INSTITUTE FOR DEFENSE ANALYSES

Predicting Trust in Automated Systems – An Application of TOAST

Daniel J. Porter, Project Leader

Caitlan A. Fealing

July 2022

IDA Document NS D-33188

Log: H 2022-000321

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task C9089, "Trust in Automation," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Dr. V. Bram Lillard and consisted of Mr. Andrew Flack and Dr. John T. Haman from the Operational Evaluation Division.

For more information:
Dr. Daniel J. Porter, Project Leader
dporter@ida.org • (703) 578-2869

Dr. V. Bram Lillard, Director, Operational Evaluation Division
vlillard@ida.org • (703) 845-2230

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-33188

Predicting Trust in Automated Systems – An Application of TOAST

Daniel J. Porter, Project Leader

Caitlan A. Fealing

Executive Summary

The number of people using autonomous systems for everyday tasks has increased steadily since the 1960s and has increased dramatically since the invention of devices that can be controlled via smartphone. Until recently, researchers have been able to gain insights on trust levels only by observing a human's reliance on the system. Researchers needed a method of quantifying how much an individual trusts the automated system they are using.

Trust is "a psychological state or behavior in which one person is willing to make him- or herself vulnerable … because he or she is sufficiently confident that the other person will not exploit him or her" (Nave et al., 2015, p. 774). As applied to automated systems, trust is the state in which a person makes themselves vulnerable because they are confident that the automated system is capable and reliable enough to complete the task. This definition of trust adapted for the autonomy context indicates that a person needs to understand the capabilities, limitations, and performance of a system to trust it.

Humans inherently base their trust in automated systems on the amount of risk associated with failing a task. The fundamental differences across automated systems in potential failure penalties and potential failure frequencies mean researchers have to view trust on a scale that depends on the combination of an expected penalty score and the system's level of accuracy. A lack of either component of this combination should correlate with a person overtrusting or undertrusting the system.

Researchers may estimate levels of trust if they create a situation in which they completely understand the autonomous system and all plausible situations while the other participants do not. To estimate trust levels, the nine-item Trust of Automated Systems Test (TOAST) scale has two main categories of questions: understanding and performance, with four and five questions, respectively. This scale has been initially validated to detect differences in trust among military-affiliated operators, and current research focuses on extending this validation to civilians.

Following Wojton et al.'s (2020) research on TOAST, we aimed to determine how well this scale performs when used outside a military context. Specifically, we sought to answer three research questions:

1. How much does a person's trust in the system change as they use it?
2. How much time does it take for a person to rely on an automated system to the extent that they will?

3. Does the TOAST scale correlate with reliance when participants have to make real choices rather than self-report their expected choices?

We randomized participants into a high understanding or low understanding condition via the text they saw as an introduction to the experiment, and we further randomized them into a high performance or low performance condition according to which version of the automated system designed to help with the task, named Helper (90 percent accuracy or 60 percent accuracy), they would see throughout all trials of the experiment. The experiment contained three phases of trials: Training; Single Trials – Time to Steady State; and Block Trials – Measuring the Steady State. We administered the TOAST scale before training, after the single trials, and after the block trials. In a single trial, the participant had to decide whether to use Helper for each decision, and in a block trial, the participant had to decide whether to use Helper for groups of four decisions.

We found three main results:

1. Participants who used a poorly performing automated system (the version of Helper with 60 percent accuracy) trusted the system less than expected when deciding to use that system for each decision; however, those who used a high performing system trusted the system to the degree expected. Additionally, both participants who used the poorly performing system and those who used the high performing system lost a significant amount of trust after deciding to use the system for groups of four decisions.
2. On average, participants reached their steady state of trust on trial 15.4 out of an average of 49 completed trials, indicating they spent 70 percent of the single trials that they completed in their steady state of trust. We also found a non-significant trend that participants who used the low performance system reached their reliance steady state more slowly than those who used the high performance system.
3. A logarithmic trend line accounts for 78.3 percent of the variance in the average percentage of participants who chose to manually complete the block of trials per penalty associated with an incorrect trial response. Even though the trend line fit the data well, the log of participants' level of steady states of reliance did not correlate with their TOAST scores collected at the end of the experiment.

As observed in the experiment, understanding a well-performing autonomous system is the key to maintaining trust on a case-by-case basis. Maintaining trust that is well calibrated to the system is a key aspect of creating highly effective human-machine teams. Widespread use of the TOAST scale would allow researchers to (1) better predict how users will accept new autonomous systems and (2) determine whether specific human-machine teams are appropriate for the tasks they are designed to complete.

Contents

1. Introduction

The prevalence of people using autonomous systems for everyday tasks has been steadily increasing since the 1960s and has increased dramatically with the invention of devices that can be controlled via smartphone (Marikyan et al., 2019). As people become accustomed to machines affecting more of the tasks they perform in their lives, the general public's reliance on autonomous systems increases (Marikyan et al., 2019). Two notable examples of these autonomous systems are Apple's Siri and Tesla's Autopilot. Both systems are designed to make processes easier for the user, but the possible consequences when these systems act incorrectly differ vastly. Siri and other voice assistants can interact with search engines, send messages, and provide directions to a home address if one has been set (Cervantes, 2021). While those features can be dangerous, Tesla's Autopilot and other similar autonomous driving systems can kill people by crashing the vehicle (Nayak, 2022). Inherently, humans base their trust in automated systems on the amount of risk associated with a perceived penalty (Wojton et al., 2020). Thus, it is more natural for a human to trust a system like Siri more than Tesla's Autopilot because the penalty of the worst-case scenario while using Siri is much less than the penalty of the worst-case scenario while using Autopilot.

A. Trust versus Reliance

The fundamental differences in potential penalties and frequencies of failure across automated systems means researchers have to view trust on a scale that depends on the combination of an expected penalty score and the system's level of accuracy. Trust is "a psychological state or behavior in which one person is willing to make him- or herself vulnerable … because he or she is sufficiently confident that the other person will not exploit him or her" (Nave et al., 2015). As applied to automated systems, trust is the state in which a person makes themselves vulnerable because they are confident that the automated system is capable and reliable enough to complete the task. This definition of trust, adapted to fit the autonomy context, indicates that for a person to accurately trust a system, that person needs to understand the capabilities and limitations of the system while knowing the system's performance. A lack of either of these things should correlate with a person overtrusting or undertrusting the system.

It is not possible to accurately measure trust without having a person respond to a validated scale, but researchers can gain insights into trust levels by observing a person's reliance on the system. As noted by Lee and See (2004), "Trust guides reliance when

complexity and unanticipated situations make a complete understanding of the automation impractical." Given the case in which researchers completely understand the automation and all possible situations while other individuals do not, the researchers should be able to estimate the extent to which a person's trust guides how much they rely on a system.

It is important to differentiate between the concepts of trust and reliance. Trust is an inherently psychological state whereas reliance is a measure of use. "Reliance is the state of being dependent on someone or something" (Merriam-Webster, n.d.). Therefore, reliance on a system does not inherently require the user to be vulnerable and does not require that the user believe the system is capable and trustworthy enough to complete the task. For example, someone may rely on a system because they are bored and do not want to complete the task or because the task is too strenuous for the human to complete alone so that human requires assistance regardless of how trustworthy that assistance is. A person's reliance level is determined by the ultimate degree or amount they will use a system in a given scenario. Reliance level can change between person, system, and type of scenario, but a person's reliance on one specific machine for one type of scenario should remain the same.

While theory states that humans should reach a certain steady state of reliance, that state may not be the ideal level for a system given that system's performance. In order for the user to rely on the system to a degree that is appropriate for the system's demonstrated performance, the user must accurately understand how the system should perform. According to Wojton et al. (2020), the two main components of measuring trust are measuring how much the person understands how the system should work and measuring how well the person expects the system to perform. These researchers specifically include an understanding component because "clarifying the conditions under which the automation is likely to perform as expected" encourages appropriate levels of trust (Wojton et al., 2020).

