Is this nuclear material secure? Examining trust in voice user interfaces for international nuclear safeguards seal examination

Kristin M. Divis
Sandia National Laboratories
kmdivis@sandia.gov

Jamie L. Coram
Sandia National Laboratories
jcoram@sandia.gov

Breannan C. Howell
Sandia National Laboratories
bchowel@sandia.gov

Zoe N. Gastelum
Sandia National Laboratories
zgastel@sandia.gov

Abstract

Voice user interfaces (VUIs) as decision aids are becoming increasingly popular in both everyday interactions (e.g., mobile assistants on personal phones) and high-risk, high-consequence security settings such as international nuclear safeguards. It is important that users have appropriately calibrated trust in these VUIs. Here, we bridge the domains of international nuclear safeguards, trust in technology, and VUI guidelines by examining human performance and trust in a VUI digital assistant for a safeguards seal examination task. This study serves as the foundation for future work investigating the impact of factors such as explainability, provenance, confidence, and granularity information on user trust in VUIs. This research will help establish best practice guidelines for VUIs within the context of international nuclear safeguards, which may also be applied to other national security VUI applications.

Keywords: Trust, voice user interface (VUI), digital assistant, cognitive science, artificial intelligence (AI)

1. Introduction

1.1. Nuclear Safeguards Domain

Safeguards inspectors from the International Atomic Energy Agency (IAEA) perform audits on commercial nuclear facilities around the world as part of treaty-based nuclear nonproliferation commitments. If inspectors find indications of nuclear material diversion or misuse, it could indicate the facilities are surreptitiously being used to develop nuclear weapons programs. Because of the broad political and economic ramifications of their findings, safeguards inspectors are faced with the challenging task of accurately detecting subtle signals of diversion or misuse while avoiding falsely accusing innocent countries. Nuclear material that is appropriately accounted for increases confidence between states that they do not require their own nuclear weapons program to counter potential illicit programs from their neighbors or adversaries. The international nuclear safeguards research community is invested in identifying new technology to aid inspectors in their tasks, such as voice user interface (VUI) digital assistants (e.g., Smartt, Gastelum, Rutkowski, Peter-Stein, & Shoman, 2021). Further research is needed to better understand *appropriate* levels of trust in VUIs prior to their adoption for specialized safeguards inspection tasks (e.g., seal examination or nuclear materials measurement). Findings within the safeguards domain may also be extended to other high-consequence inspection decision spaces, with implications for border security, physical protection, nuclear arms control, and export control, among others.

1.2. Trust in Technology

Over the past thirty years, the cognitive science community has produced a substantial body of research on trust in technologies such as automation and/or robotics (Hancock, et al., 2011; Hancock, Kessler, Kaplan, Brill, & Szalma, 2021; Hoff & Bashir, 2015; Kohn, de Visser, Wiese, Lee, & Shaw, 2021; Lee & Moray, 1994; Schaefer, Chen, Szalma, & Hancock, 2016; Sheridan, 2019). Mobile digital assistants such as Apple's Siri or Amazon's Alexa have become ubiquitous as artificial intelligence (AI) becomes increasingly sophisticated, and the global market for VUIs is expected to expand substantially in the coming years (Research & Markets, 2023). In response, researchers have extended investigations of trust in technology to AI and cognitive assistants (Anton, Oesterreich, Schuir, & Teuteberg, 2022; Kaplan, Kessler, Brill, & Hancock, 2021; National Academies of Sciences, Engineering, and Medicine, 2022; Siddike & Kohda, 2019). The user experience (UX) and computer science communities have simultaneously delved into best practices for conversational assistants

URI: https://hdl.handle.net/10125/106488
978-0-9981331-7-1

(Pearl, 2016; Platz, 2020; Deibel & Evanhoe, 2021; Murad, Cowan, Munteanu, & Clark, 2019).

