# Situational Cues for Continuous Trust Calibration in Automated Systems

Alexandra Forsey-Smerek[1], Katya Arquilla[2], and Julie Shah[3]
*Massachusetts Institute of Technology, Cambridge, MA 02139*

**Appropriate user trust calibration in automated systems is critical for optimizing system usage, improving task performance, and reducing user workload. While undertrust in a system may lead to system disuse and suboptimal performance, overtrust in a system can result in reduced user situation awareness and susceptibility to consequences of system failure. In dynamic domains, the reliability of an automated system may fluctuate based on environmental conditions and the type of task being performed. Fluctuation in system reliability demands that user trust in the system adapts to optimize system usage and task performance. Recent attention has been focused on the presentation of adaptive trust calibration cues based on quantified human operator trust to promote appropriate trust calibration. In dynamic domains such as future human and robotic space exploration missions, real-time quantification of operator trust in an automated system may not be possible. In this work, we expand the application space of trust calibration cues through the introduction of situational trust cues (STCs), presented independent of user real-time trust in the system. Situational trust calibration cues are presented if environmental conditions or task type is changed in a manner that significantly affects performance of the automated system. We present the design of an experiment investigating the application of STCs to a Lunar rover path planning automated decision support tool in a dynamic mission operations environment. The experiment is designed to assess the impact of STCs on user selection of the most appropriate level of automation to afford an automated decision support tool, where the most appropriate level of automation fluctuates between three automation levels based on the type of planning task being executed and the state of the simulated environment. We ultimately plan to investigate the utility of STCs in mitigating user overtrust and undertrust in automated systems.**

## Nomenclature

| | | |
|---|---|---|
| *DEM* | = | digital elevation model |
| *J* | = | joule |
| *kg* | = | kilogram |
| *km* | = | kilometer |
| *km/h* | = | kilometers per hour |
| *LRV* | = | Lunar Roving Vehicle |
| *PDDL* | = | Planning Domain Definition Language |
| *O-STCs* | = | over-cued situational trust cues |
| *SEXTANT* | = | Surface Exploration Traverse Analysis and Navigation Tool |
| *STCs* | = | situational trust cues |
| *TCCs* | = | trust calibration cues |
| *VIPER* | = | Volatiles Investigating Polar Exploration Rover |
| *W* | = | watt |

## I. Introduction

HUMAN operation of aerospace vehicles requires increasing interaction with varying levels of automation. As automation becomes ubiquitous in all areas of aerospace—including aircraft, spacecraft, and unmanned vehicle operations—appropriate user affordance of autonomy to decision support tools for a given scenario are key[1]. In some circumstances, the capabilities of automated systems are favorable for task execution, while in others, the unique skills and adaptability of a human operator are essential for task completion. Appropriate delegation of tasks to automated systems can also reduce inter-operator variability in task performance. The tradeoff between human and automation

[1] Graduate Student, Aeronautics and Astronautics, 77 Massachusetts Ave, Cambridge, MA 02139, aforsey@mit.edu
[2] Postdoctoral Fellow, Aeronautics and Astronautics, 77 Massachusetts Ave, Cambridge, MA 02139
[3] Associate Professor, Aeronautics and Astronautics, 77 Massachusetts Ave, Cambridge, MA 02139

is governed by the user's trust in the system and ability to navigate between different levels of automation in varying task types and environmental conditions. Within this tradeoff, user overtrust and undertrust in automated systems presents a significant risk to overall task performance.

One method of mitigating this risk is to enable the human to select the appropriate level of automation given the current conditions and system capabilities by delivering cues during the decision-making process, which is an open area of research. One way to do this is to provide trust calibration cues (TCCs), as developed by Okamura and Yamada, that aim to identify user overtrust or undertrust in a system and subsequently steer user trust with discrete cues to increase or decrease trust during task execution[2]. However, directional steering is limited by the need to characterize human performance in relation to automated system performance for each combination of task and environment. Additionally, directional steering requires real-time identification of human operator trust state in the automated system. Real-time, non-interruptive measurements of operator trust state are often based on behavioral indicators such as system reliance, which has been found to be significantly positively correlated with reported user trust in a system[3]. However, in many operational scenarios, real-time calculation of human reliance on an automated system may be infeasible or limited in its ability to capture the construct of trust.

Our approach builds from the concept of TCCs, but the proposed Situational Trust Cues (STCs) aim to bypass the need for real-time trust state identification by triggering the human to examine their own level of trust and recalibrate how they are using the system based on both task type and environmental conditions. In this work, we present STCs as a method to promote appropriate user trust calibration in automated systems, and discuss an experiment design devised to investigate the effectiveness of STCs applied to a rover path planning task. We consider the limitations of STCs in the experimental design, with specific attention allotted to the potential negative effects of presenting cues too frequently.

## II. Related Work

A common approach to improving user trust in an automated system is to provide confidence information to the user. This confidence information usually provides some assessment of the system's capabilities in a given environment. Verame et al., 2016 conducted a study to investigate the effectiveness of such confidence information on a user's trust in an autonomous software system designed to translate hand-written text to typed documents. Confidence information was provided as the system's perceived confidence on each task and was presented as a binary "yes" or "no" value, and they found that the provision of high-confidence information encouraged blind use of the autonomous system[4]. McGuirl and Sarter, 2006, shows the benefits of continually updating confidence information throughout the task to improve trust calibration and appropriate use of an autonomous system for pilots handling in-flight icing[5]. Despite these studied benefits of system confidence information, other approaches to trust calibration have been investigated and show promise for calibrating trust in situations where interruptive cues during task completion are potentially more impactful than a constant presentation of system confidence information.