According to Nourani et al. (2019), these clarifications must also be meaningful to the system's users because "whether explanations are human-meaningful can significantly affect perception of a system's accuracy independent of the actual accuracy observed from system usage." If the users have meaningful explanations that allow them to better understand the system, "users are capable of estimating the system accuracies reasonably well and gradually adapting their trust levels to the system performance within 30 trials" (Yu et al., 2019). Therefore, altering how much a user understands a system would affect that user's perception of how well the system performs.

Additionally, Yu at al. (2019) mention that "70% is the system accuracy threshold that determines whether users will trust and use the system with high self-confidence." Thus, when choosing performance levels for highly accurate and less accurate systems, the high performance system should have an accuracy of greater than 70 percent and the low performance system should have an accuracy of less than 70 percent. In theory, providing

systems with these performance levels will ensure that people using the high performance system have greater trust and reliance than those using the low performance system.

Ideally, a user would reach a level of reliance on a system that would remain constant through infinite amounts of further interactions with that system. To accurately predict a user's reliance on a system, it is essential for that user to reach a steady state of reliance. In their research, Yu et al. (2019) found that "After 25 trials, both the trust level and the perceived system performance reached a stable level, and we can infer that no significant change of them will happen if the user continues interacting with the systems" (Yu et al., 2019). Once the user has reached this steady state, further levels of reliance on the system should not change and it is theoretically possible to measure that exact reliance level by varying the amount of risk associated with a task.

Perceiving risk for a given task is based in how much perceived penalty is associated with that task. Tasks with higher penalty are seen as a bigger risk because there is a greater potential for danger, harm, or loss to the individual. Humans who have a standard risk response prefer to minimize risk whenever possible, within reason. In 2018, Rossi et al. determined that "there is correlation between the magnitude of an error performed by a robot and the corresponding loss of trust by the human towards the robot." Therefore, it is expected that humans who use systems that make errors while performing tasks with large penalties will lose more trust than those who use systems that make fewer errors on similar tasks. Because perceived penalty is such a large contributing factor to perceived risk, it is important that the users fully understand the risk associated with the tasks they are performing with the system. If the user does not believe there is a high risk associated with poor performance on a task, the user is more likely to let the system complete tasks because it is less taxing on the user (Morando et al., 2020).

B. Development and Use of TOAST

A wide variety of jobs and professions use automated systems to help with workload and personal performance. Such systems include automated help chat features that talk to clients and answer questions with little supervision, processing equipment that flags errors, and autopilot systems in aircraft and other vehicles. "Opportunities to automate common workplace processes are everywhere, which is why automation is becoming a common element of every business," says Uzialko (2022) in his article on workplace automation. In addition to using automation in the workplace, average consumers frequently use automated systems (e.g., voice assistants) for everyday tasks (Marikyan et al., 2019). People who play video games typically are familiar with the concept of a sidekick system that helps the player complete the game (Cerny, 2015), and some workers use sorting systems as part of their jobs (Khojastehnazhand et al., 2010). Various amounts of research are dedicated to these types of systems, but little of this research uses a validated trust in autonomy scale to evaluate how much their users trust the systems.

To measure trust, researchers at the Institute for Defense Analyses (IDA) created the Trust of Automated Systems Test (TOAST), a validated scale designed to measure trust by analyzing system performance and understanding. The TOAST scale already has been initially validated for use within the defense community and in some civilian scenarios (Wojton et al., 2020). This nine-item scale contains both understanding and performance components, with four and five questions, respectively. To further explore the results from Wojton et al.'s first TOAST research paper, we planned to test whether the TOAST scale predicts real reliance behaviors. Additionally, we planned to test how long it would take a person to trust a system to the extent that they will and to see whether trust increases with system use. This research would provide insight into how well users will accept new autonomous systems based on their TOAST scores and how appropriate a team consisting of a user and an automated system is for completing a given task.

Combining ideas from Wojton et al.'s, Nourani et al.'s, and Yu et al.'s research, we tested three main concepts: (1) how much a person's trust in the system changes as they use it, (2) how much time it takes for a person to reach their steady state of reliance, and (3) whether the TOAST scale matches up with reliance when users are required to make real choices rather than self-report their expected choices. We anticipated that as participants used the system more, they would trust high performing systems more and trust low performing systems less. Also, we expected that participants who used the low performing system would reach their reliance steady state faster than those who used the high performance system. Finally, we believed that a user's level of steady state reliance would be correlated with a TOAST score collected at the end of the experiment.

2. Experimental Method

A. Participants

Participants were recruited through Amazon's Mechanical Turk, an online recruitment service. Of the 419 participants who attempted the experiment, 120 chose to exit the experiment early and 5 were removed for failing to meet the experiment's requirements. Of the remaining 294 participants, 74 were in the High Understanding + High Performance condition, 74 were in the Low Understanding + High Performance condition, 73 were in the High Understanding + Low Performance condition, and 73 were in the Low Understanding + Low Performance condition.

Participants were paid $1.00 for attempting the experiment. They could earn a $0.50 bonus for the single trials if they obtained at least 50 of 60 possible points in the high performance condition or at least 45 of 60 possible points in the low performance condition. They could also earn a $0.50 bonus for the block trials if their total penalty was ≤ 10 points in the high performance condition or ≤ 20 points in the low performance condition. Participants could earn both bonuses with qualifying scores, for a total bonus of $1.00.

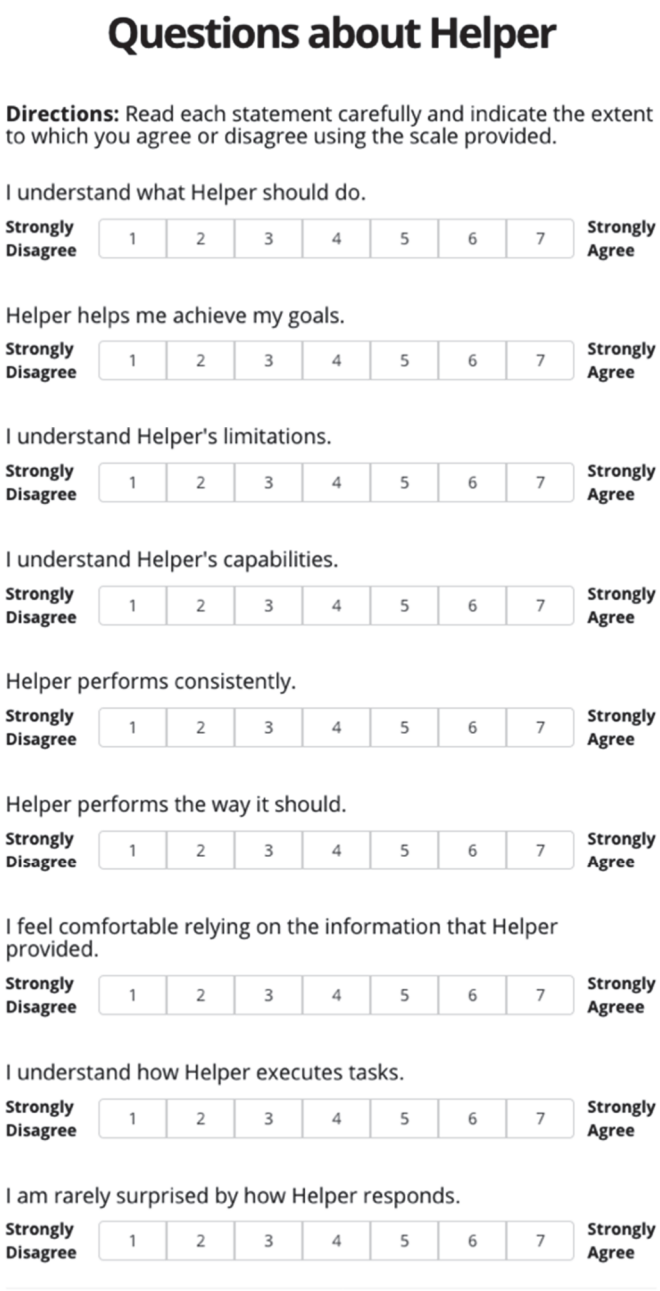
In total, 174 participants received no bonus, 84 received a $0.50 bonus for the single trials only, 17 received a $0.50 bonus for the block trials only, and 19 received $1.00 for both sets of trials.

B. Materials

We used Gorilla.sc, an online experiment creation tool, to collect data from participants.