Trust is a varied and multi-faceted concept. Siddike and Kohda (2019) define trust in cognitive assistants (a close proxy to VUIs for safeguards inspectors) as the belief that the technology will help one reach a desired decision. Lee and See (2004) note the importance of trust when uncertainty and vulnerability are inherent to the situation. Trust in technology often also incorporates measures of reliance on the technology as well as compliance with its suggestions.

Performance of the system, predictability of the system (match to user expectations), system reliability and dependability (consistency and effectiveness), transparency of decisions (verification), and provenance (source of information) are common factors highlighted in trust taxonomies (Schaefer, Chen, Szalma, & Hancock, 2016; Siddike & Kohda, 2019; Chien, Lewis, Semnani-Azad, & Sycara, 2014; McGuinness, Glass, Wolverton, & da Silva, 2007; Maier, 2021; Jung, Dorner, Weindhardt, & Pusmaz, 2018). Trust is also built over time through consistent positive experiences (Lee & See, 2004; Merritt & Ilgen, 2008). Finally, VUIs should be designed using UX best practices for both content and interaction patterns, since a system that is seen as both useful and usable will be more likely to be adopted and trusted (Anton, Oesterreich, Schuir, & Teuteberg, 2022).

1.3. Research Focus

This work targets the intersection of trust in technology with VUI digital assistants in the nuclear safeguards application domain. The experiments reported here are the first of a larger research line aimed at examining the role of trust factors such as explainability, granularity (level of detail provided), provenance, and confidence for VUI digital assistants in the context of key nuclear inspections tasks (seal examination and nuclear materials measurement). Subsequent work in this research line will also extend and replicate these findings with nuclear safeguards subject matter experts.

Here, we quantitatively measure trust in the VUI as compliance with the voice assistant's recommendations and its interaction with participant accuracy and response time on the inspection task. We also collect subjective ratings of trust and reliability (compared against actual reliability) as well as responses to the Trust of Automated Systems Test (TOAST) questionnaire. TOAST is a scale for measuring trust in automated systems that aligns well with our targeted tasks and has previously been applied to national security settings (Wojton, Porter, Lane, Bieber, & Madhavan, 2020). Individual differences in propensity toward technology adoption were also measured via the Affinity for Technology (ATI) scale (Franke, Attig, & Wessel, 2018).

We hypothesize that participant decisions will be influenced by input from the VUI, and that their performance on the inspection task will correlate with subjective measures of trust. These experiments will also serve as a baseline against which to test direct manipulations of additional factors we hypothesize will positively influence trust in the VUI (e.g., providing explainability information on the location of detected issues with a seal) as well as the impact of professional knowledge (e.g., VUIs may need to provide more or less detailed information due to inspectors' tamper detection skill). They also build upon prior work examining the cognitive impact of errors from *visual* machine learning-based assistance for visual analysis tasks within the international nuclear safeguards domain (Gastelum, Matzen, Divis, & Howell, 2022; Divis, Howell, Matzen, Stites, & Gastelum, 2022) and expand burgeoning work on trust in VUIs from industry applications to a high-consequence, global security space. Factors important for trusting a VUI may manifest differently when global security is on the line (e.g., detecting illicit nuclear materials or a network attack) compared to everyday digital assistant tasks (e.g., starting a timer while baking).

2. Methodology

We conducted two experiments examining user trust in a VUI within the context of an international nuclear safeguards seal examination task. Participants were asked to imagine they worked for the United Nations as part of a specialized team of safeguards inspectors. Inspectors secure containers with seals that indicate if containers have been opened since the last inspection. After a brief training session on tamper detection that included practice trials, participants were shown simulated seals and asked to decide whether to keep or remove each seal. Seals showing normal wear and tear only were to be kept in place. Seals showing signs of tamper were to be removed and replaced. Participants viewed each seal and heard recommendations provided by the Voice Assistant Laboratory ("VAL") digital assistant. VAL read the seal ID and indicated whether her analysis identified signs of tamper. Participants were told that VAL was not always right but would provide information intended to help them make their decision. Experiment 2 also varied the accuracy of VAL's vocalization of the seal ID. These studies were designed to evaluate the impact of manipulations of the seal checking task on user trust in the system.

