assist With Task Assignments in Human-Supervised Multi-Robot Teams Operating in Challenging Environments. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 11245–11252. https://doi.org/10.1109/IROS45743.2020.9341588 ISSN: 2153-0866.

[4] Izz aldin Hamdan, August Capiola, Gene M. Alarcon, Joseph B. Lyons, Keitaro Nishimura, Katia Sycara, and Michael Lewis. 2021. Exploring the Effects of Swarm Degradations on Trustworthiness Perceptions, Reliance Intentions, and Reliance Behaviors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 65, 1 (Sept. 2021), 1141–1145. https://doi.org/10.1177/1071181321651057 Publisher: SAGE Publications Inc.

[5] Salvatore Andolina and Jodi Forlizzi. 2014. The design of interfaces for multi-robot path planning and control. In *2014 IEEE International Workshop on Advanced Robotics and its Social Impacts*. IEEE, Evanston, IL, USA, 7–13. https://doi.org/10.1109/ARSO.2014.7020972

[6] J. Elin Bahner, Monika F. Elepfandt, and Dietrich Manzey. 2008. Misuse of Diagnostic Aids in Process Control: The Effects of Automation Misses on Complacency and Automation Bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52, 19 (Sept. 2008), 1330–1334. https://doi.org/10.1177/154193120805201906 Publisher: SAGE Publications Inc.

[7] Adrian Banuls, Anthony Mandow, Ricardo Vazquez-Martin, Jesus Morales, and Alfonso Garcia-Cerezo. 2020. Object Detection from Thermal Infrared and Visible Light Cameras in Search and Rescue Scenes. In *2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, Abu Dhabi, United Arab Emirates, 380–386. https://doi.org/10.1109/SSRR50563.2020.9292593

[8] Shishir Bashyal and Ganesh Kumar Venayagamoorthy. 2008. Human swarm interaction for radiation source search and localization. In *2008 IEEE Swarm Intelligence Symposium*. IEEE, St. Louis, MO, USA, 1–8. https://doi.org/10.1109/SIS.2008.4668287

[9] Stephanie Bonadies and S. Andrew Gadsden. 2019. An overview of autonomous crop row navigation strategies for unmanned ground vehicles. *Engineering in Agriculture, Environment and Food* 12, 1 (Jan. 2019), 24–31. https://doi.org/10.1016/j.eaef.2018.09.001

[10] Leonard A. Breslow, Daniel Gartenberg, J. Malcolm McCurry, and J. Gregory Trafton. 2014. Dynamic Operator Overload: A Model for Predicting Workload During Supervisory Control. *IEEE Transactions on Human-Machine Systems* 44, 1 (Feb. 2014), 30–40. https://doi.org/10.1109/TSMC.2013.2293317 Conference Name: IEEE Transactions on Human-Machine Systems.

[11] S. Breznitz. 2013. *Cry Wolf: The Psychology of False Alarms*. Psychology Press. Google-Books-ID: lxwLVy16wqoC.

[12] Jeffrey B. Brookings, Glenn F. Wilson, and Carolyne R. Swain. 1996. Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology* 42, 3 (Feb. 1996), 361–377. https://doi.org/10.1016/0301-0511(95)05167-8

[13] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–21. https://doi.org/10.1145/3449287

[14] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[15] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. 2023. How Time Pressure in Different Phases of Decision-Making Influences Human-AI Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–26. https://doi.org/10.1145/3610068

[16] August Capiola, Izz aldin Hamdan, Joseph B. Lyons, Michael Lewis, Gene M. Alarcon, and Katia Sycara. 2024. The Effect of Asset Degradation on Trust in Swarms: A Reexamination of System-Wide Trust in Human-Swarm Interaction. *Human Factors* 66, 5 (May 2024), 1475–1489. https://doi.org/10.1177/00187208221145261 Publisher: SAGE Publications Inc.

[17] Roger A. Chadwick. 2006. Operating Multiple Semi-Autonomous Robots: Monitoring, Responding, Detecting. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 3 (Oct. 2006), 329–333. https://doi.org/10.1177/154193120605000325 Publisher: SAGE Publications Inc.