In total, there were 120 stimuli, with 60 shown in the single trials and 60 shown in the block trials. The stimuli were images of full bullet cartridges and empty bullet casings.

Additionally participants were administered the TOAST scale, shown in Figure 2-1 (Wojton, 2020). This nine-question scale has two components—understanding and performance—with four and five questions, respectively, that are shown in a random order.



Questions about Helper

Directions: Read each statement carefully and indicate the extent to which you agree or disagree using the scale provided.

I understand what Helper should do.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

Helper helps me achieve my goals.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

I understand Helper's limitations.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

I understand Helper's capabilities.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

Helper performs consistently.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

Helper performs the way it should.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

I feel comfortable relying on the information that Helper provided.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agreee

I understand how Helper executes tasks.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

I am rarely surprised by how Helper responds.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

Figure 2-1. The TOAST Scale

C. Procedure

According to our experimental story, participants are performing a critical step of sorting objects in a factory. Their task is to make sure bullets get sent to soldiers to use in combat and bullet casings get sent to a warehouse so the warehouse employees can refill the casings and create more bullets. Sending bullet casings to the soldiers is problematic because the soldiers are in combat and cannot refill the casings themselves, making casings useless to them. Sending bullets to the warehouse is equally problematic because the

warehouse cannot do anything with the bullets besides send them back to the factory, and the soldiers did not receive the bullets they need. During the experiment, participants could choose to sort the stimuli manually or let an autonomous system, named Helper, sort the objects for them.

We randomized participants into high and low understanding conditions via the text they saw as an introduction to the experiment. Those in the high understanding condition were given text that explained how their version of Helper (high or low performance) would work. Those in the low understanding condition were given a placebo text of similar length that did not explain how Helper works.

We also randomized participants into high and low performance conditions according to which version of Helper they saw throughout all trials of the experiment. Those in the high performance condition saw a version of Helper with an accuracy rate of 90 percent and those in the low performance condition saw a version with an accuracy rate of 60 percent.

The experiment had three phases of trials: Training; Single Trials – Time to Steady State; and Block Trials – Measuring the Steady State. We administered the TOAST scale before training, after the single trials, and after the block trials.

1. Training

Participants used a tutorial to show them how to use Helper during the single trials, then were asked to test their knowledge using a fake version of Helper that sorts cupcakes and donuts. After they became familiar with the fake Helper, they were shown how to differentiate the real stimuli of bullets and bullet casings. They then had the opportunity to sort bullets and bullet casings on their own to ensure they knew the difference between the two.

2. Single Trials – Time to Steady State

Following a Go/No-Go design, participants first viewed a box, then text would appear in front of the box telling the participants where Helper has decided to send the box depending on Helper's accuracy rate and the stimulus in the trial. When this message was shown, the participant had the option to press nothing, allowing Helper to send the unknown object to the location it has indicated, or to press the space bar, allowing the participant to manually decide where the object should go.

An accurate response by either Helper or the participant would earn the participant one point and an inaccurate response would earn the participant zero points. The participant's points earned during the trial would display once Helper or the participant had decided where the box should go. To incentivize participants to use Helper, a one-second lag would occur between when the participant pressed the space bar to complete the task

manually and when they saw the stimulus for that trial. At the end of the single trials, the participant's final score was shown, along with the amount of times they chose to use Helper and the number of times they completed the task manually.

To ensure that participants paid attention to the experiment instead of letting it run, selecting Helper each time, a biohazard box was shown after a variable amount of trials. To clear the biohazard box and return to the regular sorting task, participants needed to press enter on their keyboard.

The single trial phase had a five-minute time limit so the participant might not have seen all 60 possible trials.

3. Block Trials – Measuring the Steady State

The participant received instructions that the game has changed now that they have had the chance to evaluate Helper's performance. They will be paired with the same Helper from the single trials. There were four trials per block and 15 possible blocks, but these trials also had a five-minute time limit.

Before each block, the participant decided whether Helper should complete all four trials in the block or whether they would manually complete the four trials. Each block had a different penalty per trial, ranging from 1 to 100 {1, 1, 2, 2, 5, 6, 8, 10, 12, 15, 20, 20, 24, 50, and 100}.

If the participant chose to let Helper complete the task, each decision took three seconds. The amount of points the participant received was displayed at the end of each block, before the participant had to decide whether Helper should complete the next block. At the end of the block trials, the participant saw their total score and the amount of times they chose to let Helper sort the groups of stimuli.

The block trials also had a biohazard box appear after a variable amount of trials to ensure that participants are paying attention to the experiment.

3. Experimental Results

A. Trust Changes with System Use

1. Single Trials

We performed three independent samples t-tests to analyze the difference in trust from before the participants had used the low performance Helper (TOAST 1) to after the participants had used the low performance Helper for the single trials (TOAST 2). Results indicate that the difference between TOAST 1 and TOAST 2 is significantly lower than 0 for participants who used the low performance Helper when considering the understanding component, $t(146) = -2.288$, $p = .024$; performance component, $t(146) = -5.020$, $p < .001$; and the complete TOAST score, $t(146) = -4.735$, $p < .001$ (Table 3-1).

Table 3-1. Difference between TOASTs 1&2 from 0 in t-tests – Single Trials

Understanding					
	T-statistic	df	P-value	Confidence Interval	Mean
High Performance	1.3348	146	0.184	[-0.03760275, 0.19406533]	0.07823129
Low Performance	-2.2884	146	0.02355	[-0.30109529, -0.02203397]	-0.1615646
Performance					
	T-statistic	df	P-value	Confidence Interval	Mean
High Performance	0.31631	146	0.7522	[-0.1071029, 0.1479192]	0.02040816
Low Performance	-5.0201	146	1.483e-06	[-0.8039768, -0.3497647]	-0.5768707
All					
	T-statistic	df	P-value	Confidence Interval	Mean
High Performance	0.89191	146	0.3739	[-0.05605958, 0.14827425]	0.04610733
Low Performance	-4.735	146	5.135e-06	[-0.5560296, -0.2285509]	-0.3922902

We also performed three independent samples t-tests to analyze the difference in trust from TOAST 1 to TOAST 2 for participants using the high performance Helper. The difference in trust for participants who used the high performance Helper is not significantly different from 0 for the understanding component, $t(146) = 1.335$, $p = .184$; performance component, $t(146) = 0.316$, $p = .752$; or the complete TOAST score, $t(146) = 0.892$, $p = .374$ (Table 3-1).

2. Block Trials

We performed three independent samples t-tests to analyze the difference in trust from after participants had used Helper for the single trials (TOAST 2) to after the participants had used their assigned version of Helper for the block trials (TOAST 3). Results indicate that the difference between TOAST 2 and TOAST 3 is significantly lower than 0 for participants who used the low performance Helper when considering the understanding component, t (146) = −2.812, p = .006; performance component, t (146) = −4.343, p < .001; and the complete TOAST score, t (146) = −4.368, p < .001 (Table 3-2).

Table 3-2. Difference between TOASTs 2&3 from 0 in t-tests – Block Trials

Understanding					
	T-statistic	df	P-value	Confidence Interval	Mean
High Performance	-3.1766	146	0.001819	-0.38071128 -0.08867647	-0.2346939
Low Performance	-2.8115	146	0.005609	-0.29540847 -0.05153031	-0.1734694
Performance					
	T-statistic	df	P-value	Confidence Interval	Mean
High Performance	-4.3698	146	2.347e-05	-0.4939710 -0.1863011	-0.3401361
Low Performance	-4.3433	146	2.612e-05	-0.4196822 -0.1571885	-0.2884354
All					
	T-statistic	df	P-value	Confidence Interval	Mean
High Performance	-4.25	146	3.794e-05	-0.4296511 -0.1568946	-0.2932729
Low Performance	-4.3677	146	2.367e-05	-0.3447341 -0.1299447	-0.2373394

We also performed three independent samples t-tests to analyze the difference in trust from TOAST 2 to TOAST 3 for participants using the high performance Helper. Results indicate that the difference between TOAST 2 and TOAST 3 is significantly lower than 0 for participants who used the high performance Helper when considering the understanding component, t (146) = −3.177, p = .002; performance component, t (146) = −4.370, p < .001; and the complete TOAST score, t (146) = −4.250, p < .001 (Table 3-2).