2.1. Participants

Data was collected on Pavlovia (www.pavlovia.org) and Prolific (www.prolific.co) using participants in America fluent in English with a 90% minimum prior approval rate. Participants were compensated at greater than or equal to New Mexico minimum wage ($6 for 15-30 minutes of participation). As an incentive to perform the task carefully, participants were also provided with a three-cent bonus for each trial they got correct (up to $2.25). Experiment 1 included data from 30 participants (after removing and replacing data from two participants who either failed attention check questions or had technical difficulties with the study). Fifteen participants reported their gender as male (15 female), and their mean age was 31.8 (*SD* = 8.0) years old. Experiment 2 included data from 41 participants (after removing and replacing data from three participants who failed attention check questions). Nineteen participants reported their gender as male (22 female), and their mean age was 38.7 (*SD* = 12.2) years old.

2.2. Materials

Stimuli for the seal examination task were created by layering a generic seal image, seal ID, wear and tear pattern, and (optional) tamper pattern[1]. See Figure 1 for an example seal. Wear and tear patterns were represented with thick, rounded marks that occurred on the face of the seal. We included both low wear and tear patterns (fewer markers with more transparency) and high wear and tear patterns (more marks with less transparency). Tamper marks were represented by thin, jagged lines which could occur on the bottom or front face of the seal. They could also be high difficulty (more transparent) or low difficulty (less transparent). Seal IDs were randomly generated unique 5-digit numbers.

Participants were shown descriptions and examples of both wear and tear patterns and tamper patterns during the instructions. Practice trials provided additional feedback to help participants calibrate their understanding of tamper detection within this task. The wear and tear patterns and tamper patterns were also normed for detection difficulty in an independent study prior to running the experiments reported here. The norming study was designed to calibrate difficulty of the patterns and did not include assistance from a VUI. We tested 40 wear and tear patterns (20 low; 20 high) and 30 tamper patterns (15 within a quadrant on the front of the seal; 15 within either side of the bottom of the seal). Those tamper patterns were tested at four levels of transparency to vary detection difficulty. Across three between-subjects counterbalancing conditions, 45 participants indicated the presence or absence of tamper on seals with a wear and tear pattern only, tamper pattern only, a combination of wear and tear and tamper patterns, or a control (no wear and tear or tamper patterns). Based on participant accuracy in the norming study, patterns were selected to create stimuli for subsequent experiments so that difficulty was neither too high (i.e., detection at floor in the norming study) or too low (i.e., detection at ceiling in the norming study). Average tamper detection in the norming study was 48% for difficult patterns and 91% for easier patterns. This selection led to tampered seals that could be quite challenging—but not impossible—to detect, with independently calibrated difficulty metrics.

Participants were also provided with a visual representation of VAL. When VAL was run on a seal, the visual representation was animated and VAL provided the analysis (e.g., "Analyzing seal. Seal 46152 shows normal wear and tear." or "Analyzing seal. Seal 35489 shows signs of tamper."). VAL had a female voice with an American accent.

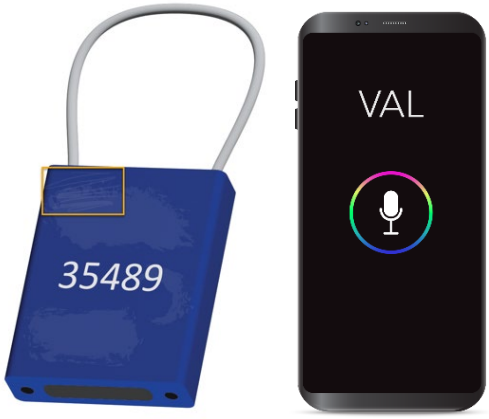

Figure 1. Example of seal and the Voice Assistant Laboratory (VAL). Larger blotches on front face of the seal are examples of wear and tear. Narrow scratches in upper left corner are an example of tamper (outlined with yellow box for demonstration purposes only). Images are not to scale.

