Quantifying Trust in Human-Robot Interaction for Advanced Air Mobility Systems

Darya Zanjanpour
University of Toronto, Institute for Aerospace Studies
darya.skywalker@robotics.utias.utoronto.ca

Sana Kokate
University of Toronto, Department of Psychology
sana.kokate@mail.utoronto.ca

Hugh H.T. Liu
University of Toronto, Institute for Aerospace Studies
hugh.liu@utoronto.ca

Jason E. Plaks
University of Toronto, Department of Psychology
jason.plaks@utoronto.ca

***Abstract*—Trust is a foundational element in human interaction with Unmanned Aerial Vehicles (UAVs). However, quantifying trust has posed a significant challenge. This research aims to establish a framework for defining and quantitatively measuring trust within the domain of Advanced Air Mobility (AAM), integrating engineering and psychological perspectives. The primary objective of this study is to design empirical experiments aimed at evaluating and measuring trust during unmanned aerial vehicles or drone operations. To achieve this goal, we categorize trust dimensions and use line integrals for quantitative trust assessment. The second objective of this research is to capture the calibration of trust levels for users, enabling them to place an appropriate level of trust in a system based on the system's capabilities. The results confirm the possibility of quantitatively measuring trust, highlighting experimental participants' increasing reliance on the system's capabilities over time. Future research will focus on further experimentation to comprehend the impact of various factors on trust.**

***Index Terms*—UAV, AAM, Human-Robot Interaction, Trust, Safety, Performance**

I. Introduction

Trust is a basis of human interaction with Unmanned Aerial Vehicles (UAVs) or Unmanned Aerial Systems (UAS), particularly in the growing field of Advanced Air Mobility (AAM) systems. AAM, defined as a safe and automated air transportation system, has seen rapid advancements in UAS technology [1]. One of the significant challenges in this domain is guaranteed obstacle clearance or collision avoidance from a safety perspective. Consequently, operators find themselves in complex and demanding roles, relying on their own judgment to perform a safe and efficient operation while avoiding collision. Autonomous or semi-automated flight may help to address the concern, to the extent that the machine's autopilot function is perceived to be trustworthy.

Quantifying trust and determining the optimal level of trust in automation represent significant challenges in the automation industry. Addressing these challenges requires further research and experimentation to develop effective trust models. These models are essential for ensuring the design of safe, efficient and reliable automated systems [2].

979-8-3503-1579-0/24/$31.00 ©2024 IEEE

II. Related Works

As depicted in Figure 1, the relationship between Automation Capacity and Trust is central. Drones often operate in close proximity to various objects when performing mission tasks such as surveys, mapping, and delivery. This situation may force operators to override automation, inadvertently elevating the risk of improper actions, thus jeopardizing overall system safety and reliability [3]. Conversely, excessive trust in automation can also lead to issues as operators may fail to recognize system misbehaviour, leading to failures to prevent preventable accidents [4]. This underscores the critical need for a balanced approach to automation, requiring operators to make informed decisions regarding system reliance and ensuring that users maintain trust levels in agreement with the system's capabilities.

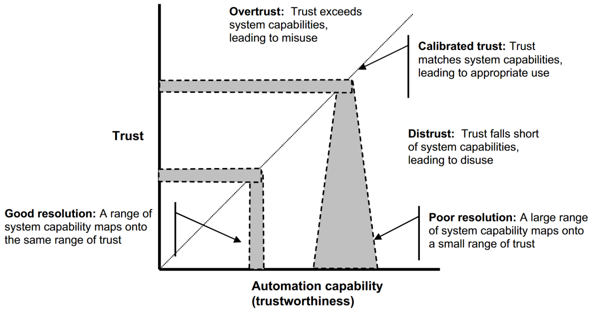
Overtrust: Trust exceeds system capabilities, leading to misuse
Calibrated trust: Trust matches system capabilities, leading to appropriate use
Trust
Distrust: Trust falls short of system capabilities, leading to disuse
Good resolution: A range of system capability maps onto the same range of trust
Poor resolution: A large range of system capability maps onto a small range of trust
Automation capability (trustworthiness)

Fig. 1: An illustration of the relationship between Automation Capacity and Trust [5].

The study of human-robot trust emphasizes the importance of transparency in automation's capabilities, allowing users to calibrate their trust levels accurately. Achieving the correct level of trust requires users to possess "Calibrated trust," aligning their trust with the system's capabilities [4].