B. Time to Steady State

We analyzed data from each participant to determine on which number of the single trials they reached the point where reliance is predicted to remain constant through infinite amounts of further interactions with their version of Helper.

On average, participants reached their steady state of trust on trial 15.4 out of an average of 49 completed trials, indicating they spent 70 percent of the single trials that they completed in their steady state of trust. Of the 294 participants, 43 percent (125/294) maintained their steady state through all trials and 12 percent (36/294) never reached a steady state.

A one-tailed t-test shows a non-significant trend that participants who used the low performance Helper reached their reliance steady state more slowly than those who used the high performance Helper, t (292) = 1.491, p = .069. Participants who used the high performance Helper spent 72 percent of the single trials they completed in their steady state of trust and participants who used the low performance Helper spent 66 percent of the single trials they completed in their steady state of trust.

C. TOAST and Reliance

A logarithmic trend line accounts for 78.3 percent of the variance in the average percentage of participants who chose to manually complete the block per penalty (Figure 3-1). The TOAST 3 understanding component, r (292) = 0.034, p = .566; performance component, r (292) = −0.066, p = .267; and the complete TOAST score, r (292) = −0.032, p = .585, had no correlation with the logarithmic slope (Table 3-3).

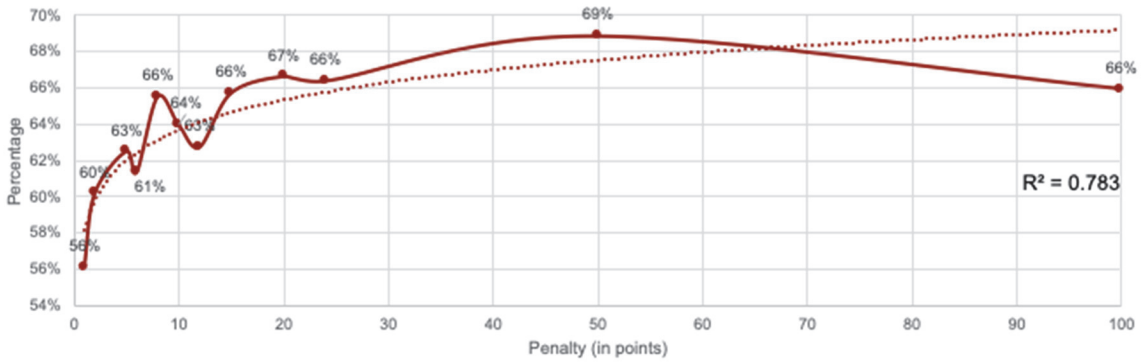


Figure 3-1. Average Percentage of Participants Who Chose Manual per Penalty

Table 3-3. Correlation of TOAST 3 Scores with Logarithmic Slope

Scale Component	Correlation with ln Slope
TOAST 3 – Understanding	0.033597354
TOAST 3 – Performance	-0.065729388
TOAST 3 – All	-0.032110191

Scale Component	Correlation with ln Slope
TOAST Average – Understanding	0.11197389
TOAST Average – Performance	-0.01434953
TOAST Average – All	0.039320146

4. Discussion

We expected that as participants used the system more, they would trust the high performing systems more and the low performing systems less. In accordance with our hypothesis, participants who used the low performance Helper trusted the system less than expected when using Helper on a case-by-case basis; however, those who used the high performance Helper trusted the system the same as expected rather than more than expected. Additionally, participants who used the low performance version of Helper lost a significant amount of trust after using Helper on a group-case basis, but those who used the high performance version of Helper lost a statistically equivalent amount of trust. Both results indicate that having a high performance system is important for trust but only when the user can decide for each case whether to trust or distrust the system.

Contrary to what Madhavan and Wiegmann (2007) found in their research, we observed a trend that participants who used the low performing system reached their reliance state more slowly than those who used the high performing system. If further analysis finds significant evidence that this is true, such evidence would show that it takes longer to break trust than to maintain trust. This discrepancy between research findings might be because the consequence for a participant of getting one trial wrong is one point. In the grand scale of risk, one point is not much and may not be enough of a penalty to cause a rapid drop in trust when using either system. Therefore, those who used the low performance system may have taken more time to decide that they would trust Helper less.

A logarithmic trend line accounts for 78.3 percent of the variance in the average percentage of participants who chose to manually complete the block of trials per penalty. Even though the trend line fit the data well, the log of participants' level of steady states of reliance were not correlated with their TOAST 3 scores. It is possible that there is no evidence supporting the hypothesis because participants might have heavily based their decisions on which trials they saw first. Because the block trials were randomized, participants saw trials with given penalty points in a different order. For example, one participant might have seen a trial with 50 penalty points first and then 1 penalty point, and another could have seen a trial with 1 penalty point first and then 50 penalty points. In this case, even if the participants both decide to stop trusting Helper on trial 2, the amount of points that those participants risked would be vastly different, but evidence of this difference may not appear in their TOAST 3 scores.

Further research could include a second round of the single trials without the penalty system after participants have completed the block trials. Adding this section would allow

us to determine whether the loss in trust seen after the block trials is attributable to the fact that participants cannot change their mind after each trial or to their changed perception based on observing Helper's performance in the block trials.

Additionally, future research could include a single trial section with a variable penalty to distinguish the effects of the variable penalties from the fact that those penalties apply to block of trials. In this new section, participants would see the amount of penalty points they would lose for a wrong answer before deciding whether Helper should sort that that one stimulus. As in the original experiment, participants would have the opportunity to earn a real-world performance bonus based on the amount of penalty points they accumulate. Data from this section of the experiment could be compared with TOAST data to determine whether participants' steady state from single trials with penalties is correlated with their reported trust in the system.

As observed in the experiment, understanding a well-performing autonomous system is the key to maintaining trust on a case-by-case basis. Maintaining trust that is well calibrated to the system is a key aspect of creating highly effective human-machine teams. Widespread use of the TOAST scale would allow researchers to (1) better predict how users will accept new autonomous systems and (2) determine whether specific human-machine teams are appropriate for the tasks they are designed to complete.

Whether the system in question is Apple's Siri, Tesla's Autopilot, or one of the many autonomous systems currently used by the general population, it is important to understand how compatible these systems are with human users, both physically and psychologically.

Appendix A.
Additional Data

Table A-1. High vs. Low Performance t-tests for TOASTs 1–3

Understanding						
	T-statistic	df	P-value	Confidence Interval	Mean (High Performance)	Mean (Low Performance)
TOAST 1	-1.0323	292	0.3028	[-0.30153795, 0.09405495]	5.562925	5.666667
TOAST 2	1.2386	292	0.2165	[-0.08012691, 0.35223575]	5.641156	5.505102
TOAST 3	0.61922	202	0.5363	[-0.1630081, 0.3126680]	5.406463	5.331633
Performance						
	T-statistic	df	P-value	Confidence Interval	Mean (High Performance)	Mean (Low Performance)
TOAST 1	-0.50104	292	0.6167	[-0.2346697, 0.1394316]	5.487075	5.534694
TOAST 2	3.8138	292	0.000167	[0.2660071, 0.8333126]	5.507483	4.957823
TOAST 3	3.0899	202	0.002195	[0.1807788, 0.8151396]	5.167347	4.669388
All						
	T-statistic	df	P-value	Confidence Interval	Mean (High Performance)	Mean (Low Performance)
TOAST 1	-0.80552	292	0.4212	[-0.2498528, 0.1047281]	5.520786	5.593348
TOAST 2	3.1974	292	0.001539	[0.1406486, 0.5910219]	5.566893	5.201058
TOAST 3	2.4072	202	0.0167	[0.05652453, 0.56327895]	5.273621	4.963719

Table A-2. High vs. Low Performance t-tests for the Differences between TOASTs 1&2 and 2&3

Understanding						
	T-statistic	df	P-value	Confidence Interval	Mean (High Performance)	Mean (Low Performance)
TOAST 1&2	2.6133	292	0.009431	[0.05920468, 0.42038716]	0.07823129	-0.16156463
TOAST 2&3	-0.63605	292	0.5252	[-0.2506701, 0.1282211]	-0.2346939	-0.1734694