[18] Jessie Y. C. Chen and Michael J. Barnes. 2014. Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues. *IEEE Transactions on Human-Machine Systems* 44, 1 (Feb. 2014), 13–29. https://doi.org/10.1109/THMS.2013.2293535

[19] Jessie Y. C. Chen, Michael J. Barnes, and Michelle Harper-Sciarini. 2011. Supervisory Control of Multiple Robots: Human-Performance Issues and User-Interface Design. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 4 (July 2011), 435–454. https://doi.org/10.1109/TSMCC.2010.2056682

[20] Patrick Christ, Florian Lachner, Axel Hösl, Bjoern Menze, Klaus Dieopold, and Andreas Butz. 2016. *Human-Drone-Interaction: A Case Study to Investigate the Relation Between Autonomy and User Experience*. Vol. 9914. https://doi.org/10.1007/978-3-319-48881-3_17 Pages: 253.

[21] Anders Lyhne Christensen, Kasper Andreas Rømer Grøntved, Maria-Theresa Oanh Hoang, Niels van Berkel, Mikael Skov, Alea Scovill, Gareth Edwards, Kenneth Richard Geipel, Lars Dalgaard, Ulrik Pagh Schultz Lundquist, Ioanna Constantiou, Christiane Lehrer, and Timothy Merritt. 2022. The HERD project: Human-multi-robot interaction in search & rescue and in farming. In *Adjunct Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. 1–4.

[22] J.W. Crandall, M.A. Goodrich, D.R. Olsen, and C.W. Nielsen. 2005. Validating Human–Robot Interaction Schemes in Multitasking Environments. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 35, 4 (July 2005), 438–449. https://doi.org/10.1109/TSMCA.2005.850587

[23] Mary Cummings. 2004. Automation bias in intelligent time critical decision support systems. In *AIAA 1st intelligent systems technical conference*. 6313.

[24] M.L. Cummings, C. Mastracchio, K.M. Thornburg, and A. Mkrtchyan. 2013. Boredom and Distraction in Multiple Unmanned Vehicle Supervisory Control. *Interacting with Computers* 25, 1 (Jan. 2013), 34–47. https://doi.org/10.1093/iwc/iws011

[25] M.L. Cummings and P.J. Mitchell. 2008. Predicting Controller Capacity in Supervisory Control of Multiple UAVs. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 38, 2 (March 2008), 451–460. https://doi.org/10.1109/TSMCA.2007.914757

[26] M. L. Cummings and P. J. Mitchell. 2007. Operator scheduling strategies in supervisory control of multiple UAVs. *Aerospace Science and Technology* 11, 4 (May 2007), 339–348. https://doi.org/10.1016/j.ast.2006.10.007

[27] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. Explainable AI for Robot Failures: Generating Explanations that Improve User Assistance in Fault Recovery. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Boulder CO USA, 351–360. https://doi.org/10.1145/3434073.3444657

[28] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 251–258. https://doi.org/10.1109/HRI.2013.6483596 ISSN: 2167-2148.

[29] Zsolt Domozi, Daniel Stojcsics, Abdallah Benhamida, Miklos Kozlovszky, and Andras Molnar. 2020. Real time object detection for aerial search and rescue missions for missing persons. In *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)*. IEEE, Budapest, Hungary, 000519–000524. https://doi.org/10.1109/SoSE50414.2020.9130475

[30] Mica R. Endsley. 1995. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors* 37, 1 (March 1995), 32–64. https://doi.org/10.1518/001872095779049543 Publisher: SAGE Publications Inc.

[31] Mica R. Endsley and Mark D. Rodgers. 1994. Situation Awareness Information Requirements Analysis for En Route Air Traffic Control. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 38, 1 (Oct. 1994), 71–75. https://doi.org/10.1177/154193129403800113 Publisher: SAGE Publications Inc.