Visser et al., 2014 broadly defines a trust cue as "any information element that can be used to make a trust assessment about an agent", which can be presented as an approach to calibrating appropriate trust of a system[6]. A method recently proposed by Okamura and Yamada, 2020, uses a behavioral indicator to detect user overtrust and undertrust in an automated system, and adaptively presents "trust calibration cues (TCCs)" to steer users to properly adjust their trust in the system[2]. One study where TCCs were applied simulated a pothole inspection task, where users decided whether to either manually inspect a pothole, or delegate the task to an automated system. The automated system was expected to outperform the human on the task except in inclement weather scenarios, which resulted in degraded performance of the automated system to the point where the human was expected to outperform the system. TCCs were presented if user trust, measured by the selection of whether or not to delegate the task to the automated system, matched what was appropriate for the environment at a certain time. TCCs were demonstrated to have a significant effect on user reliance on the automated system, helping users to successfully alter their behavior between weather changes. Presentation of constant system confidence information for the same task did not show a significant effect on user reliance on the automated system[7]. However, this method is limited in application by its dependence on the ability to detect human overtrust and undertrust in real-time and reliance on the preliminary quantification of appropriate human reliance on the automated system throughout all scenarios that will be encountered.

Cai and Lin, 2010, previously introduced "cognitive cues", which present system confidence information at decision points using visual, auditory, or tactile cues. The presentation of cognitive cues had a significant effect on operator compliance with system recommendations when completing a driving task using an automated driving assistant[8]. Unlike TCCs, these cues are presented independent of real-time human trust measurements and focus on

communicating system confidence information. An example of a cue would be a sound or steering wheel vibration when a simulated collision was close to happening. The volume of the noise or intensity of the vibration would scale with the confidence level of the system. We seek to use a similar framework in cue presentation that is independent of real-time quantification of human trust state. However, unlike cognitive cues, our proposed method does not aim to directionally steer users toward one decision. Additionally, our proposed method takes into account situational changes caused by both updates to system confidence information and changes in task type.

## III. Situational Trust Cues

We propose STCs as a method to actively assist users in appropriately recalibrating their trust in an automated system throughout continuous usage in a dynamic domain. Cues are presented to users based on situational changes throughout task completion that may affect the performance and reliability of the automated system. Situational changes occur in two categories: 1) task type, and 2) environmental conditions. STCs are presented solely based on situational updates and are not reliant on real-time measurement of user trust, unlike TCCs. Therefore, STCs are dependent upon users' prior knowledge of the limitations of the automated system, but they do not require continuous quantification of trust in the system, instead empowering the user to assess their own level of trust and leverage their understanding of the system to make adjustments accordingly. The researched impact of confidence information previously discussed suggests that humans have the ability to assess and adjust their usage of a system based on system qualities and performance. STCs are non-directional in that they do not inform the user in which direction they need to adjust (i.e. trust the system less or trust the system more). Instead, STCs only suggest that the user recalibrate trust in the system based on situational updates. STCs expand the application of TCCs to scenarios where real-time calibration of user trust in an automated system is not available through measurable behavioral indicators, for example in scenarios where users are not explicitly delegating an automated system a certain level of responsibility and user reliance on a system is difficult to quantify.

In this work we present an experimental design to assess if STCs are effective in appropriately influencing user trust calibration in an automated decision support tool. We aim to understand if non-directional cues presented adaptively based on situational updates have a significant effect on the level of automation a user will afford an automated decision support tool during task completion. We purposefully utilize an automated decision support tool with three different levels of automation so that the user must make a choice between the three levels, and cueing does not become directional due to a binary decision space. We chose a mission operations scenario for our empirical evaluation. STCs could be applied to an array of aerospace environments in application to both automated decision support tools and general operator interactions with automated systems. Once such application for STCs would be a spaceflight scenario where crewmembers must rely on automated systems to complete tasks. For example, a non-expert crew member may be required to conduct a medical exam using semi-automated instruments. STCs could assist in appropriately calibrating user trust in the system based on previous training to ensure the user relies on the system when appropriate, but seeks additional support from a flight doctor when necessary.

We also seek to understand the limitations of STCs in an operational environment. STCs rely on a user's ability to accurately re-examine the operating scenario and appropriately recalibrate trust after being presented a cue. Therefore, the effectiveness of STCs are dependent on user understanding of system capabilities and performance in different tasks and environmental scenarios. An assessment of the impact of user understanding of system capabilities is included in our proposed experimental framework. Another potential disadvantage of STC relates to the risk of under-cueing and over-cueing. If an initial cue does not affect behavior, there is no feedback into the cueing system to initiate the delivery of a follow-up cue, as in the TCC method. Alternatively, over-cueing, or presenting cues more frequently than appropriate, may degrade a user's ability to select the appropriate level of automation. These effects are important to understand when operationally integrating the method.

## IV. Experimental Design

### A. Overview

This section provides a detailed outline of an experiment designed to assess the effect of STCs on continuous trust calibration when applied to a dynamic mission operations environment. Participants are asked to complete a series of tasks selecting paths for a Lunar rover. Throughout task completion, participants are assisted by an automated decision support tool. At the start of each task, participants select the level of autonomy they would like to afford the decision support tool for that task. Participant selection of automation level provides a behavioral measurement of user reliance on the system and is used as a proxy for participant trust in the decision support tool at that point in time.

The automated decision support tool is simulated using pre-determined solutions generated offline. The experiment is administered as survey available online through the Qualtrics system. This format allows participants to complete the experimental protocol outside the laboratory environment, enabling a relatively large sample size.