Performance						
	T-statistic	df	P-value	Confidence Interval	Mean (High Performance)	Mean (Low Performance)
TOAST 1&2	4.5322	292	8.523e-06	0.3379088 0.8566490	0.02040816	-0.57687075
TOAST 2&3	-0.5053	292	0.6137	-0.2530742 0.1496729	-0.3401361	-0.2884354

All						
	T-statistic	df	P-value	Confidence Interval	Mean (High Performance)	Mean (Low Performance)
TOAST 1&2	4.4893	292	1.03e-05	0.2462016 0.6305935	0.04610733	-0.39229025
TOAST 2&3	-0.63682	292	0.5247	-0.2287987 0.1169317	-0.2932729	-0.2373394

Appendix B.
Predicting Trust in Automated Systems:
Application of the Trust of Automated Systems Test (TOAST)

Caitlan A. Fealing

June 29, 2022

Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

Understanding the trust between a human and a well-performing autonomous system is the key to creating the most effective human-machine teams

Outline

- Background Information
- The Experiment
- Data
- Analysis
- Results
- Next Steps

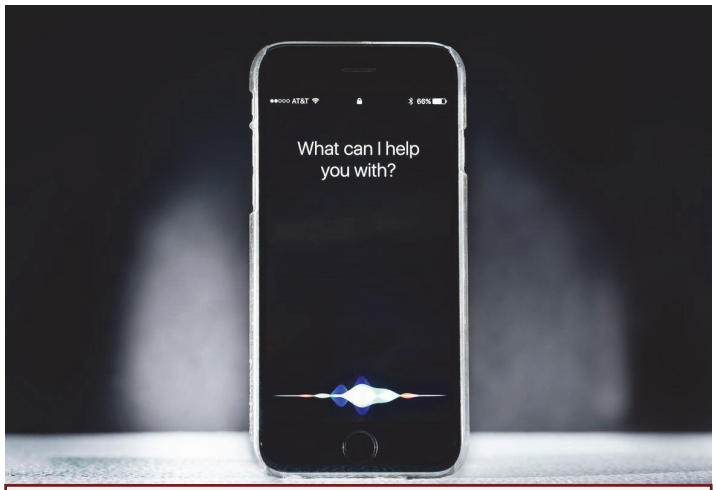
Trust is key to determining the circumstances under which people will rely on automated systems

It is important to have a standard trust of autonomy scale so systems can be compared to one another

Trust of autonomous systems is a psychological state in which a person makes themselves vulnerable because they are confident that the automated system is capable and reliable enough to complete the task.[1]

[1] Nave et al., 2015; Button, 2017; Hoff & Bashir, 2015

Trust depends on context and cannot be adequately measured by observing reliance



Apple's Siri

Tesla's Autopilot

Less risk, low penalty

More risk, high penalty

- Sometimes the outcome is less risky, indicating less penalty for failure, but other times, failure can be catastrophic

- "Trust guides reliance when complexity and unanticipated situations make a complete understanding of the automation impractical."[2]

[2] Lee & See, 2004

The TOAST scale enables researchers to estimate levels of users' trust

TOAST decomposes trust into understanding and performance

TOAST Questions	
Understanding	**Performance**
• I understand what the system should do. • I understand the system's limitations. • I understand the system's capabilities. • I understand how the system executes tasks.	• The system helps me achieve my goals. • The system performs consistently. • The system performs the way it should. • I feel comfortable relying on the information that the system provided. • I am rarely surprised by how the system responds.
Wojton et al., 2020	

Questions and Hypotheses

1. How much does a person's trust in the system change as they use it?
   - H_1: As participants use the system more, they trust high performing systems more and trust low performing systems less.
2. How much time does it take for a person to rely on an automated system to the extent that they will?
   - H_2: In accordance with Madhavan and Wiegmann's findings, participants who use the low performing system will reach their reliance steady state faster than those using the high performance system.
3. Does the TOAST scale correlate with reliance when participants have to make real choices/not self-report?
   - H_3: The participant's level of steady state reliance will be correlated with their TOAST score at the end of the experiment.

The Experiment

- Scenario: Participants simulated working at a sorting facility. They could either choose to let Helper sort the stimuli to the appropriate location or they could do it manually.
- Helper is a simulated AI system that participants could use to sort stimuli during their tasks.
- 2 × 2 Design = 4 Conditions

	High Understanding • Participant is given instructions on how Helper works	**Low Understanding** • Participant is given a brief text that does not explain Helper
High Performance • Helper has an accuracy rate of 90%		
Low Performance • Helper has an accuracy rate of 60%		

Data from 294 participants were analyzed

	High Understanding	Low Understanding
High Performance	74	74
Low Performance	73	73

Score Category	Payment	Participants
No Bonus	$1.00	174
Single Trial Bonus	$1.50	84
Block Trial Bonus	$1.50	17
Both Bonuses	$2.00	19

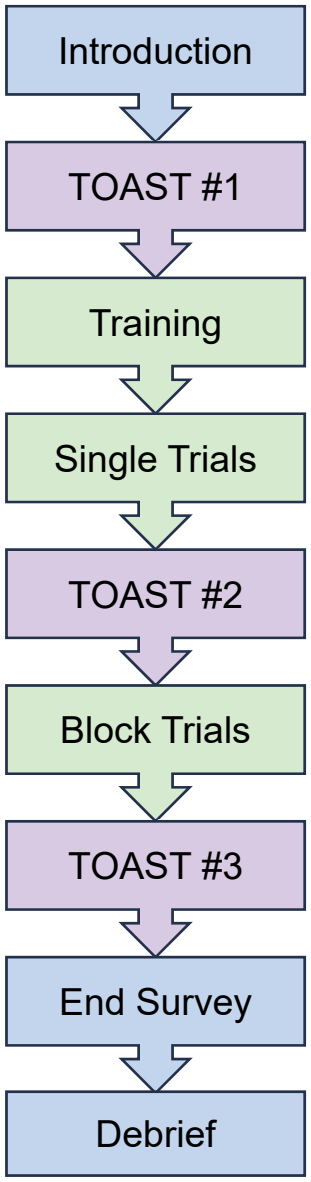
419 participants attempted the experiment

- Participants were recruited through Amazon's Mechanical Turk (MTurk)

Attrition

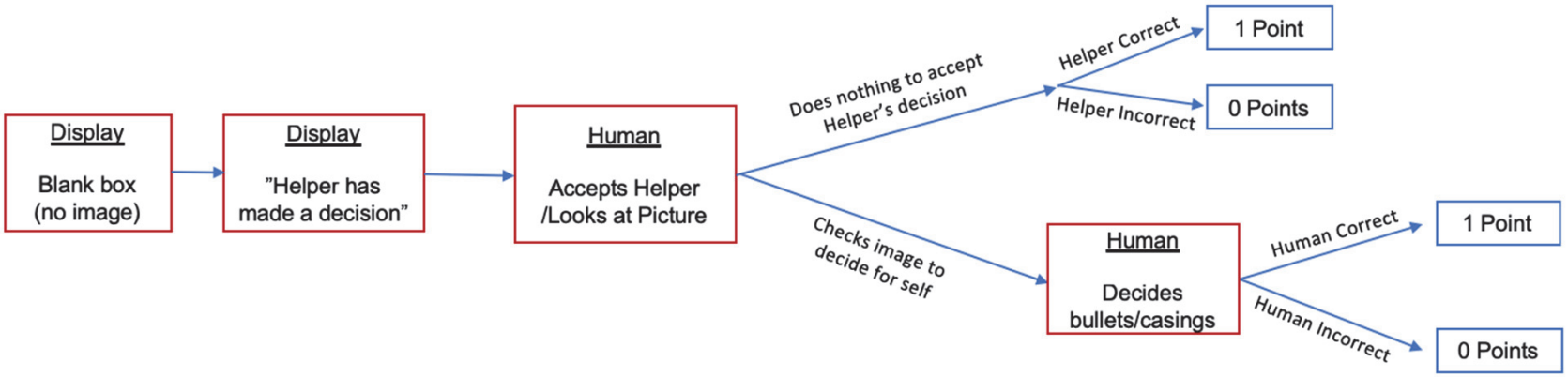
- 120 participants chose to exit the experiment early
  - Most attrition (63%) occurred in the High Understanding condition
- 5 participants were removed for failing to meet experiment requirements