Previous research has explored trust in human-robot interaction. O'Neil [6] emphasized the distinction between trust and being trustworthy, highlighting the need for trust to be earned. Recognizing the need for trust to be earned, it is understood that a single experiment with a single set of measurements cannot solely reveal the level of trust an operator places in a system; rather, it measures the system's trustworthiness. To address this, our experimental method implemented multiple runs to provide operators with opportunities to earn trust.

Psychological research conducted by Plaks [7] emphasizes the significance of psychological factors in shaping trust, factors such as humanness, autonomy, and emotions. Acknowledging the substantial role that psychological factors can play in influencing an operator's trust in a system, it becomes apparent that our experimental design must prioritize simplicity and naturalness, avoiding any interventions that could potentially trigger emotional responses from the operators. This careful approach is crucial in ensuring that our 'baseline' measurements remain unaffected by emotional factors, allowing us to gauge trust in its purest form.

The classification of trust into three distinct categories of 1) Rational, 2) Affective and 3) Normative provides a comprehensive framework for understanding its multi aspects nature. Within the Rational category, trust finds its foundation in the process of logical decision-making, as trustors carefully consider past actions and evaluate the advantages and disadvantages associated with the trustee. Affective trust, in contrast, delves into the realm of emotional bonds, as it is predicated on the belief that the trustee holds goodwill towards the trustor. This form of trust places a strong emphasis on the underlying motivations of the trustee. Lastly, Normative trust is based on moral values and ethical principles. Trustors who subscribe to this category anticipate that the trustee's actions will consistently align with these shared values [8]. Applying these trust classifications to UAV-human trust presents challenges due to the absence of emotions and moral values in UAVs. Therefore, our research primarily centers on rational trust, where individuals make logical decisions based on prior actions, advantages, and disadvantages.

Measuring a qualitative feature of a system, such as trust, is challenging but achievable through various methods. Different aspects of trust between humans and the system can be measured [9] [10]. Logical trust, based on rational decision-making, has been explored previously. Freddy et al. [11] associated the number of operator overrides of automation as an indication of distrust in the device's autonomous functioning. [12] They introduced a variable, Relative Expected Loss (REL), to assess this type of trust. In a related study, de Visser et al. [13], assessed users' total time spent flying in manual mode and similar parameters. The results indicated that parameters related to performance are acceptable metrics to assess trust. In this study, we aim to measure logical trust in greater depth by operationalizing the concepts of optimality and efficiency. Our goal is to quantify and measure rational trust by comparing past actions to current ones and conducting multiple runs to quantify changes in users' trust with repeated experience [14].

III. Methods

We note that the fundamentals explored in this research are not limited to our specific experimental paradigm. The primary objective is to develop tools capable of quantitatively measuring trust.

Although we implemented our concepts in a MATLAB simulation, we suggest that the methods described here are generalizable across different experimental settings, whether simulations or actual flight tests. These methods and analytic techniques allow researchers to assess trust levels within their own customized experiments, thereby opening doors to innovative approaches for enhancing and calibrating trust in a wide range of contexts.

A. Basic Thought Experiment

The primary objective was to mathematically model the development of rational trust. Rational trust relies on a logical decision-making process by the trustor, involving the consideration of past actions and an assessment of the pros and cons of the trustee [8]. Building on this definition and recognizing the distinction between trustworthiness and trust, we designed a method that allows researchers to quantitatively measure earned trust over time. To achieve this, the technique must be repetitive, spanning multiple runs, ensuring that trust is measured cumulatively.

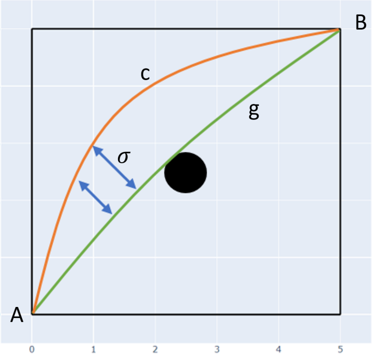

Fig. 2: An illustration of a map with an obstacle located at the center alongside two fight path trajectories. The c trajectory represents a potential flight path. The g trajectory illustrates the shortest flight path, and optimal, trajectory to complete the task.