### B. Path Planning Tasks

The experiment is framed as the participant's first day operating a rover on the Lunar surface. The rover is described to the participant using size and weight specifications similar to that of the NASA Volatiles Investigating Polar Exploration Rover (VIPER) rover[9]. Each path planning task is completed using an open data Lunar digital elevation model (DEM)[10], which represents a topographic map of the Lunar surface. The rover's path is planned using a series of waypoints, represented by dots overlaid on the DEM. The participant is responsible for completing a series of high-level strategic path planning tasks, on maps about 4 km by 4 km.

Figure 1 presents an example map, which shows a Lunar DEM with overlaid waypoints. Regions on the map shown in black represent areas where the slope is greater than the rover's maximum traversable slope of 15°, and therefore too steep for the rover to traverse. These areas are referred to as obstacles. If an obstacle exists between any two waypoints, the rover cannot travel between the two waypoints. The designated start and end waypoints for a task are depicted in teal and purple, respectively. Participants complete a path planning task by selecting intermediate waypoints between the designated start and end waypoints for the rover to travel between. The rover can only move between non-diagonal, adjacent waypoints.

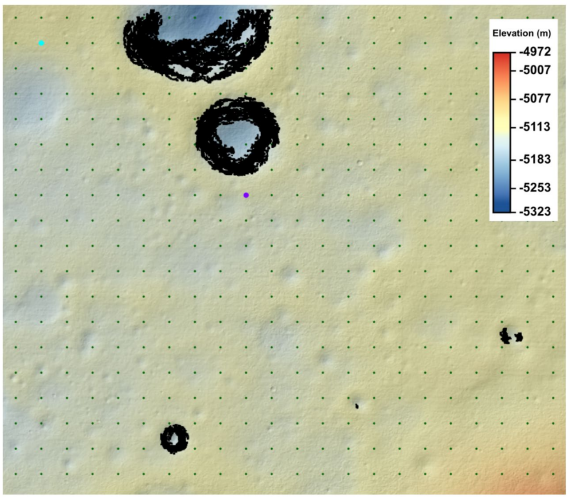



**Figure 1. Lunar DEM with waypoints overlaid. The start (purple) and end (teal) waypoints for an example path planning task are marked.**

Participants are only given visual information about the slopes using the DEM to induce a difficult path planning problem, in which the optimal solution is not obvious. This setup motivates reliance on an automated decision support tool in order to increase performance and reduce inter-operator variance in path selection. Path planning tasks were represented simplistically using waypoints to enable participants to complete a series of many path planning tasks in a timely manner.

The path planning tasks were encoded in the Planning Domain Definition Language (PDDL)[11] for use by the automated decision support tool, which will be described later in the experimental design in more detail. Throughout the experiment, participants completed two types of path planning tasks – traverse tasks and science tasks – each task type with slightly different objectives. Encoding path planning tasks in PDDL additionally allowed for a rich expression of different objectives and executable actions among the two different types of path planning tasks.

*1. Traverse Task*

The objective of the traverse task is to use the waypoints to select a path for the rover to travel between the start and end waypoints that minimizes the total energy expended by the Lunar rover. The total energy expended by the rover (Eq.1) is calculated using the energy rate of the rover and the time the rover spends traveling, calculated using the distance of the selected path and the rover's velocity. The velocity of the rover is assumed to stay constant over all traversable slopes at 0.72 km/h, which reflects the reported top speed of NASA's VIPER rover.

$$\text{total energy [J]} = \text{energy rate [W]} * \text{time [s]} \quad (1)$$

The energy rate of the rover is positively correlated with the magnitude of the terrain slope over which it is traveling. Energy rate equations were adopted from the rover energy rate equations derived for the Surface Exploration Traverse Analysis and Navigation Tool (SEXTANT), an integrated traverse planner and analysis tool[12]. The energy rate equations were formulated using historical Lunar Roving Vehicle (LRV) data[13] and normalized to Lunar gravity. Equation inputs were altered to resemble reported VIPER rover mass (430 kg), and estimated average electronics energy consumption rate while traversing (150 W). Given the complex nature of the energy rate equations, a simplified form is generated by combining constant variables for presentation to participants. The energy rate equation provided to participants is shown below in Eq.2, where α refers to the terrain slope in degrees.

$$\text{energy rate}(\alpha) = \begin{cases} 171, & \alpha = 0 \\ 171 + 2.3 * \alpha, & \alpha > 0 \\ 171 + (-0.7 * \alpha), & \alpha < 0 \end{cases} \quad (2)$$

Providing a simplified version of the cost function allows participants to focus on the key takeaways important to path planning tasks: both traveling uphill and downhill increase the rover's energy rate, but downhill less so than uphill. Participants are also informed that it is not possible for the rover to run out of energy. Having an understanding of the rover's energy rate allows participants to make informed decisions when selecting a path for the rover.

*2. Science Task*

Similarly, the science task requires the participant to select a path for the Lunar rover between the start and end waypoints; however, participants must additionally select three science objective waypoints for science sample collection. The three selected science objective waypoints must be included on the path selected between the start and end waypoints. In every science task, each waypoint on the map will be given a "scientific cost", designating the overall cost to mission objectives if the waypoint is selected for science sample collection. Low scientific cost means a waypoint is scientifically interesting to the mission. Participants are informed that the scientific costs of waypoints are generated by scientist colleagues to provide information about the most valuable points to select as science objective waypoints for science sample collection. The smaller the scientific cost of the waypoint, the more valuable science sample collection at the point will be. Therefore, it is important to select scientific objective waypoints with the lowest scientific cost. The scientific value of waypoints is described in terms of a cost to allow for the representation of science objective waypoint selection as a cost function to be minimized by the automated system.