The Experiment – Outline



- The introduction text randomizes participants into High and Low Understanding conditions
- Participants are randomized into High and Low Performance conditions based on which version of Helper they see in the single and block trials
- Participants are given the TOAST 3 times
- 3 trials: Training, Single, and Block
- General AI usage data and participant demographics are collected via the end survey
- Participants are told the purpose of the study in the debrief

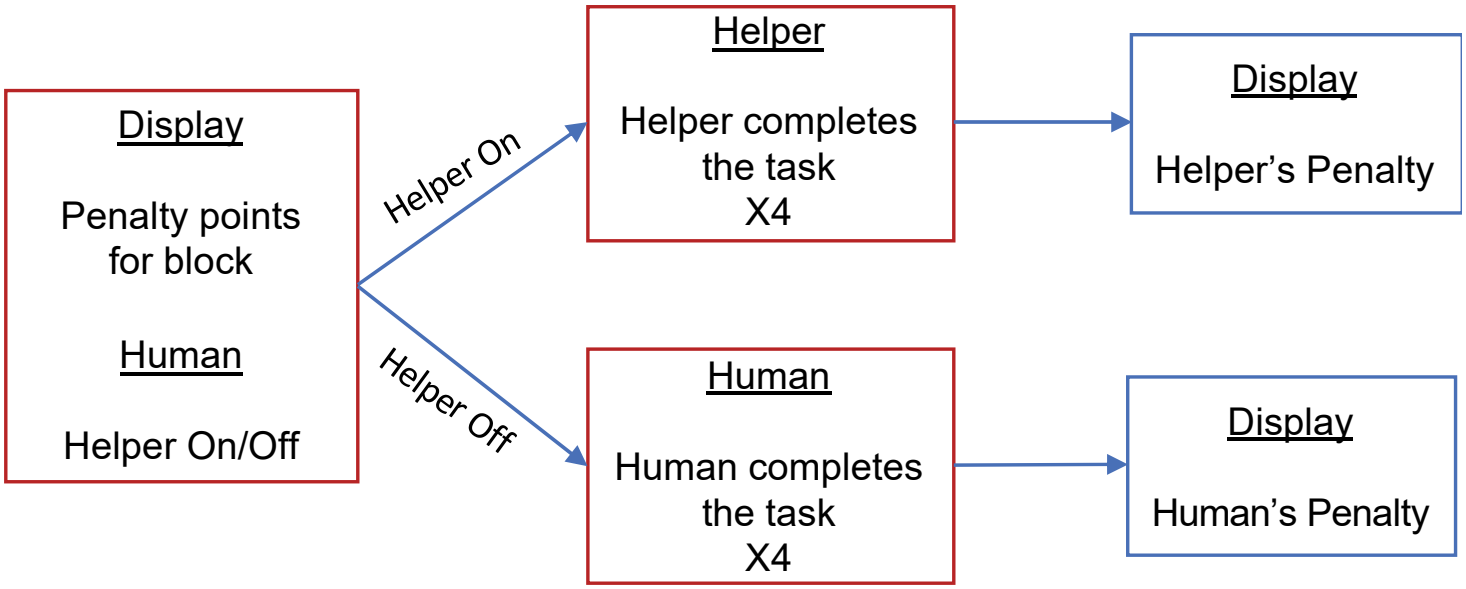
The Experiment – Single Trials



Goal: Measure time to reliance state with Helper

60 Go/No-Go trials

The Experiment – Block Trials



Goal: Measure reliance on Helper when given the choice

4 trials per block, 15 blocks

This experiment had a high prevalence of digital assistant users

11	x	128	=	1,408
Batches of participants		nodes (parts) of the experiment		separate csv files of data

Data

of Observations: 294

- 1 per participant

Questionnaires:

- Introduction
- TOASTs #1-3
- End Survey
- Debrief

Tasks:

- Training
- Single Trials
- Block Trials

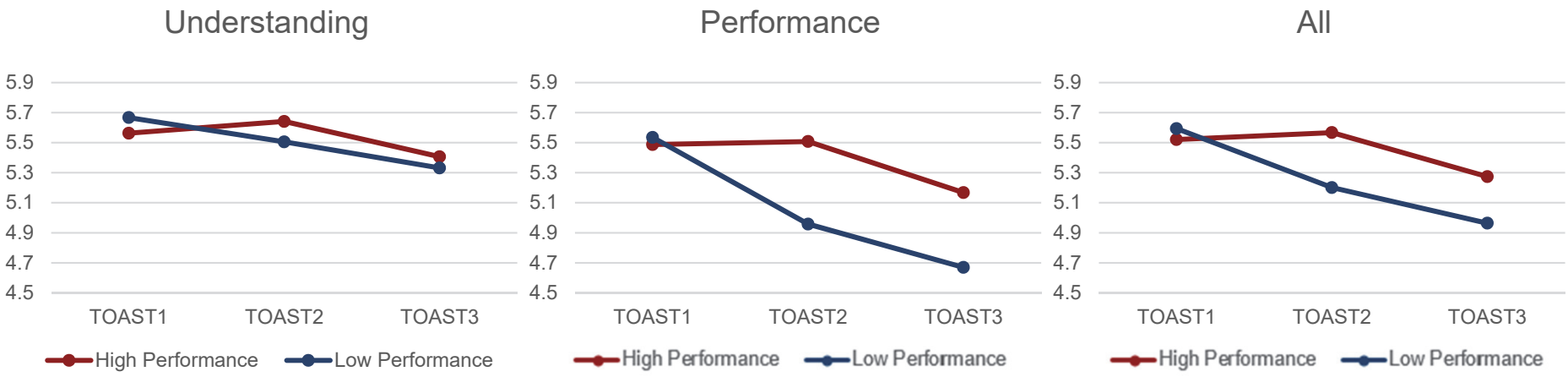
Summary Statistics

- 173/294 (59%) had used systems similar to Helper
- 197/294 (67%) use a digital assistant at least once per day
- Participants say they trust people until they have a reason not to more often than not
- Participants believe that their friends think they are tech savvy more often than not

Source: Experimental data were collected via Gorilla.sc

1. Trust declined more among operators using the low performance Helper on a case-by-case basis

How much does a person's trust in the system change as they use it?



TOAST scores range from 1 to 7, with higher scores indicating more trust.

Results

Single Trials: As participants use the system more, they trust the high performing system the same as they expected to and they trust the low performing systems less

- The difference in trust from TOAST 1 to TOAST 2 is significantly lower than 0 for participants who used the Low Performance Helper

Block Trials: Participants who used the high performing system lost the same amount of trust as those who used the low performing system

- The difference in trust from TOAST 2 to TOAST 3 is significantly lower than 0 for both groups of participants and not significantly different between the groups. That is, the slope of the blue line between TOAST 1 to TOAST 2 is statistically significantly less than 0 in all three graphs.

2. On average, participants reached their steady state of trust on trial 15.4 out of an average of 49 completed trials

How much time does it take for a person to rely on an automated system to the extent that they will?

Results

On average, participants spent 70% of the single trials they completed in their steady state of trust

- 43% (125/294) participants maintained their steady state through all single trials
- 12% (36/294) participants never reached a steady state

Trend – Participants who use the low performing system reached their reliance steady state slower than those using the high performing system

- Participants using the high performing system spent 72% of the single trials they completed in their steady state of trust
- Participants using the low performing system spent 66% of the single trials they completed in their steady state of trust
- One-tailed t-test = 0.0686 → More testing needed

Note: Participants using the low performing system chose to complete the task manually more often than those using the high performing system

3. The participants' level of steady states of reliance did not correlate with their TOAST 3 scores

Does the TOAST scale match up with reliance when participants have to make real choices/not self-report?