In our thought experiment, we commence with the most basic interaction involving a drone: flying directly from point A to point B. To introduce the concept of trust into this scenario, we add an obstacle. The task entails navigating the drone from point A to B while avoiding any collision with the obstacle. It's natural to anticipate that the operator will opt for a flight path that maintains a comfortable margin of distance from the obstacle to ensure a safe and collision-free path.

Next, the operators are asked to repeat the task multiple times. They are instructed to prioritize safety and obstacle avoidance while simultaneously striving for an optimal flight path (in terms of distance). We restrict the flight speed to a single setting, meaning that operators can optimize their approach only by flying more directly to the destination. As depicted in Figure 2, the most optimal, direct, trajectory involves choosing a diagonal path that passes directly through the obstacle. However, they must still deviate from the obstacle to ensure no collision occurs. Our hypothesis is that as users gain more experience and consider their past actions, they will gradually become more adept at playing optimally, reflecting increasing higher rational trust. Our expectation is that users will start with a flight path denoted as "c" which maintains a considerable distance from the obstacle to ensure safety. Over multiple runs, we anticipate that their trajectories will converge toward a more optimal trajectory, denoted as "g" as shown in Figure 2. Thus, we operationalize rational trust as the gradual convergence of trajectories over multiple runs. This convergence can be evaluated by measuring change in the distance difference, denoted as D in the following.

$$|c - g| = D$$

In the following sections, we will present our method for measuring this form of trust, which builds on this basic principle.

B. Formulation of Rational Trust

In our methodology, we employ mathematical equations to quantitatively model rational trust. Operators are placed in a scenario in which they must navigate from point (x_1,y_1) to (x_2,y_2), avoiding obstacles, as depicted in Figure 2. We denote the subtraction between the actual trajectory and the shortest possible trajectory as "σ." To determine this difference, we utilize the principles of line integrals. Subtracting the line integrals of these two paths can yield a value that serves as our trust factor:

$$\left| \int_C f(x,y)\,ds - \int_g g(x,y)\,ds \right| = \sigma \tag{1}$$

To facilitate comparisons of σ for multiple runs with varying map settings, such as different optimal trajectory lengths, we use division rather than subtraction between the actual and shortest paths. The division always results in a ratio, enabling us to compare different runs:

$$\left| \frac{\int_C f(x,y)\,ds}{\int_g g(x,y)\,ds} \right| = \sigma' \tag{2}$$

This division between the actual trajectory and the shortest possible trajectory provides an alternative measure of the trust coefficient, denoted as "σ'." Higher trust is associated with a trajectory closer to the optimal path, resulting in a σ' value closer to 1. Conversely, as users deviate further from the optimal path, σ' increases, reaching values significantly greater than 1. As operators engage in more runs, we hypothesize that they will gain increasing trust. We anticipate an inverse relationship between the frequency of task repetition and the value of σ'. The observed negative trend between the number of runs and σ' serves as the final trust measurement, confirming that the quantified value represents the trust earned during these runs.

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(\sigma'_i - \bar{\sigma'})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(\sigma'_i - \bar{\sigma'})^2}} = r \tag{3}$$

To quantify this relationship, we utilize Pearson correlation, denoted as "r", where n represents the total number of runs, x_i stands for the run number.

We hypothesize that the calculated "r" value will be negative. That is, participants' mean deviation from the optimal trajectory will decrease as the number of attempts increases.

C. Simulation Implementation

In the simulation, participants are tasked with flying from point A to B while prioritizing obstacle avoidance, much like operating a real drone. We acknowledge that participant reactions may differ between a simulation and real-world drone operation, but our goal is to test the reliability and robustness of our measurement technique for application in both simulation and real flight tests.

Participants are instructed to fly the shortest path while avoiding collisions in a situated $3D$ MATLAB environment. The simulation is presented to them in a simplified $2D$ top-down view. This simplification minimizes complexity and allows users to focus exclusively on their assigned tasks.

Dijkstra's search algorithm is employed to calculate the shortest path efficiently by iteratively examining nodes with the smallest distance from the source node and adjusting distances to their neighbouring nodes [14].

The task is repeated 30 times to enable participants to gradually build trust. Limiting the runs to 30 prevents participant fatigue, which could impact performance. The starting and ending points are consistent, but random obstacles with varying numbers and radii are introduced in each run to prevent map memorization. To minimize external influences on participants' responses, the simulation maintains uniformity in starting and ending points and map complexity.