[32] Maik Friedrich, Maresa Biermann, Patrick Gontar, Marcus Biella, and Klaus Bengler. 2018. The influence of task load on situation awareness and control strategy in the ATC tower environment. *Cognition, Technology & Work* 20, 2 (May 2018), 205–217. https://doi.org/10.1007/s10111-018-0464-4

[33] Zohar Gilad, Ofra Amir, and Liat Levontin. 2021. The Effects of Warmth and Competence Perceptions on Users' Choice of an AI System. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. https://doi.org/10.1145/3411764.3446863

[34] Nitin Gupta, Ann M. Bisantz, and Tarunraj Singh. 2001. Investigation of Factors Affecting Driver Performance Using Adverse Condition Warning Systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 45, 23 (Oct. 2001), 1699–1703. https://doi.org/10.1177/154193120104502329 Publisher: SAGE Publications Inc.

[35] Mustapha Mouloua Hilburn, Brian Robert Molloy, and Raja Parasuraman. 1996. Monitoring of Automated Systems. In *Automation and Human Performance*. CRC Press. Num Pages: 25.

[36] Maria-Theresa Oanh Hoang, Kasper Andreas Rømer Grøntved, Niels van Berkel, Mikael B Skov, Anders Lyhne Christensen, and Timothy Merritt. 2023. Drone Swarms to Support Search and Rescue Operations: Opportunities and Challenges. *Cultural Robotics: Social Robots and Their Emergent Cultural Ecologies* (2023), 163–176.

[37] Maria-Theresa Oanh Hoang, Niels van Berkel, Mikael B Skov, and Timothy Merritt. 2022. Challenges Arising in a Multi-Drone System for Search and Rescue. In *Adjunct Proceedings of the 12th Nordic Conference on Human-Computer Interaction*. 1–5.

[38] Maria-Theresa Oanh Hoang, Niels van Berkel, Mikael B. Skov, and Timothy R. Merritt. 2023. Challenges and Requirements in Multi-Drone Interfaces. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3544549.3585673

[39] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (May 2015), 407–434. https://doi.org/10.1177/0018720814547570 Publisher: SAGE Publications Inc.

[40] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development. *Frontiers in Psychology* 9 (2018). https://doi.org/10.3389/fpsyg.2018.00861

[41] Aya Hussein, Sondoss Elsawah, and Hussein A. Abbass. 2020. Trust Mediating Reliability–Reliance Relationship in Supervisory Control of Human–Swarm Interactions. *Human Factors* 62, 8 (Dec. 2020), 1237–1248. https://doi.org/10.1177/0018720819879273 Publisher: SAGE Publications Inc.

[42] Jean-Paul Imbert, Helen M. Hodgetts, Robert Parise, François Vachon, Frédéric Dehais, and Sébastien Tremblay. 2014. Attentional costs and failures in air traffic control notifications. *Ergonomics* 57, 12 (Dec. 2014), 1817–1832. https://doi.org/10.1080/00140139.2014.952680

[43] Rune Hylsberg Jacobsen, Lea Matlekovic, Liping Shi, Nicolaj Malle, Naeem Ayoub, Kaspar Hageman, Simon Hansen, Frederik Falk Nyboe, and Emad Ebeid. 2023. Design of an Autonomous Cooperative Drone Swarm for Inspections of Safety Critical Infrastructure. *Applied Sciences* 13, 3 (Jan. 2023), 1256. https://doi.org/10.3390/app13031256 Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[44] Weiwei Jiang, Zhanna Sarsenbayeva, Niels Van Berkel, Chaofan Wang, Difeng Yu, Jing Wei, Jorge Goncalves, and Vassilis Kostakos. 2021. User Trust in Assisted Decision-Making Using Miniaturized Near-Infrared Spectroscopy. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445710

[45] Debra G. Jones, Betty Bolte, , and Mica R. Endsley. 2003. *Designing for Situation Awareness: An Approach to User-Centered Design*. CRC Press, London. https://doi.org/10.1201/9780203485088