Results

- A logarithmic trend line (shown by the red dashed line) accounts for 78.3% of the variance in the average percentage of participants who chose to manually complete the block per penalty
- The Understanding component (r = 0.03), Performance component (r = –0.07), and complete TOAST 3 score (r = – 0.03) had no correlation with ln slope (the slope coefficient in the logarithmic trendline equation)

Results

1. How much does a person's trust in the system change as they use it?
   - Single Trials: As participants use the system more, they trust the high performing system the same as they were expected to and they trust the low performing systems less
   - Block Trials: Participants who used the high performing system lost the same amount of trust as those who used the low performing system

2. How much time does it take for a person to rely on an automated system to the extent that they will?
   - On average, participants reached their steady state of trust on trial 15.4
   - On average, participants spent 70% of the single trials they completed in their steady state of trust
   - Trend – Participants who use the low performing system reached their reliance steady state more slowly than those using the high performing system

3. Does the TOAST scale correlate with reliance when participants have to make real choices/not self-report?
   - The participants' level of steady states of reliance were not correlated with their TOAST 3 scores

Next Steps

- Steady state – Find a more rigorous way to calculate when participants reach their steady state of reliance
  - Chaos theory
  - Ensure reaching a steady state of using Helper does not indicate that the participant has become bored with the experiment

- Correlation of steady state reliance with TOAST scores – Further test how trust changes based on penalty per trial to evaluate the TOAST scale

Understanding a well-performing autonomous system is the key to the most effective human-machine teams

Key benefits of studying trust of autonomous systems are:

1. Better predictions about when people will not accept new autonomous systems
2. An emphasis on creating more human-compatible systems (e.g., for combat scenarios)

References

- Button, R. W. (September 7, 2017). *Artificial Intelligence and the Military*. RAND Corporation. https://www.rand.org/blog/2017/09/artificial-intelligence-and-the-military.html.
- Devitt, S. K. (2018). Trustworthiness of autonomous systems. In H. A. Abbass, J. Scholz. & Darryn Reid (eds.), *Foundations of Trusted Autonomy* (pp. 161-184). Springer Cham. https://doi.org/10.1007/978-3-319-64816-3.
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407-434. https://doi.org/10.1177/0018720814547570.
- Jessup, S. A. (2018). "Measurement of the Propensity to Trust Technology." MS Thesis, Wright University.
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153-184. https://doi.org/10.1006/ijhc.1994.1007.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392.
- Lees, M. N., & Lee, J. D. (2007). The influence of distraction and driving context on driver response to imperfect collision warning systems. *Ergonomics*, *50*(8), 1264-1286. https://doi.org/10.1080/00140130701318749.
- Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277-301. https://doi.org/10.1080/14639220500337708.
- Nave, G., Camerer, C., & McCullough, M. (2015). Does oxytocin increase trust in humans? A critical review of research. *Perspectives on Psychological Science*, *10*(6), 772-789. https://doi.org/10.1177/1745691615600138.
- Nourani, M., Kabir, S., Mohseni, S., & Ragan, E. D. (2019, October). The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 7, pp. 97-105). https://ojs.aaai.org/index.php/HCOMP/article/view/5284.
- Soffar, H. (January 19, 2021). "Military Artificial Intelligence (Military Robots) Advantages, Disadvantages & Applications." *Science Online.* https://www.online-sciences.com/robotics/military-artificial-intelligence-military-robots-advantages-disadvantages-applications/.
- Wojton, H. M., Porter, D., Lane, S. T., Bieber, C., & Madhavan, P. (2020). Initial validation of the Trust of Automated Systems Test (TOAST). *The Journal of Social Psychology*, *160*(6), 735-750. https://doi.org/10.1080/00224545.2020.1749020.
- Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019, March). Do I trust my machine teammate? An investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 460-468). https://doi.org/10.1145/3301275.3302277.

Backup

TOAST Scale

Questions about Helper

Directions: Read each statement carefully and indicate the extent to which you agree or disagree using the scale provided.

I understand what Helper should do.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

Helper helps me achieve my goals.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

I understand Helper's limitations.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

I understand Helper's capabilities.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

Helper performs consistently.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

Helper performs the way it should.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

I feel comfortable relying on the information that Helper provided.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agreee

I understand how Helper executes tasks.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

I am rarely surprised by how Helper responds.

Strongly Disagree	1	2	3	4	5	6	7	Strongly Agree

Participants – Bonuses

- Participants earned a $0.50 bonus for the single trials if they obtained at least 50/60 possible points in the High Performance condition and at least 45/60 possible points in the Low Performance condition.
- Participants earned a $0.50 bonus for the block trials if their total penalty was ≤ 10 points in the High Performance condition and ≤ 20 points in the Low Performance condition.
- Participants could earn both bonuses with qualifying scores for a $1.00 total bonus.

Summary Statistics

Participants

- 282 Participants
- 166/282 (59%) had used systems similar to Helper
- 188/282 (67%) use a digital assistant at least once per day
- Participants say they trust people until they have a reason not to more often than not
- Participants believe that their friends think they are tech savvy more often than not

Single Trials

- Average Percent Correct for Single Trials: 78%
- Average Percent of Trials That Used Helper: 61%
- Average Number of Trials Where Participant Chose Manual: 27
- Average Number of Trials Total: 49

Block Trials

- Average Percent Correct for Block Trials: 75%
- Average Percent of Trials That Used Helper: 42%
- Average Number of Trials Where Participant Chose Manual: 8
- Average Number of Trials Total: 12
- Average Total Penalty: 172

Question 1 – Statistics – High vs. Low Performance

Understanding						
	T-statistic	df	P-value	Confidence Interval	Mean (High Performance)	Mean (Low Performance)
TOAST 1	−1.0323	292	0.3028	[−0.30153795, 0.09405495]	5.562925	5.666667
TOAST 2	1.2386	292	0.2165	[−0.08012691, 0.35223575]	5.641156	5.505102
TOAST 3	0.61922	202	0.5363	[−0.1630081, 0.3126680]	5.406463	5.331633

Performance						
	T-statistic	df	P-value	Confidence Interval	Mean (High Performance)	Mean (Low Performance)
TOAST 1	−0.50104	292	0.6167	[−0.2346697, 0.1394316]	5.487075	5.534694
TOAST 2	3.8138	292	0.000167	[0.2660071, 0.8333126]	5.507483	4.957823
TOAST 3	3.0899	202	0.002195	[0.1807788, 0.8151396]	5.167347	4.669388

All						
	T-statistic	df	P-value	Confidence Interval	Mean (High Performance)	Mean (Low Performance)
TOAST 1	−0.80552	292	0.4212	[−0.2498528, 0.1047281]	5.520786	5.593348
TOAST 2	3.1974	292	0.001539	[0.1406486, 0.5910219]	5.566893	5.201058
TOAST 3	2.4072	202	0.0167	[0.05652453, 0.56327895]	5.273621	4.963719

Orange rows indicate significant findings.

Question 1 – Statistics – High vs. Low Performance

Understanding						
	T-statistic	df	P-value	Confidence Interval	Mean (High Performance)	Mean (Low Performance)
TOAST 1&2	2.6133	292	0.009431	[0.05920468, 0.42038716]	0.07823129	−0.16156463
TOAST 2&3	−0.63605	292	0.5252	[−0.2506701, 0.1282211]	−0.2346939	−0.1734694

Performance						
	T-statistic	df	P-value	Confidence Interval	Mean (High Performance)	Mean (Low Performance)
TOAST 1&2	4.5322	292	8.523e-06	[0.3379088, 0.8566490]	0.02040816	−0.57687075
TOAST 2&3	−0.5053	292	0.6137	[−0.2530742, 0.1496729]	−0.3401361	−0.2884354

All						
	T-statistic	df	P-value	Confidence Interval	Mean (High Performance)	Mean (Low Performance)
TOAST 1&2	4.4893	292	1.03e-05	[0.2462016, 0.6305935]	0.04610733	−0.39229025
TOAST 2&3	−0.63682	292	0.5247	[-0.2287987, 0.1169317]	−0.2932729	−0.2373394

Orange rows indicate significant findings.