At the conclusion of the experiment, participants were given a post-task questionnaire. Participants rated their trust in VAL ("On a scale from 0 to 10, how much did you trust VAL?") and the reliability of VAL ("On a scale from 0 to 100, how reliable did you find VAL's information?"). They then completed the TOAST questionnaire (Wojton, Porter, Lane, Bieber, & Madhavan, 2020) and the ATI scale (Franke, Attig, & Wessel, 2018).

[1] The seals were created for experimental purposes only. They intentionally were not representative of actual IAEA or U.S. Government seals, nor were the markings representative of actual wear and tear or tamper events.

2.3. Procedure

Participants provided consent to participate and then worked through instructions and practice trials before starting the main seal examination task. For each seal, participants saw a representation of VAL and clicked to run her on the next seal. VAL verbally provided her analysis, and the seal was displayed. Participants then decided whether to keep the seal, remove and replace the seal, or re-run VAL. Participants could only re-run VAL once per seal.

Participants saw 75 experiment trials plus 3 attention check trials. Seals could either be normal (no tamper) or tampered. VAL would indicate either that the seal showed normal wear and tear or that it showed signs of tamper. Forty seals were *true negatives* (normal seals that VAL said were normal). Twenty seals were *true positives* (tampered seals that VAL said were tampered). Five seals were *false negatives* (tampered seals that VAL said were normal). Ten seals were *false positives* (normal seals that VAL said were tampered). Therefore, 67% of the seals were normal. VAL correctly detected tamper (presence or absence of tamper) 80% of the time. If participants re-ran VAL on a seal where her initial analysis was correct (true negatives and true positives), then her analysis remained the same. If participants re-ran VAL on a seal where her initial analysis was incorrect (false negatives and false positives), then her analysis could change (50% of false negative and false positive seals were flagged *a priori* to change to the correct analysis).

All trials were counterbalanced across low and high wear and tear patterns, low and high difficulty tamper patterns, and tamper location (front or bottom) for VAL decision categories. Participants were split between two between-subjects counterbalancing conditions that included different combinations of seal patterns.

Participants completed the post-task questionnaire after finishing the seal examination ask. They were then thanked for their participation and given payment. The task took about 15-30 minutes to complete.

Experiment 2 included an additional manipulation in the seal examination task for seal ID. In 8 of the 40 true negative seals (for which the correct response would have otherwise been to leave the seal in place), VAL transposed two numbers in the seal ID (e.g., 34589 instead of 35489). Participants were told in the instructions that they should remove any seals where VAL provided the wrong seal ID. Over the entire seal examination task in Experiment 2, VAL provided the correct seal ID 89.3% of the time. VAL provided the correct seal ID and correctly detected tamper 69.3% of the time.

3. Results

All statistical tests were held at an $\alpha = .05$ level. Unless otherwise noted, mixed-effects models with Tukey corrections for multiple comparisons were used for all statistical analyses.

3.1. Experiment 1

Data were dropped for 12 (out of 2250) trials due to response times (RTs) greater than three standard deviations beyond the mean.

Participants rarely re-ran VAL (3.75% of seals); half of the participants re-ran VAL at least one time. There were no statistically significant differences in the post-task questionnaire measures between participants who re-ran VAL at least once and those who did not. Most re-runs occurred on false positives (11.5% of false positive seals). Since RT cannot be consistently interpreted with re-run trials and the patterns of results were similar for accuracy and RT metrics regardless of whether re-run trials were included, re-run trials were excluded from all subsequent analyses unless otherwise noted.