[46] Chanyoung Ju, Jeongeun Kim, Jaehwi Seol, and Hyoung Il Son. 2022. A review on multirobot systems in agriculture. *Computers and Electronics in Agriculture* 202 (Nov. 2022), 107336. https://doi.org/10.1016/j.compag.2022.107336

[47] Daisuke Karikawa, Hisae Aoyama, Makoto Takahashi, Kazuo Furuta, Akira Ishibashi, and Masaharu Kitamura. 2014. Analysis of the performance characteristics of controllers' strategies in en route air traffic control tasks. *Cognition, Technology & Work* 16, 3 (Aug. 2014), 389–403. https://doi.org/10.1007/s10111-013-0268-5

[48] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. "Why Do I Care What's Similar?" Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference*. ACM, Virtual Event Australia, 454–470. https://doi.org/10.1145/3532106.3533556

[49] Peter Kearney, Wen-Chin Li, Chung-San Yu, and Graham Braithwaite. 2019. The impact of alerting designs on air traffic controller's eye movement patterns and situation awareness. *Ergonomics* 62, 2 (Feb. 2019), 305–318. https://doi.org/10.1080/00140139.2018.1493151 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00140139.2018.1493151.

[50] Eliahu Khalastchi and Meir Kalech. 2019. Fault Detection and Diagnosis in Multi-Robot Systems: A Survey. *Sensors* 19, 18 (Jan. 2019), 4019. https://doi.org/10.3390/s19184019 Number: 18 Publisher: Multidisciplinary Digital Publishing Institute.

[51] Eliahu Khalastchi and Meir Kalech. 2019. On Fault Detection and Diagnosis in Robotic Systems. *Comput. Surveys* 51, 1 (Jan. 2019), 1–24. https://doi.org/10.1145/3146389

[52] Melina Klepsch, Florian Schmitz, and Tina Seufert. 2017. Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. *Frontiers in Psychology* 8 (2017). https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01997

[53] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 203–210. https://doi.org/10.1109/HRI.2010.5453195 ISSN: 2167-2148.

[54] Poornima Madhavan, Douglas A. Wiegmann, and Frank C. Lacson. 2006. Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. *Human Factors* 48, 2 (June 2006), 241–256. https://doi.org/10.1518/001872006777724408 Publisher: SAGE Publications Inc.

[55] Angela Mastrianni, Aleksandra Sarcevic, Lauren Chung, Issa Zakeri, Emily Alberto, Zachary Milestone, Ivan Marsic, and Randall S Burd. 2021. Designing Interactive Alerts to Improve Recognition of Critical Events in Medical Emergencies. In *Designing Interactive Systems Conference 2021*. ACM, Virtual Event USA, 864–878. https://doi.org/10.1145/3461778.3462051

[56] Jake N. McRae, Christopher J. Gay, Brandon M. Nielsen, and Andrew P. Hunt. 2019. Using an Unmanned Aircraft System (Drone) to Conduct a Complex High Altitude Search and Rescue Operation: A Case Study. *Wilderness & Environmental Medicine* 30, 3 (Sept. 2019), 287–290. https://doi.org/10.1016/j.wem.2019.03.004

[57] Stephanie M. Merritt, Alicia Ako-Brew, William J. Bryant, Amy Staley, Michael McKenna, Austin Leone, and Lei Shirase. 2019. Automation-Induced Complacency Potential: Development and Validation of a New Scale. *Frontiers in Psychology* 10 (2019). https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00225

[58] Stephanie M. Merritt and Daniel R. Ilgen. 2008. Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors* 50, 2 (April 2008), 194–210. https://doi.org/10.1518/001872008X288574 Publisher: SAGE Publications Inc.