The scientific cost of each waypoint is depicted by a gradient color grid map overlaid on the Lunar surface map. The square color denotes the cost of each waypoint, with darker squares signifying more valuable waypoints with smaller scientific costs. Figure 2a shows a sample map provided for a science traverse task. In each science task, there will be nine designated science points of interest, clustered in three groups of three, each with costs between zero and four. All other waypoints with no color not included in these groupings have a cost of five. Of the three groups of science points of interest, one group will always be the most scientifically valuable, with waypoint scientific costs of zero or one, while waypoints in the other two groups will have scientific costs of two and higher.

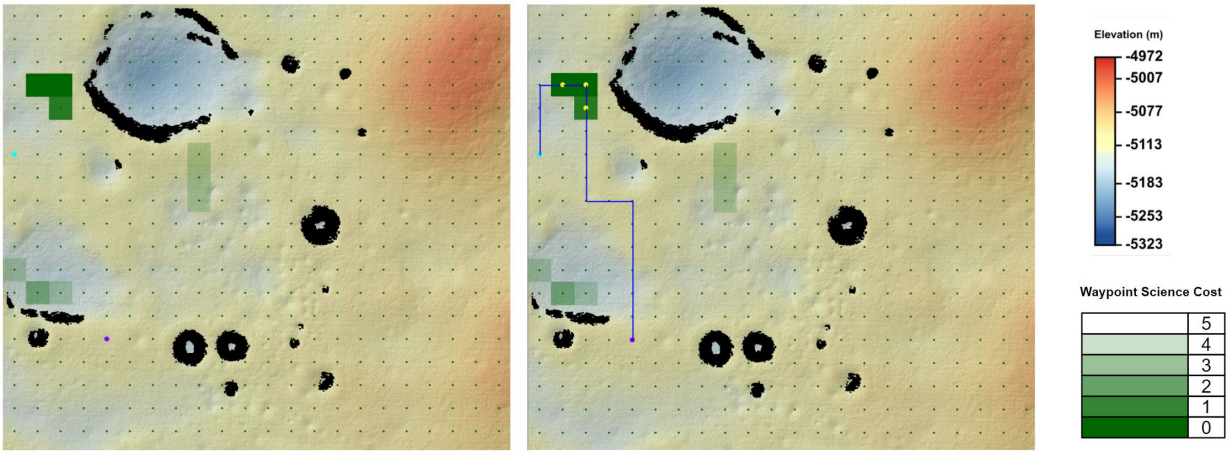
The main goal of the science task is to maximize scientific return by minimizing the total science cost. The total science cost will be measured as the sum of the scientific cost of the three science objective waypoints selected for science sample collection. The secondary goal of the science task is to minimize total rover energy expended (the same goal as in the traverse task); however, this goal is modeled as much less important than maximizing scientific return. The importance of the goals is represented by weighting each in a total cost function equation (Eq. 3). The secondary goal to minimize total rover energy expended is introduced to favor the shortest path to the destination that visits the most valuable science objective waypoints. This provides a metric to differentiate between paths that select the same science objective waypoints.

$$\text{total cost} = 1000000^* \text{ total science cost} + \text{total energy cost} \quad (3)$$

The objective of the science task is to minimize this equation for total cost. Energy cost is equivalent to the total energy calculated in the same manner as in the traverse task. Figure 2b depicts the optimal solution to the science task in 2a, where yellow dots represent selected science objective waypoints.

### C. Automated Decision Support Tool

Participants are assisted in their completion of path planning tasks by a simulated automated decision support tool. The tool provides support by generating plans to the two types of path planning tasks using a PDDL encoding of task goals and objective functions. The back-end PDDL solver employed by the tool is Scorpion[14], an optimal classical planner based on the FastDownward planner. Scorpion is able to produce optimal plans that satisfy problem requirements, such as starting at the start waypoint and ending at the end waypoint, while minimizing an overall total cost function.



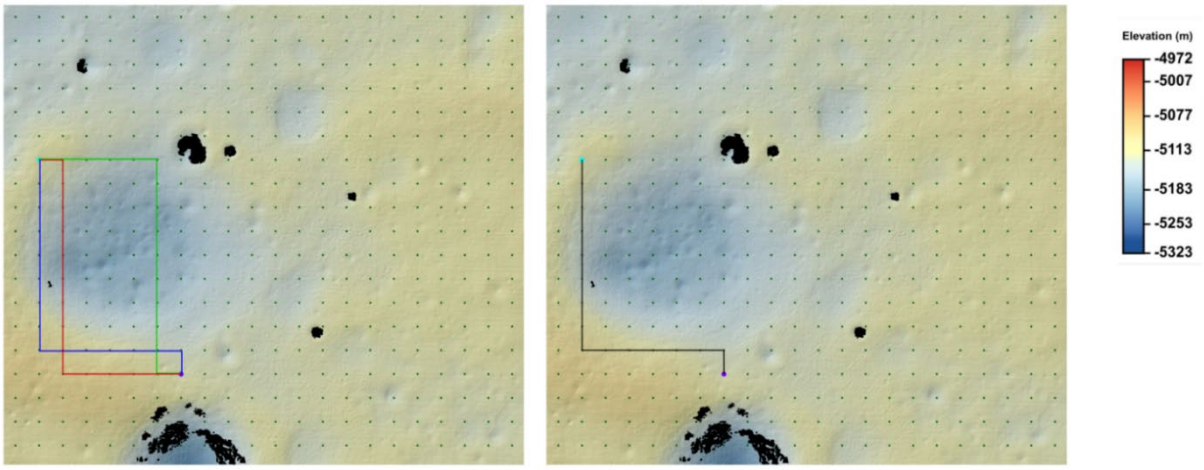

a) Automation Option 2 (Choices) b) Automation Option 3 (Decision with Veto)

**Figure 2. Example science task and optimal solution. Green shading demonstrates scientific cost of each waypoint. Scientific objective waypoints selected by the optimal solution for science sampling shown in yellow.**