D. Experimental Design

Our method allowed us to investigate how much control participants give to a simulated drone that offers both automatic and manual control modes. Experimental participants (n=40) commenced the simulation in automatic mode, whereby the drone followed a distance-optimized trajectory to reach the endpoint. However, this mode carried a notable risk of collision when encountering obstacles, and participants were unaware of the auto-generated planned trajectory until the end of each run. They understood that the automatic mode followed the most optimal path but also carried an uncertain risk of collision.

Conversely, the manual mode allowed participants to strategically place waypoints, guiding the drone along a straight path to each point. Participants could pause the game by pressing the spacebar, enabling them to switch between automatic and manual modes or to continue in their current mode.

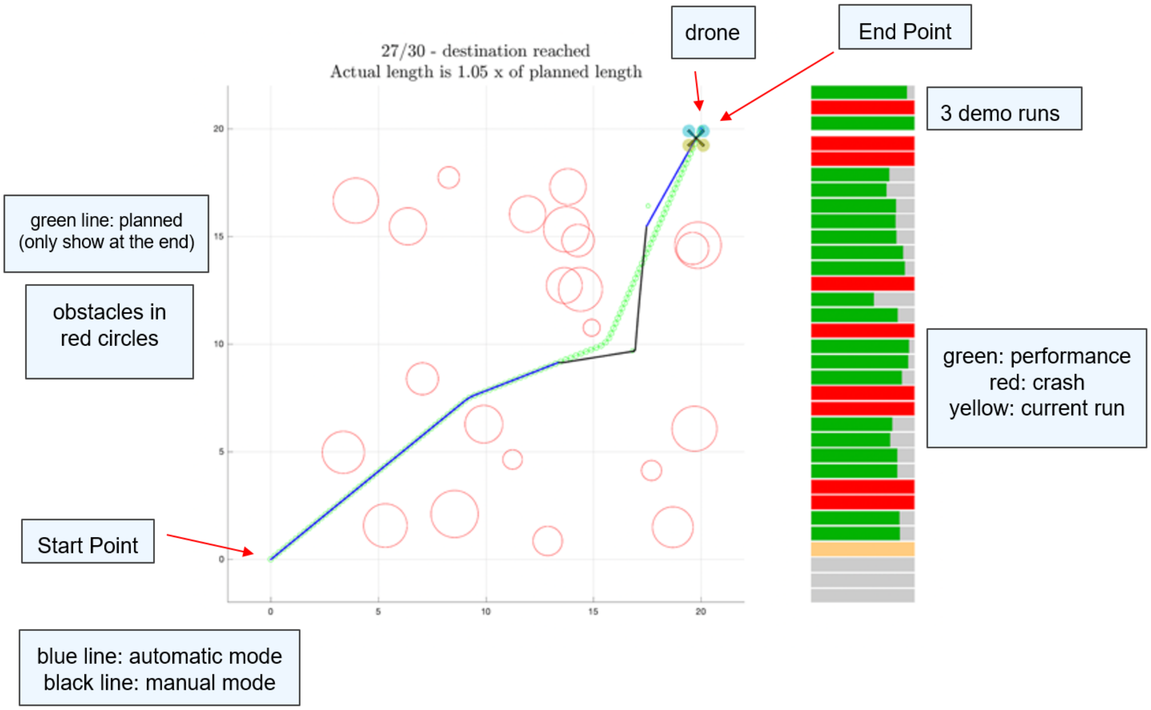

Fig. 3: The simulation panel used in the experiment featured a consistent starting point at (0,0) and an endpoint at (30,30) for all participants across various maps. Red circles represented obstacles of varying locations and sizes.

As illustrated in Figure 3, the green line represents the most optimal path starting from the initial point. The blue segments of the flight path indicate when the automatic mode was active, while the black line signifies the time when the manual mode was engaged.

Importantly, participants had the freedom to switch between automatic and manual control modes. All participants were provided with the same set of maps for navigation. The flight trajectory combines both automatic and manual modes. In the preceding section, we referred to participants' chosen trajectory as the "actual trajectory." In contrast, the green trajectory, labelled as the "planned trajectory," represents the most optimal path.

As shown in Figure 3, a feedback panel on the left displayed performance indicators: a full red bar for collisions, a yellow bar for the ongoing run, and a green bar for successful runs. The length of the green bar reflected participant performance, with longer bars indicating better performance and shorter bars indicating poorer performance. This visual feedback aided participants in assessing their performance.