[59] Mark Miller and Sam Holley. 2021. Air Traffic Controller Resource Management: An Approach for Reducing Cognitive Loading and Increasing Situational Awareness. In *Advances in Human Aspects of Transportation*, Neville Stanton (Ed.). Springer International Publishing, Cham, 535–542. https://doi.org/10.1007/978-3-030-80012-3_61

[60] Cecilia G. Morales, Elizabeth J. Carter, Xiang Zhi Tan, and Aaron Steinfeld. 2019. Interaction Needs and Opportunities for Failing Robots. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. Association for Computing Machinery, New York, NY, USA, 659–670. https://doi.org/10.1145/3322276.3322345

[61] Dan R. Olsen and Stephen Bart Wood. 2004. Fan-out: measuring human control of multiple robots. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Vienna Austria, 231–238. https://doi.org/10.1145/985692.985722

[62] Lisa Ordóñez and Lehman Benson. 1997. Decisions under Time Pressure: How Time Constraint Affects Risky Decision Making. *Organizational Behavior and Human Decision Processes* 71, 2 (Aug. 1997), 121–140. https://doi.org/10.1006/obhd.1997.2717

[63] Murillo Pagnotta, David M. Jacobs, Patricia L. de Frutos, Ruben Rodríguez, Jorge Ibáñez-Gijón, and David Travieso. 2022. Task difficulty and physiological measures of mental workload in air traffic control: a scoping review. *Ergonomics* 65, 8 (Aug. 2022), 1095–1118. https://doi.org/10.1080/00140139.2021.2016998 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00140139.2021.2016998.

[64] Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors* 52, 3 (June 2010), 381–410. https://doi.org/10.1177/0018720810376055 Publisher: SAGE Publications Inc.

[65] Raja Parasuraman, Robert Molloy, and Indramani L. Singh. 1993. Performance Consequences of Automation-Induced 'Complacency'. *The International Journal of Aviation Psychology* 3, 1 (Jan. 1993), 1–23. https://doi.org/10.1207/s15327108ijap0301_1 Publisher: Taylor & Francis _eprint: https://doi.org/10.1207/s15327108ijap0301_1.

[66] Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. 2008. Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making* 2, 2 (June 2008), 140–160. https://doi.org/10.1518/155534308X284417 Publisher: SAGE Publications.

[67] Gloria Phillips-Wren and Monica Adya. 2020. Decision making under stress: the role of information overload, time pressure, complexity, and uncertainty. *Journal of Decision Systems* 29, sup1 (Aug. 2020), 213–225. https://doi.org/10.1080/12460125.2020.1768680 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/12460125.2020.1768680.

[68] Stephen Rice, David Keller, Gayle Hunt, and David Trafimow. 2009. Automation Dependency Under Time Pressure. *2009 International Symposium on Aviation Psychology* (Jan. 2009), 611–616. https://corescholar.libraries.wright.edu/isap_2009/14

[69] Jennifer M. Riley and Laura D. Strater. 2006. Effects of Robot Control Mode on Situation Awareness and Performance in a Navigation Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 3 (Oct. 2006), 540–544. https://doi.org/10.1177/154193120605000369 Publisher: SAGE Publications Inc.

[70] David A. Robb, José Lopes, Stefano Padilla, Atanas Laskov, Francisco J. Chiyah Garcia, Xingkun Liu, Jonatan Scharff Willners, Nicolas Valeyrie, Katrin Lohan, David Lane, Pedro Patron, Yvan Petillot, Mike J. Chantler, and Helen Hastie. 2019. Exploring Interaction with Remote Autonomous Systems using Conversational Agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. Association for Computing Machinery, New York, NY, USA, 1543–1556. https://doi.org/10.1145/3322276.3322318

[71] Adam Rule and Jodi Forlizzi. 2012. Designing interfaces for multi-user, multi-robot systems. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, Boston Massachusetts USA, 97–104. https://doi.org/10.1145/2157689.2157705

[72] Fabrice Saffre, Hanno Hildmann, and Hannu Karvonen. 2021. The Design Challenges of Drone Swarm Control. In *Engineering Psychology and Cognitive Ergonomics*, Don Harris and Wen-Chin Li (Eds.). Vol. 12767. Springer International Publishing, Cham, 408–426. https://doi.org/10.1007/978-3-030-77932-0_32 Series Title: Lecture Notes in Computer Science.