At the start of each path planning task, participants choose the level of automation to afford the automated decision support tool. The tool has three available options of automation, each option corresponding to a different level within Sheridan's levels of automation[15]. Three levels of automation were selected to prevent directionality of cueing as a byproduct of a binary decision space. Only three levels of automation were selected to limit task conditions and ensure appropriate justification for each level as the most appropriate option under a simulated operational scenario could be achieved. The first automation option, Automation Option 1, corresponds to Sheridan's first level of automation, which describes a scenario in which the computer provides no assistance, and all decision-making and action is left to the human. Therefore, Automation Option 1 describes a scenario in which the participant manually completes the path planning task without any assistance from the decision support tool. Automation Option 2 corresponds to Sheridan's fourth level of automation, in which the computer assists in decision-making by narrowing down possible actions to a subset of a few actions for the human operator to select from. Selection of Automation Option 2 will provide the participant with a choice between three potential path planning task solutions from which the participant will choose one. The third and final automation option, Automation Option 3, corresponds to Sheridan's sixth level of automation. This level describes a scenario in which the computer makes a decision about the action to take and automatically executes the action unless the action is vetoed by a human operator. This automation level is simulated by having the decision support tool select a path planning task solution and execute the solution unless the participant vetoes the path selection. In the case the participant vetoes the chosen solution, the participant moves on to the next task. Representations of the two task types within the decision support tool are formulated so that different automation options are the most appropriate for the two task types if the decision support tool is working optimally. For the remainder of this section, description of the behavior of the decision support tool will assume optimal functioning.

When completing the traverse task, the decision support tool is given complete information of the path planning problem. The decision support tool has access to the elevation data from the DEM, knowledge of the obstacles, knowledge of the rover initial state and end goal state, and knowledge of the traverse task cost function. With complete information the back-end planner is able to produce an optimal solution to the traverse task. If Automation Option 3 (Decision with Veto) is selected when the participant is completing a traverse task, the decision support tool will select and execute the optimal solution. If the participant selects Automation Option 2, the decision support tool will present the optimal path as well as two additionally generated, non-optimal paths. The two additional paths are generated by the planner by altering the PDDL description of the problem to ensure the second and third solutions travel between different subsets of waypoints. If the participant selects Automation Option 1, they will be prompted to manually select a path between the start and end waypoint. Figures 3a and 3b depict maps shown to a participant after the selection of Automation Option 2 and Automation Option 3, respectively. Given the planner will generate an optimal plan based on the defined task objective, the most appropriate automation option selection will be Automation Option 3. While Automation Option 2 provides the optimal solution as one of the possible choices, this option also provides the chance for the participant to choose a non-optimal solution, and therefore is deemed a less appropriate choice than

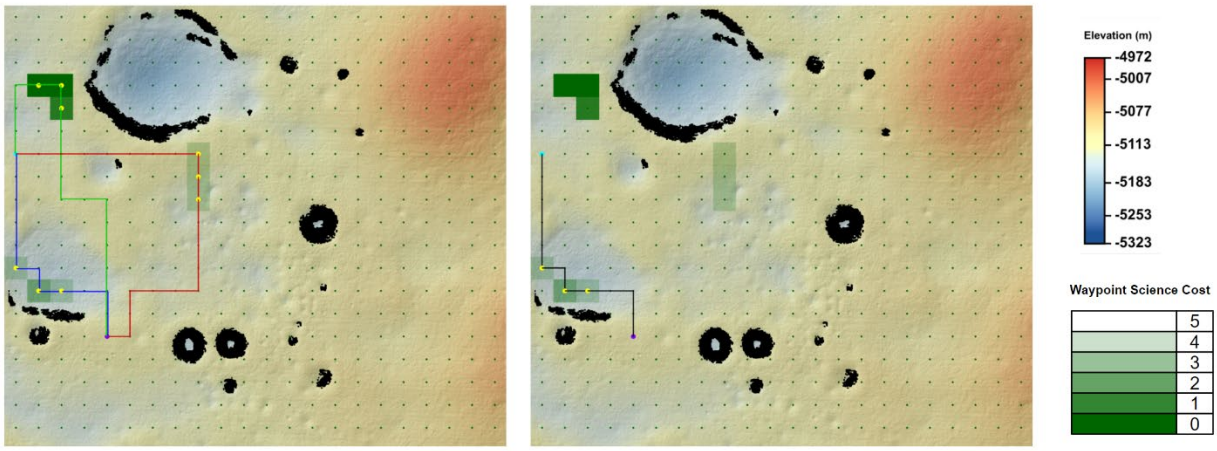
Automation Option 3 in terms of task performance. The planning task was also intentionally structured so that a manually-generated solution produced by choosing Automation Option 1 will likely be non-optimal. The cost function humans are attempting to minimize is both complex and dependent on slope information, which is not directly provided to the human planner. Instead humans are only given visual information regarding map elevation and left to extrapolate slope information. A pilot study of 12 participants demonstrated that while human planners are capable of manually producing near-optimal plans for certain tasks, plans produced by humans were not optimal on average; the costs of generated human plans also showed high variance. Therefore, Automation Option 1 is the most appropriate option to guarantee the optimal plan is selected as well as reduce inter-operator variability.




a) Automation Option 2 (Choices) b) Automation Option 3 (Decision with Veto)

**Figure 3. Maps presented to participant upon selection of Automation Option 2 or Automation Option 3 when completing a traverse task. Automation Option 3 selects the optimal path, while Automation Option 2 presents a choice between three paths, one of which is the optimal path (represented in blue here).**