To familiarize participants with the game dynamics, they underwent three demo runs before the main data collection phase. Each subsequent run featured distinct obstacle maps. Throughout the experiment, various types of data were collected, including interaction points (spacebar presses), location information, actions taken, distances to the closest obstacle, and the tracked trajectory, represented as a sequence of locations with associated lengths. With this comprehensive dataset, our analysis sought to provide quantitative insights into participant behaviour and to measure trust within the experimental context.

The advantage of using the manual mode is a reduced risk, as participants have the freedom to place waypoints that allow for direct flight. However, the disadvantage is that the drone will likely not follow the most optimal trajectory, as participants tend to create a new trajectory that steers well clear of a looming obstacle. In contrast, the fr-valueautomatic mode offers the advantage of following the optimal path, but it comes with a safety risk. Combining these two modes (by switching between the two) provides participants with the opportunity to balance these two risk contingencies, between safety and optimality.

E. Participants

Data were collected from a sample of 40 participants, primarily composed of undergraduate students enrolled in the introductory psychology course at the University of Toronto. Participants willingly took part in the experiment to earn partial course credit. The study adhered to strict privacy standards and data protection rules, governed by Canadian laws and local regulations related to human subject research and data privacy [16]. No data were collected until the university's Research Ethics Board (REB) granted formal approval. The participants' prior experience with drone operation was not mandatory, as the simulation provided a dynamic 2D view and required no specific expertise. Control commands were limited to selecting points and choosing a strategy to play. To facilitate meaningful comparisons, all 30 maps were maintained identical for every participant.

IV. Results

We assessed σ' across 30 runs by dividing the actual trajectory's length by the planned length. This measurement enabled us to assess the difference between the paths and determine whether, by the end of 30 runs, their actual path converged to the most optimal path, aligning with our expectations of their trust development.

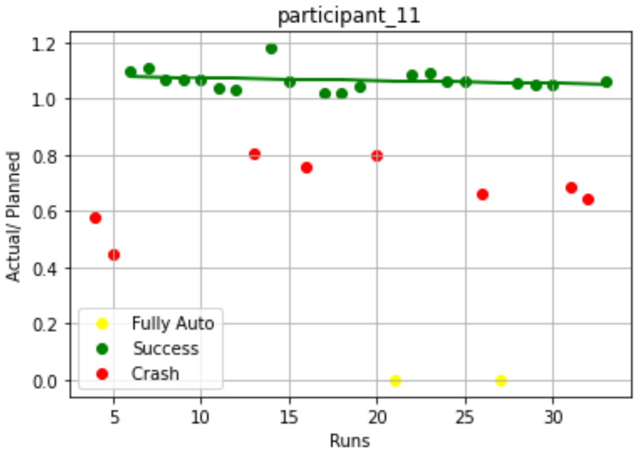

Fig. 4: The"σ' vs. run number" plot for a random participant

Figure 4 serves as a representative example of the data collected for a randomly selected participant, number 11. In this visualization, green dots are successful runs with no crashes, yellow dots represent fully autonomous mode runs where the participant did not switch modes and remained entirely in auto mode, while red dots denote runs where the participant's simulated drone crashed before reaching the destination. The green plotted run indicates a negative trend for successful runs.

When considering only the successful runs for this participant, the decreasing values of σ' throughout the runs indicate that the user is gradually earning trust. A challenge arises when a participant crashes into an obstacle, as indicated by the red data points in Figure 4. In such runs, the actual trajectory is shorter than the planned trajectory, not due to differences in skill or trust, but because the participant was unable to reach the destination and crashed mid-flight. This disparity makes direct comparisons between crash runs and successful runs problematic. To address this issue, we proposed a solution: terminating the planned trajectory to the closest point where the simulated drone crashed. This adjustment allowed us to align the planned length of crash runs with the rest of the data, making the trajectories comparable for analysis.