Participants' responses were counted as accurate if they aligned with the ground truth for the seal (normal or tamper). A mixed effects model was run predicting participant accuracy from the fixed effect of VAL decision category (true negative, true positive, false negative, or false positive) with random effects for participant. All differences were statistically significant (all $Z > 3.72$ and $p < .0001$) except for false negatives relative to false positives. Accuracy was highest for true negatives followed by true positives. Accuracy was lowest when VAL provided an incorrect recommendation (false negatives and false positives). See Figure 2.

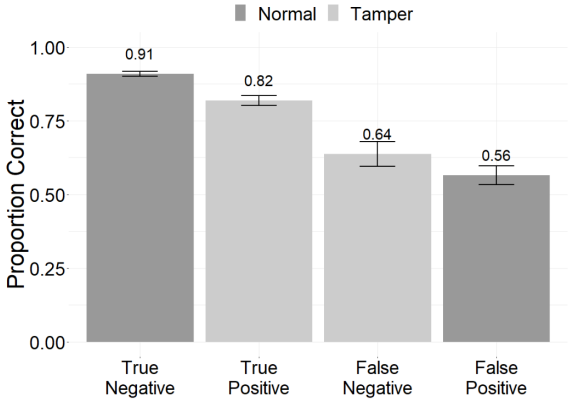

Figure 2. Proportion correct in Experiment 1 by ground truth (normal or tamper) and category of response from VAL (excluding re-runs). Error bars represent standard error of the mean.

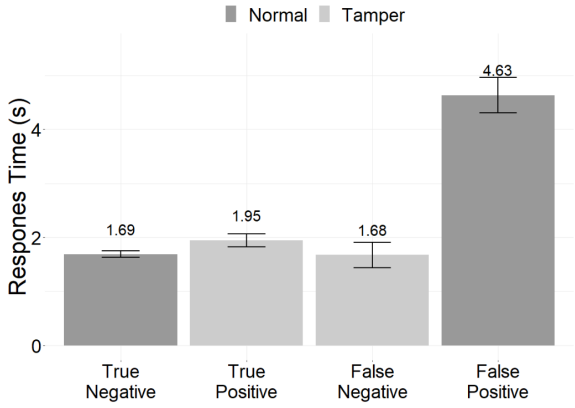

Figure 3. Response time in seconds for correct trials in Experiment 1 by ground truth (normal or tamper) and category of response from VAL (excluding re-runs). Error bars represent standard error of the mean.

Analysis of RT data was limited to accurate trials. A mixed effects model was conducted predicting RT from the fixed effect of VAL decision category with random effects for participant. Decisions for false positive trials were significantly slower than all other categories (all $Z > 3.72$ and $p < .0001$). No other significant differences were found between decision categories. See Figure 3.

Participants' mean trust rating was 6.50 on a scale from 0 to 10 ($SD = 1.94$); their mean reliability rating was 68.33 on a scale from 0 to 100 ($SD = 18.72$). Not accounting for re-runs[2], VAL's actual reliability was 80%; participants significantly underestimated VAL's reliability by 11.67% ($t = -3.41$, $df = 29$, $p = .002$).

The TOAST included 9 questions rated on a 7-point Likert scale, with higher ratings indicating higher trust. Participants' mean ratings were 5.82 ($SD = 0.93$) for the System Understandability subcomponent and 3.93 ($SD = 0.92$) for the System Performance subcomponent. The ATI questions were rated on a 6-point Likert scale, with higher ratings indicating higher affinity for technology interaction. Participants' mean ATI rating was 3.65 ($SD = 0.85$).

Pearson's correlations between overall accuracy and the post-task questionnaire are shown in Table 1. Trust and reliability ratings were highly correlated, indicating those questions were treated similarly by participants. Participants' accuracy correlated with the trust and reliability scores they assigned to VAL, but not their TOAST ratings nor ATI self-assessments. The TOAST System Performance subcomponent was also highly correlated with trust and reliability scores. The ATI and TOAST System Usability scores only significantly correlated with one another.