The science task is formulated so that the decision support tool has incomplete information about the costs of science objectives. This reflects a scenario where input from real-time scientific analysis provides the human operator with more complete information about the situation than the decision support tool. When completing a science task, the decision support tool will again have access to elevation data from the DEM, knowledge of the obstacles, knowledge of the rover initial state and end goal state, and knowledge of the science task cost function. However, it will evenly value all nine science points of interest, without information of updated costs based on scientific analysis. Therefore, if Automation Option 3 is selected during a science task, the automated decision support tool will not necessarily return an optimal path (Figure 4b). Because the automated planner has information that all science objective waypoints cost the same, it will return the selection of the three science objective points of interest and the associated traverse path that minimizes overall rover energy expenditure. If Automation Option 2 (Choices) is selected during a science task, then the decision support tool generates two additional solutions based on the added constraints that it may not reselect any of the same science objective waypoints (Figure 4a). The additional two generated solutions represent paths that select science objective waypoints in the two other sets of clustered science points of interest while minimizing overall rover energy expenditure. One of these two additional paths will select the actual least-costly group of science objective points while minimizing the overall rover energy expenditure, producing the optimal solution. Therefore, Automation Option 2 becomes the most appropriate automation option when completing a science task, given that it provides participants the ability to utilize information unknown to the decision support tool to select the optimal solution, which is not recognized by the decision support tool as the optimal solution. In the pilot study, we found that some participants were capable of manually producing near-optimal paths for science tasks. This is unsurprising, given the weight placed on the selection of science objective waypoints makes any path that selects the least costly science objective waypoints near-optimal. However, variance among path scores selected by participants remained high, making Automation Option 2 a better choice than Automation Option 1 in order to both ensure the optimal path is selected and reduce inter-operator variance.




a) Automation Option 2 (Choices) b) Automation Option 3 (Decision with Veto)

**Figure 4. Maps presented to participant upon selection of Automation Option 2 or Automation Option 3 when completing a science task. Automation Option 3 selects a non-optimal path, while Automation Option 2 presents a choice between three paths, one of which is the optimal path (represented in green here).**

### D. Conditions

In addition to changes in task type, two binary environmental conditions will change throughout the course of the experiment to influence the appropriate level of automation for each task. The first environmental condition describes the presence or absence of significant Lunar dust due to disturbance from science activities completed by the rover. Lunar dust has been found to significantly affect camera image quality, with the potential to stick to camera lenses[16]. In this experiment Lunar dust will be simulated to induce the identification of additional, non-existent obstacles by the rover's hazard cameras, which the automated decision support tool will avoid when selecting a path for the rover. Avoidance of non-existent obstacles will result in longer, costlier plans generated by the path planner. Information regarding this environmental condition will be presented to the participant as constant hazard camera system confidence information, which will be either low or high for a specific task. If significant Lunar dust is present, hazard camera system confidence will be low. When hazard camera system confidence is low, paths manually selected by human operators will generally be better optimized than plans generated by the automated decision support tool, and Automation Option 1 will be the most appropriate automation option selection. This assumption was confirmed in our pilot study. The second binary environmental condition represents whether it is Lunar day or Lunar night. This condition presents a spurious change that will not have an effect on the most appropriate automation option selection for a specific task, but is necessary to include into experiment design in order to assess potential effects of over-cueing.

**Table 1 All possible combinations of task and environmental conditions and their corresponding most appropriate automation option.**

| | High Hazard Camera System Confidence | | Low Hazard Camera System Confidence | |
|---|---|---|---|---|
| | Lunar day | Lunar night | Lunar day | Lunar night |
| Traverse Task | Automation Option 3 (Decision with Veto) | Automation Option 3 (Decision with Veto) | Automation Option 1 (Manual) | Automation Option 1 (Manual) |
| Science Task | Automation Option 2 (Choices) | Automation Option 2 (Choices) | Automation Option 1 (Manual) | Automation Option 1 (Manual) |

Table 1 shows each of the possible combinations of task and environmental conditions and the corresponding most appropriate automation option. When hazard camera system confidence is high, the most appropriate automation option is dependent on the type of task the participant is completing. If the participant is completing a traverse task (automated decision support tool has complete information), then Automation Option 3 is the most appropriate, affording the decision support tool the highest available level of automation. If the participant is completing a science task (automated decision support tool has incomplete information), the most appropriate automation option is Automation Option 2, which allows the participant to inject their knowledge of the most valuable science points by choosing a path among three path choices. However, when hazard camera system confidence is low, the most appropriate automation option selection is always Automation Option 1, representing the option in which participants complete the planning task completely manually with no assistance from the decision support tool. Whether it is Lunar day or Lunar night has no effect on the most appropriate level of automation.

### E. Procedures

A between-subjects experiment will be conducted with a goal of sixty total participants. Twenty participants will be randomly assigned to each of three groups: STCs group, No STCs group, and over-cued STCs (O-STCs) group. All participants will be presented with constant hazard camera system confidence information. Participants in the STCs group will be exposed to STCs, participants in the No STCs group will experience no STCs, and participants in the O-STCs group will be exposed to over-cued STCs. The O-STCs group is exposed to over twice the number of cues presented to the STCs group through the additional presentation of unnecessary cues when changes occur that do not affect the most appropriate automation option. The decision to conduct a between-subjects experiment was made to mitigate learning effects within each of the condition types.

At the start of the survey all participants will be oriented to the specifications and practical constraints of the rover. The domain and objectives of both traverse tasks and science tasks will then be presented and participants will practice manually completing both types of planning tasks. Next, participants will be introduced to the decision support tool and the three automation options they can select. Information will be provided about the performance of the decision support tool to allow participants to appropriately calibrate trust in the tool at the beginning and throughout the experiment. They will be told that when hazard camera system confidence is high, the decision support tool produces optimal plans based on the information available to it. Additionally, during traverse tasks the planner has complete information about the task, but during the science traverse tasks the planner has incomplete information about waypoint science costs. Participants are also informed that plans manually generated by human operators are better optimized than plans generated by the decision support tool when hazard camera system confidence is low.