Figure 5 displays the same dataset from the same participant but with a calibration process applied to the crash runs, which involves adjusting their planned trajectory lengths in relation to the crash points. This adjustment results in a more coherent and meaningful representation of the data. After the adjustment, the crash points align with the overall trend. The red line represents the overall trend of both crashes and successful runs. As depicted, the trend has a significant negative slope, a pattern consistent with increasing trust during the runs. Yellow data points represent autonomous modes where participants handed off control, allowing the system to operate independently. These runs have been excluded from our analysis as they lack direct participant interaction, and we have omitted them from the dataset for further analysis. We applied the same measurement techniques and conducted performance analysis for all 40 participants across the 30 runs, consolidating the results into a single plot as shown in Figure 6. In this final panel, all crash runs have been adjusted, and fully autonomous runs have been excluded due to no interaction. The 30 runs are plotted in order against the σ' for all 40 participants. In each run column, we have 40 data points.

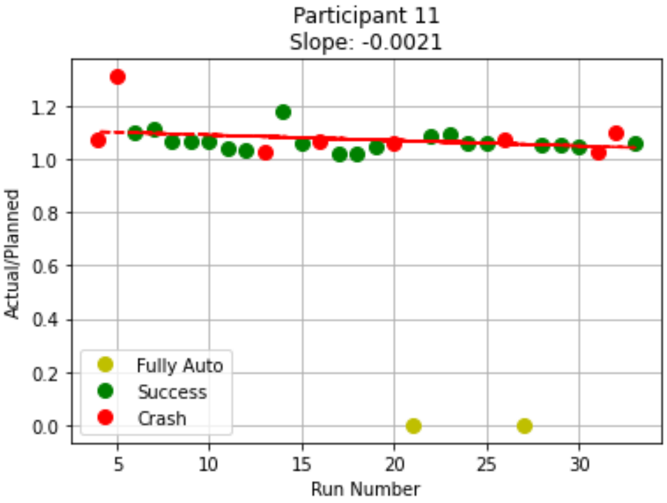

Fig. 5: The"σ' vs. run number" plot for a random participant where crash runs are corrected in relation to the crash points.

The σ' value was calculated for each participant in each run using equation (2) and is illustrated in Figure 6. The average σ' is depicted over the 30 runs for all participants in blue lines, showing a negative trend.

40 participants completed a total of 30 runs each, resulting in a dataset of 1,200 runs. The calculated r-value, using σ' obtained through Equation (2) and serving as input for Pearson's relation in Equation (3), is -0.2287. The p-value, measured for the r calculations, to validate the statistical importance for the same dataset consistently remained below 10^{-12} indicating an extremely low likelihood that this pattern occurred by chance. It is noteworthy that while the slope and average slope could be used to assess the trend of σ', we utilized r-value for easier assessment, considering the variation in slope numbers relative to the measurement setup.

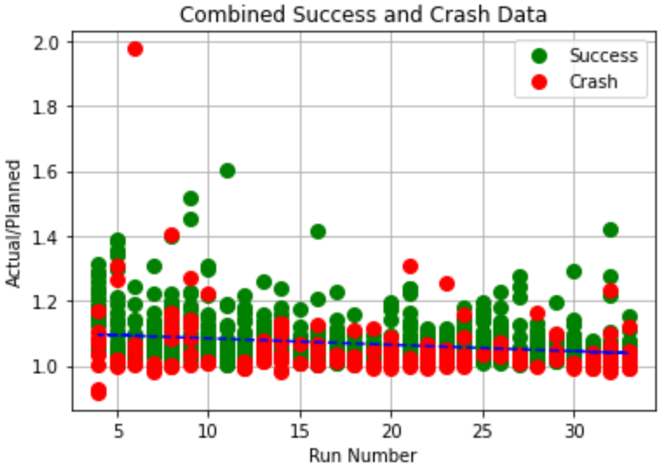

Fig. 6: The σ' vs. run number" plot for all 40 participants where crash runs are corrected in relation to the crash points

V. Conclusion

As depicted in Figure 6, we expected to observe a negative trend in σ' as a function of run numbers. This negative trend suggests that over the course of 30 runs, the actual trajectory that participants take tends to converge toward the optimal path, consistent with the development of rational trust. The trust level is calibrated over the runs by allowing participants to experiment with switching between modes, discovering the right balance of trust and improving their performance toward the optimal trajectory. The r-value is equal to -0.2287. This value aligns with our initial hypothesis. The negative value signifies the convergence of the actual flight path with the planned trajectory over 30 runs. The absolute value of 0.2287 falls within the range considered small to moderate in experimental behavioral science research [17]. The small p-value of 10^{-12} indicates that the likelihood of the results being due to chance is negligible. This supports the reliability of our findings and the suitability of a total of 40 participants for the experiment.