	Trust	Reliability	TOAST-SU	TOAST-SP	ATI
Accuracy	0.52**	0.44*	*n.s.*	*n.s.*	*n.s.*
Trust		0.98***	*n.s.*	0.74***	*n.s.*
Reliability			*n.s.*	0.74***	*n.s.*
TOAST-SU				*n.s.*	0.38*
TOAST-SP					*n.s.*

Table 1. Experiment 1 correlation matrix for overall accuracy on seal examination task, trust rating, reliability rating, TOAST System Understanding (SU) subcomponent, TOAST System Performance (SP) subcomponent, and ATI. Statistical significance is indicated by * ($p < .05$), ** ($p < .01$), *** ($p < .001$), or *n.s.* (not statistically significant).

3.2. Experiment 2

Data were dropped for 43 (out of 3075) trials due to RTs greater than three standard deviations beyond the mean.

Participants re-ran VAL on 10.6% of the seals; 80% of participants re-ran VAL at least one time. There were no statistically significant differences in the post-task questionnaire measures between participants who re-ran VAL at least once and those who did not. Re-runs were most common for false positives (19.7% of false positive seals) and true negatives where VAL transposed the seal ID (18.3% of true negative seals with a transposed ID). Once again, re-run trials were excluded from analysis of RT and accuracy data unless otherwise noted.

Participants' responses were counted as accurate if they aligned with the ground truth for the seal (normal or tamper) and if they removed the seal when the ID was transposed (even if the seal was otherwise normal). A mixed effects model was run predicting participant accuracy from the fixed effect of VAL decision category (true negative with correct ID, true negative with transposed ID, true positive, false negative, or false positive) with random effects for participant. All differences were statistically significant (all $Z > 1.64$ and $p < .05$) except for false negatives relative to false positives. Accuracy was highest for true negatives with correct seal IDs followed by true positives. Accuracy was lowest when VAL provided incorrect information, with lower accuracy when VAL read seal IDs incorrectly (true negatives with transposed IDs) than when she provided incorrect analyses about tamper detection (false negatives and false positives). See Figure 4.

[2] VAL's reliability remains similar whether re-runs are treated as an additional trial ($mean = 78.0\%$, $SD = 3.6\%$) or only VAL's final decision is used to calculate reliability ($mean = 80.7\%$, $SD = 1.7\%$). For the sake of simplicity, all subsequent analyses for Experiment 1 assume VAL has an 80% reliability.

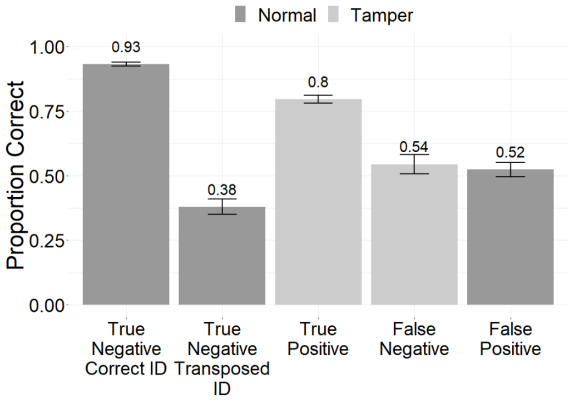

Figure 4. Proportion correct in Experiment 2 by ground truth (normal or tamper) and category of response from VAL (excluding re-runs). Errors bars represent standard error of the mean.

Re-run interactions were of note when considering trials where VAL transposed the seal ID. A mixed effects model was conducted predicting accuracy on seals where VAL transposed the seal ID, with the fixed effect of re-run interaction (no re-run, re-run without a change, and re-run with a change to the correct ID) and random effects for participant. It revealed that participants were significantly more likely to make the correct decision when they re-ran VAL on seals where she initially provided the wrong seal ID, regardless of whether she provided a correct seal ID on the re-run (all $Z > 2{,}58$ all $p < .01$). Accuracy was 38% ($SD = 29\%$) for trials where participants did not re-run VAL, 85% ($SD = 31\%$) for trials where participants re-ran VAL and her answer did not change, and 96% ($SD = 17\%$) for trials where participants re-ran VAL and her answer changed.