Participants will complete a series of 24 path planning tasks after initial instruction and training, for an estimated survey time of 90 minutes. At the beginning of each task, participants are presented with the task type, task map, and environmental conditions, and asked to select a level of automation to afford the decision support tool. Based on the selection made by the participant, the next screen will either allow the participant to manually input a path, select between three path choices, or view a selected path with the ability to veto. If the participant chooses Automation Option 3 and subsequently chooses to veto the path, the participant will continue to the next task. Regardless of the participant's selection to veto or not, Automation Option 3 will be recorded as their selection for that task.

Within the 24 path planning tasks, half of the tasks will be traverse tasks and half of the tasks will be science tasks. It is identified that the respective difficulty of each task may be governed by the map on which the task presented. When constructing the tasks, large DEMs were divided into smaller sub-maps on which tasks are presented. To account for equal difficulty among science and traverse tasks, each large DEM was split up evenly into science and traverse tasks.

All participants will be presented the same 24 tasks. To account for potential effects of learning, participants will be shown the tasks in two orders. Participants in the second ordering condition will be presented with a survey where the first and second half of the tasks are swapped. This results in six possible survey conditions derived from three cue conditions and two ordering conditions.

Out of the 24 path planning tasks, each of the three automation options will be the most appropriate selection for eight of the tasks. Therefore, if the participant selects the most appropriate automation option for each task, they will evenly select the automation options by the end of the survey. In order to prevent cueing between every task, tasks with the same most appropriate automation option will be chunked together in two groups of three and one group of two for each automation option. All tasks in the same chunk will have the same most appropriate automation option. These chunks were purposefully distributed between the first and second half of the survey so different most appropriate automation options were spread evenly throughout the experiment. The chunks are also ordered so that in both ordering conditions, the survey will start in an environment where hazard camera system confidence information

is high. This design decision was made to prevent immediate distrust of the decision support tool and begin the experiment in a setting most representative of a majority of the tasks (only a third of tasks have low hazard camera system confidence information).

STCs will be presented as visual text cues with a message that notifies participants that a situational change was detected and they should take a moment to reconsider usage of the decision support tool. Visual text cues were selected given visual TCCs were previously determined to have the most significant impact on participant behavior over auditory, pictorial, or anthropomorphic cues[17]. The selected wording for the presented cues (Figure 5) is intentionally vague to promote appropriate assessment of decision support tool usage without implying the type of operational change that has occurred or the correct next automation option. This demands users reassess the situation themselves in order to make the most appropriate selection. At the start of the tasks participants in both the STCs and O-STCs groups are informed they may see these informational pop-ups, and are instructed to use them to inform their task completion.

Cues will be presented at the start of a task, before participants are prompted to select an automation option. In the STCs group, a cue will be presented at the beginning of every task chunk except the first task chunk, because the start of each task chunk signals a new automation option becoming the most optimal. In total, eight cues will be presented to the STC group. In the O-STCs group, additional cues will be presented within each task chunk when a spurious change, such as a transition from Lunar day to Lunar night, occurs. This will introduce nine additional cues for a total of 17 cues presented to the O-STCs group.

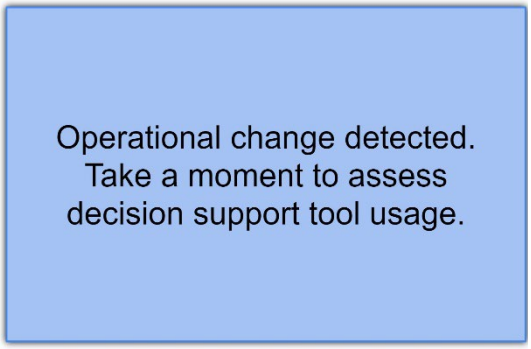



**Figure 5. Cue presented to participants in STC and O-STC groups.**

One important aim of this experiment is to understand how a participant's understanding of the automated decision support tool affects their ability to appropriately calibrate and recalibrate their trust in the automated system throughout usage. To support this aim, participants will be asked a set of assessment questions at the beginning, middle, and end of their completion of the 24 path planning tasks to quantify their understanding of the decision support tool operating capabilities. The purpose of these questions is to quantify participant's understanding of the limitations of the decision support tool, and under what conditions each automation option is the most appropriate selection. Additionally, these questions will allow us to quantify the effects of learning by participants throughout the experiment. The two ordering conditions, which swap the order in which the survey halves are presented, will allow us to account for participants gaining different knowledge about the automated system in the first half and second half of the survey. Within the three cueing condition groups of twenty participants each, ten participants will be in the first ordering condition and ten participants will be in the second ordering condition.

### F. Hypotheses

This experiment is designed to investigate the following hypotheses. They are focused on assessing the impact of STCs on the ability of the participant to continuously calibrate trust, represented by the rate of appropriate automation level selection. The hypotheses include assessments of the impact of participants' "system understanding" on the effectiveness of trust cues. An aggregate score of participant "system understanding" will be calculated from the assessment questions about participant perception of the tasks and expected performance of the decision support tool. Two hypotheses are also included that focus on the effects of O-STCs.

**[H1]** Participants presented with STCs will select the appropriate level of automation at higher rates than participants presented only with constant system confidence information.

**[H2a]** Participants with high system understanding will select the appropriate level of automation at higher rates.

**[H2b]** System understanding will be positively correlated with performance on planning tasks.

**[H3a]** Participants presented with O-STCs will select the appropriate level of automation at lower rates than participants presented with STCs.