In summary, the statistical analyses provide evidence of this method's ability to quantitatively measure rational trust. They suggest that users consistently improve their efficiency as they grow more acquainted with the task. This approach aims to improve the user experience and ensure trust aligns well with the system's capabilities and performance. Our future work will involve exploring additional methods to understand the influence of various factors on rational trust.

Future Work

The result serves as a robust validation of this method's capability to quantitatively measure rational trust in an autonomous vehicle. As we refine this method to measure rational trust, our future research endeavours will extend to a more comprehensive understanding of trust calibration. This will involve further experimentation to precisely calibrate trust levels. We aim to collect different variables during the data collection process. This will enable us to construct additional measurement tools that assess rational trust while considering more various influencing factors such as "proximity to obstacles," "usage of different modes," and more. We will employ these extra measurement tools to assess rational trust from different perspectives and consider various factors in our experiments. This approach aims to improve the user experience and ensure trust aligns well with the system's capabilities and performance.

Acknowledgment

The authors gratefully acknowledge the sponsorship of the PsychEng (Collaborative Specialization in Psychology and Engineering) program, the University of Toronto.

The authors (Zanjanpour, Liu) would also like to acknowledge that the research work presented here is sponsored by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants, RGPIN-2017-06708 and RGPIN-2023-05148.

References

[1] Patterson, Michael. "Advanced Air Mobility (AAM): An Overview and Brief History." In Transportation Engineering and Safety Conference. 2021

[2] Lee, John and Wickens, Christopher and Liu, Yili and Boyle, Linda. (2017). Designing for People: An introduction to human factors engineering.

[3] Lee, John D., and Katrina A. See. "Trust in automation: Designing for appropriate reliance." Human factors 46, no. 1 (2004): 50-80.

[4] M. Itoh, "A model of trust in automation: Why humans over-trust?," SICE Annual Conference 2011, Tokyo, Japan, 2011, pp. 198-201.

[5] Lee, John D., and Katrina A. See. "Trust in automation: Designing for appropriate reliance." Human factors 46, no. 1 (2004): 55, 2.

[6] O'Neill, O. (2002). Autonomy and trust in bioethics. Cambridge: Cambridge University Press.

[7] Plaks et al, Identifying psychological features of robots that encourage and discourage trust. Computers in Human Behavior, 134:107301, 2022.

[8] Ryan, M. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. Sci Eng Ethics 26, 2749–2767 (2020). https://doi.org/10.1007/s11948-020-00228-y.

[9] Sanders, T., Oleson, K. E., Billings, D. R., Chen, J. Y. C., and Hancock, P. A. (2011). A Model of Human-Robot Trust: Theoretical Model Development. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 55(1), 1432-1436. https://doi.org/10.1177/1071181311551298

[10] Lee J, Moray N. Trust, control strategies and allocation of function in human-machine systems. Ergonomics. 1992 Oct;35(10):1243-70. doi: 10.1080/00140139208967392. PMID: 1516577.

[11] Freedy, Amos and DeVisser, Ewart and Weltman, Gershon and Coeyman, Nicole. (2007). Measurement of trust in human-robot collaboration. 106 - 114. 10.1109/CTS.2007.4621745.

[12] Tenney, Y. J., Rogers, W. H., and Pew, R. W. (1998). Pilot opinions of cockpit automation issues. The International Journal of Aviation Psychology, 8(2), 103-120

[13] de Visser, Ewart and Parasuraman, Raja and Freedy, Amos and Freedy, Elan and Weltman, Gershon. (2006). A Comprehensive Methodology for Assessing Human-Robot Team Performance for Use in Training and Simulation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 50. 10.1177/154193120605002507.

[14] Desai, M. (2007). Sliding scale autonomy and trust in human-robot interaction (Master's thesis). University of Massachusetts Lowell.

[15] Dijkstra, E. W. (1959). "A note on two problems in connexion with graphs". Numerische Mathematik. 1: 269–271. CiteSeerX 10.1.1.165.7577. doi:10.1007/BF01386390. S2CID 123284777.

[16] https://www.sona-systems.com/privacy_sonasite_en_ca/

[17] Cohen J. (1965). Some statistical issues in psychological research. In Handbook of Clinical Psychology, ed Wolman B. B. (New York, NY: McGraw-Hill;), 95–121