[73] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Portland Oregon USA, 141–148. https://doi.org/10.1145/2696454.2696497

[74] Zhanna Sarsenbayeva, Niels van Berkel, Danula Hettiachchi, Weiwei Jiang, Tilman Dingler, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2019. Measuring the Effects of Stress on Mobile Interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 3, 1 (2019), 24:1–24:18. https://doi.org/10.1145/3314411

[75] Nadine B. Sarter and Beth Schroeder. 2001. Supporting Decision Making and Action Selection under Time Pressure and Uncertainty: The Case of In-Flight Icing. *Human Factors* 43, 4 (Dec. 2001), 573–583. https://doi.org/10.1518/001872001775870403 Publisher: SAGE Publications Inc.

[76] Juergen Sauer, Alain Chavaillaz, and David Wastell. 2016. Experience of automation failures in training: effects on trust, automation bias, complacency and performance. *Ergonomics* 59, 6 (June 2016), 767–780. https://doi.org/10.1080/00140139.2015.1094577 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00140139.2015.1094577.

[77] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3581075

[78] Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (Nov. 1999), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

[79] Hamed Taheri Gorji, Nicholas Wilson, Jessica VanBree, Bradley Hoffmann, Thomas Petros, and Kouhyar Tavakolian. 2023. Using machine learning methods and EEG to discriminate aircraft pilot cognitive workload during flight. *Scientific Reports* 13, 1 (Feb. 2023), 2507. https://doi.org/10.1038/s41598-023-29647-0 Publisher: Nature Publishing Group.

[80] B. Trouvain and H.L. Wolf. 2002. Evaluation of multi-robot control and monitoring performance. In *11th IEEE International Workshop on Robot and Human Interactive Communication Proceedings*. 111–116. https://doi.org/10.1109/ROMAN.2002.1045607

[81] Casey Tunstall, Stephen Rice, Rian Mehta, Victoria Dunbar, and Korhan Oyman. 2014. Time Pressure Has Limited Benefits for Human-Automation Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58, 1 (Sept. 2014), 1043–1046. https://doi.org/10.1177/1541931214581218 Publisher: SAGE Publications Inc.

[82] Niels van Berkel, Omer F. Ahmad, Danail Stoyanov, Laurence Lovat, and Ann Blandford. 2021. Designing Visual Markers for Continuous Artificial Intelligence Support: A Colonoscopy Case Study. *ACM Transactions on Computing for Healthcare* 2, 1 (Dec. 2021), 7:1–7:24. https://doi.org/10.1145/3422156

[83] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–18. https://doi.org/10.1145/3411764.3445432

[84] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 318–328. https://doi.org/10.1145/3397481.3450650

[85] Christopher D. Wickens, Benjamin A. Clegg, Alex Z. Vieane, and Angelia L. Sebok. 2015. Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors* 57, 5 (Aug. 2015), 728–739. https://doi.org/10.1177/0018720815581940 Publisher: SAGE Publications Inc.

[86] James Wilson, Greg Chance, Peter Winter, Suet Lee, Emma Milner, Dhaminda Abeywickrama, Shane Windsor, John Downer, Kerstin Eder, Jonathan Ives, and Sabine Hauert. 2023. Trustworthy Swarms. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*. ACM, Edinburgh United Kingdom, 1–11. https://doi.org/10.1145/3597512.3599705

[87] 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*. ACM, Vancouver BC Canada, 143–146. https://doi.org/10.1145/1978942.1978963

[88] Choon Yue Wong and Gerald Seet. 2017. Workload, awareness and automation in multiple-robot supervision. *International Journal of Advanced Robotic Systems* 14, 3 (May 2017), 1729881417710463. https://doi.org/10.1177/1729881417710463 Publisher: SAGE Publications.

[89] Peter Wright. 1974. The harassed decision maker: Time pressures, distractions, and the use of evidence. *Journal of Applied Psychology* 59, 5 (1974), 555–561. https://doi.org/10.1037/h0037186 Place: US Publisher: American Psychological Association.