**[H3b]** Participants presented with O-STCs will select the appropriate level of automation at higher rates than participants presented only with constant system confidence information.

### G. Data Analysis Approach

The independent variable in this study represents the level of cueing the participant is exposed to with three possible levels: No Cueing, STCs, and O-STCs. All participants will have access to constant system confidence information. The impact of each cueing level will be assessed through the measurement of the percent of appropriate automation option selections. Additional dependent variables measured will include task duration and task performance, where task performance is measured as the similarity of the generated path to the optimal path. In addition, subjective self-assessments of propensity to trust automation will be collected prior to the experiment, and subjective self-assessments of trust in the system will be collected post- experiment.

Linear mixed effects models will be implemented for data analysis, with a separate model for each response variable. Each model will include level of cueing and task type as fixed effects, with participant identifiers included as random effects. Other demographic information may be included in the model if after visual inspection of the data these metrics appear to be responsible for variability within the data set.

## V. Conclusion

Calibrating trust in automated systems remains a major challenge in the implementation of automated systems across aerospace domains. In this work we have described our experimental framework designed to investigate the effectiveness of STCs in reducing the risk of over and undertrust in an automated path planner. STCs represent a novel approach to trust cueing that have potential applications across a diverse set of environments and tasks. These situational cues offer the user information about potential changes in the performance of an automated path planner, but they do not give explicit information on the performance of the planner on each task, which differs from common approaches to trust calibration. In this experiment, we will investigate not only the effectiveness of the cues themselves, but also the impact of the user's understanding of the automated system on the cues' effectiveness. Answering both questions through human participant studies will provide a path forward for designing situation-agnostic automated decision support tools for future teams combined of humans and automated agents. Next research steps are testing the application of STCs in an operational environment, such as in a mission analog setting requiring human supervisory control of a robotic system. This would involve end-to-end integration of cues, including the preliminary assessment of when cues should be presented throughout task completion.

## Acknowledgments

We would like to thank the Solar System Exploration Research Virtual Institute (SSERVI) which partially funded this work as part of the Resource Exploration and Science of OUR cosmic Environment (RESOURCE) project.

## References

1. Frank JD, McGuire K, Moses HR, Stephenson J. Developing decision AIDS to enable human spaceflight autonomy. *AI Mag*. 2016;37(4):46-54. doi:10.1609/aimag.v37i4.2683
2. Okamura K, Yamada S. Calibrating Trust in Autonomous Systems in a Dynamic Environment. *Proc 42nd Annu Meet Cogn Sci Soc*. Published online 2020:1-6.
3. Daronnat S, Azzopardi L, Halvey M, Dubiel M. Inferring Trust From Users' Behaviours; Agents' Predictability Positively Affects Trust, Task Performance and Cognitive Load in Human-Agent Real-Time Collaboration. *Front Robot AI*. 2021;8(July):1-14. doi:10.3389/frobt.2021.642201
4. Verame JKM, Costanza E, Ramchurn SD. The effect of displaying system confidence information on the usage of autonomous systems for non-specialist applications: A lab study. *Conf Hum Factors Comput Syst - Proc*. Published online 2016:4908-4920. doi:10.1145/2858036.2858369
5. McGuirl JM, Sarter NB. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Hum Factors*. 2006;48(4):656-665. doi:10.1518/001872006779166334
6. De Visser EJ, Cohen M, Freedy A, Parasuraman R. A design methodology for trust cue calibration in cognitive agents. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2014;8525 LNCS(PART 1):251-262. doi:10.1007/978-3-319-07458-0_24
7. Okamura K, Yamada S. Empirical Evaluations of Framework for Adaptive Trust Calibration in Human-AI Cooperation. *IEEE Access*. 2020;8:220335-220351. doi:10.1109/ACCESS.2020.3042556
8. Cai H, Lin Y. Tuning trust using cognitive cues for better human-machine collaboration. In: *Proceedings of the Human*

*Factors and Ergonomics Society*. Vol 3. ; 2010:2437-2441. doi:10.1518/107118110X12829370499727
9. NASA. Volatiles Investigating Polar Exploration Rover. Published online 2021:12. doi:10.1511/2022.110.1.12
10. Henriksen MR, Manheim MR, Speyerer EJ, Robinson MS. Extracting accurate and precise topography from Lroc Narrow Angle Camera stereo observations. *Int Arch Photogramm Remote Sens Spat Inf Sci - ISPRS Arch*. 2016;41(July):397-403. doi:10.5194/isprsarchives-XLI-B4-397-2016
11. Fox M, Long D. PDDL2.1: An extension to PDDL for expressing temporal planning domains. *J Artif Intell Res*. 2003;20:61-124. doi:10.1613/jair.1129
12. Johnson AW, Hoffman JA, Newman DJ, Mazarico EM, Zuber MT. An integrated Traverse Planner and Analysis tool for planetary exploration. *AIAA Sp Conf Expo 2010*. Published online 2010. doi:10.2514/6.2010-8829
13. Heiken GH, Vaniman DT, French BM. *Lunar Sourcebook, A User's Guide to the Moon*.; 1991.
14. Seipp J, Keller T, Helmert M. Saturated cost partitioning for optimal classical planning. *J Artif Intell Res*. 2020;67:129-167. doi:10.1613/jair.1.11673
15. Sheridan TB. *Telerobotics, Automation, and Human Supervisory Control*.; 1992.
16. Gaier JR. The Effects of Lunar Dust on EVA Systems During the Apollo Missions. *Nasa/Tm-2005-213610/Rev1*. 2007;(March):1-16.
17. Okamura K, Yamada S. Adaptive trust calibration for human-AI collaboration. *PLoS One*. 2020;15(2). doi:10.1371/journal.pone.0229132