Question 1 – Statistics – High/Low Performance in Single Trials: Significant Difference of Values from 0

Understanding					
	T-statistic	df	P-value	Confidence Interval	Mean
High Performance	1.3348	146	0.184	[−0.03760275, 0.19406533]	0.07823129
Low Performance	−2.2884	146	0.02355	[−0.30109529, −0.02203397]	−0.1615646
Performance					
	T-statistic	df	P-value	Confidence Interval	Mean
High Performance	0.31631	146	0.7522	[−0.1071029, 0.1479192]	0.02040816
Low Performance	−5.0201	146	1.483e-06	[−0.8039768, −0.3497647]	−0.5768707
All					
	T-statistic	df	P-value	Confidence Interval	Mean
High Performance	0.89191	146	0.3739	[−0.05605958, 0.14827425]	0.04610733
Low Performance	−4.735	146	5.135e-06	[−0.5560296, −0.2285509]	−0.3922902

Orange rows indicate significant findings.

Question 1 – Statistics – High/Low Performance in Block Trials: Significant Difference of Values from 0

Understanding					
	T-statistic	df	P-value	Confidence Interval	Mean
High Performance	−3.1766	146	0.001819	[−0.38071128, −0.08867647]	−0.2346939
Low Performance	−2.8115	146	0.005609	[−0.29540847, −0.05153031]	−0.1734694
Performance					
	T-statistic	df	P-value	Confidence Interval	Mean
High Performance	−4.3698	146	2.347e-05	[−0.4939710, −0.1863011]	−0.3401361
Low Performance	−4.3433	146	2.612e-05	[−0.4196822, −0.1571885]	−0.2884354
All					
	T-statistic	df	P-value	Confidence Interval	Mean
High Performance	−4.25	146	3.794e-05	[−0.4296511, −0.1568946]	−0.2932729
Low Performance	−4.3677	146	2.367e-05	[−0.3447341, −0.1299447]	−0.2373394

Orange rows indicate significant findings.

Question 2 – Data

System Performance	Average Percent of Trials in Steady State	Variance	# Participants Who Started Their Steady State on the First Trial	# Participants Who Never Reached Steady State
High Performance	0.723316359	0.131209056	78	17
Low Performance	0.661288522	0.123340875	47	19
t-test (one-tailed)	0.068574019			

	Average Percent of Trials in Steady State	Variance	# Participants Who Started Their Steady State on the First Trial	# Participants Who Never Reached Steady State
All Participants	0.69230244	0.127805726	125	36

Question 3 – Data

TOAST Data – Scale Component	Correlation with ln Slope
TOAST 3 Average – Understanding	0.033597354
TOAST 3 Average – Performance	−0.065729388
TOAST 3 Average – All	−0.032110191
TOAST Average – Understanding	0.11197389
TOAST Average – Performance	−0.01434953
TOAST Average – All	0.039320146

References

Cerny, M. (2015, September). Sarah and Sally: Creating a likeable and competent ai sidekick for a videogame. In *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference*. https://ojs.aaai.org/index.php/AIIDE/article/view/12815.

Cervantes, E. (2021, December 25). The Best Siri Commands for Productivity, Information, Laughter, and More. *Android Authority*. https://www.androidauthority.com/best-siri-commands-1094484/.

Desai, M., Stubbs, K., Steinfeld, A., & Yanco, H. (2009, April). Creating trustworthy robots: Lessons and inspirations from automated systems. In *Proceedings of AISB '09 Convention: New Frontiers in Human-Robot Interaction* (pp. 1–8).

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570.

Jessup, S. A. (2018). The Measurement of the Propensity to Trust Technology. MS Thesis, Wright University.

Jian, J., Bisantz, A., Drury, C., & Llinas, J. (2009). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 53–71. https://doi.org/10.1207/S15327566IJCE0401_04.

Khojastehnazhand, M., Omid, M., & Tabatabaeefar, A. (2010). Development of a lemon sorting system based on color and size. *African Journal of Plant Science*, *4*(4), 122-127. https://doi.org/10.5897/AJPS.9000061.

Körber, M. (2018, August). Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association* (pp. 13–30). Springer Cham.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153−184. https://doi.org/10.1006/ijhc.1994.1007.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.

Lees, M. N., & Lee, J. D. (2007). The influence of distraction and driving context on driver response to imperfect collision warning systems. *Ergonomics*, *50*(8), 1264–1286. https://doi.org/10.1080/00140130701318749.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708.

Marikyan, D., Papagiannidis, S., & Alamanos, E. (2019). A systematic review of the smart home literature: A user perspective. *Technological Forecasting and Social Change*, *138*, 139–154. https://doi.org/10.1016/j.techfore.2018.08.015.

Merriam-Webster. (n.d.). Reliant. In *Merriam-Webster.com dictionary*. Retrieved July 8, 2022, from https://www.merriam-webster.com/dictionary/reliant.

Morando, A., Gershon, P., Mehler, B., & Reimer, B. (2020, September). Driver-initiated Tesla Autopilot disengagements in naturalistic driving. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 57–65). https://doi.org/10.1145/3409120.3410644.

Nave, G., Camerer, C., & McCullough, M. (2015). Does oxytocin increase trust in humans? A critical review of research. *Perspectives on Psychological Science*, *10*(6), 772–789. https://doi.org/10.1177/1745691615600138.

Nayak, M. (2022, July 8). Musk Admitted That Tesla Made a Mistake in Removing Speed Limiter Before Crash That Killed Teenager, Father Tells Court. *Fortune*. https://fortune.com/2022/07/08/musk-admitted-tesla-made-mistake-removing-speed-limiter-before-crash-that-killed-teenager-father-tells-court/.

Nourani, M., Kabir, S., Mohseni, S., & Ragan, E. D. (2019, October). The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 7, pp. 97–105). https://ojs.aaai.org/index.php/HCOMP/article/view/5284.

Rossi, F. (2018). Building trust in artificial intelligence. *Journal of International Affairs*, *72*(1), 127–134.

Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2018). The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. *Paladyn, Journal of Behavioral Robotics*, *9*(1), 137–154. https://doi.org/10.1515/pjbr-2018-0010.

Uzialko, A. (2022, June 29). Workplace Automation Is Everywhere, and It's Not Just About Robots. *Business News Daily*. https://www.businessnewsdaily.com/9835-automation-tech-workforce.html.

Wojton, H. M., Porter, D., T. Lane, S., Bieber, C., & Madhavan, P. (2020). Initial validation of the Trust of Automated Systems Test (TOAST). *The Journal of Social Psychology*, *160*(6), 735–750. https://doi.org/10.1080/00224545.2020.1749020.

Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019, March). Do I trust my machine teammate? An investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 460–468). https://doi.org/10.1145/3301275.3302277.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
25-07-2022	IDA Publication	

4. TITLE ANDSUBTITLE

Predicting Trust in Automated Systems – An Application of TOAST

5a. CONTRACT NUMBER
HQ0034-19-D-0001

5b. GRANT NUMBER
____ ____ ____

5c. PROGRAM ELEMENT NUMBER
____ ____ ____

6. AUTHOR(S)

Caitlan A. Fealing

5d. PROJECT NUMBER
ER-7-2351

5e. TASK NUMBER
C9089

5f. WORK UNIT NUMBER
____ ____ ____

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
730 East Glebe Road
Alexandria, Virginia 22305

8. PERFORMING ORGANIZATION REPORT NUMBER
NS-D-33188

H 2022-000321

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)
DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER
____ ____ ____

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release. Distribution Unlimited.

13. SUPPLEMENTARY NOTES
____ ____ ____

14. ABSTRACT
Following Wojton's research on the Trust of Automated Systems Test (TOAST), which is designed to measure how much a human trusts an automated system, we aimed to determine how well this scale performs when not used in a military context. We found that participants who used a poorly performing automated system trusted the system less than expected when using that system on a case by case basis, however, those who used a high performing system trusted the system the same as they expected. Additionally, both participants who used the poorly performing system and those who used the high performing system lost a significant amount of trust after using the system on a group case basis. These results indicate that having a high performance system is important for trust, but only when the user has the ability to decide to trust or distrust the system on a case-by-case basis.

15. SUBJECT TERMS
Automated system; Autonomous; reliance, trust, Trust in Automated Systems Test (TOAST)

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Daniel Porter
Unclassified	Unclassified	Unclassified	Unlimited	58	**19b. TELEPHONE NUMBER** *(include area code)* (703) 578-2869