Analysis of RT data was limited to correct trials. A mixed effects model was conducted predicting RT from the fixed effect of VAL decision category with random effects for participant. Decisions for false positive trials were significantly slower than all other categories (all $Z > 3.72$ and $p < .0001$). True negatives with transposed IDs were also significantly faster than true negatives with correct IDs or true positives (all $Z > 1.64$ and $p < .05$). No other significant differences were found between decision categories. See Figure 5.

Participants' mean trust rating was 6.15 ($SD = 2.09$). VAL's reliability[3] for getting both the tamper detection and seal ID correct was 69.3%. Participants' mean reliability rating of 66.3% ($SD = 19.6\%$) was not statistically significantly different from VAL's actual reliability. Participants' mean ratings on the 7-point Likert scale TOAST questions were 5.69 ($SD = 0.90$) for the System Understandability subcomponent and 3.69 ($SD = 0.97$) for the System Performance subcomponent. Participants' mean rating on the 6-point Likert scale ATI was 3.65 ($SD = 0.96$).

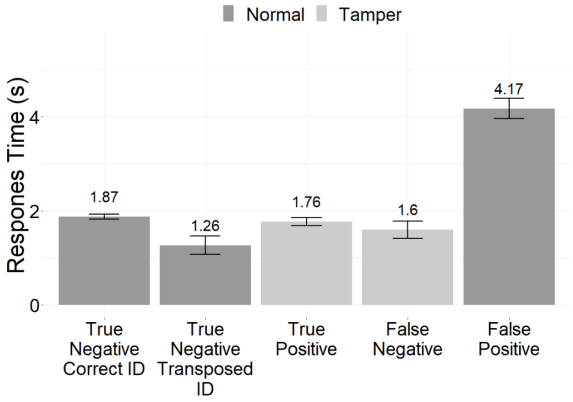

Figure 5. Response time in seconds for correct trials in Experiment 2 by ground truth (normal or tamper) and category of response from VAL (excluding re-runs). Error bars represent standard error of the mean.

See Table 2 for Pearson's correlations between accuracy on the seal examination task and the post-task questionnaire. Once again, trust and reliability were highly correlated. However, only the TOAST System Usability subcomponent correlated with accuracy in the seal examination task. The TOAST System Performance subcomponent correlated with trust, reliability, and TOAST System Understandability. The ATI was not significantly correlated with any other measure.

When comparing the post-task questionnaire responses across the two experiments there were no statistically significant differences.

	Trust	Reliability	TOAST-SU	TOAST-SP	ATI
Accuracy	*n.s.*	*n.s.*	0.36*	*n.s.*	*n.s.*
Trust		0.95***	*n.s.*	0.74***	*n.s.*
Reliability			*n.s.*	0.72***	*n.s.*
TOAST-SU				0.33*	*n.s.*
TOAST-SP					*n.s.*

Table 2. Experiment 2 correlation matrix for overall accuracy on seal examination task, trust rating, reliability rating, TOAST System Understanding (SU) subcomponent, TOAST System Performance (SP) subcomponent, and ATI. Statistical significance is indicated by * ($p < .05$), ** ($p < .01$), *** ($p < .001$), or *n.s.* (not statistically significant).

4. Discussion

Across both experiments, participants were most accurate when VAL provided correct information (with

[3] Once again, reliability was similar when counting each re-run as a new trial (*mean* = 65.6%, *SD* = 5.6%) or only using VAL's final decision in the reliability calculation (*mean* = 71.4%, *SD* = 3.7%). Analyses for Experiment 2 use the simple reliability metric which does not include re-runs (69.3%).

highest accuracy on true negatives seals with the correct ID). Participants' accuracy dropped by approximately 35-40% when VAL detected tamper on a normal (no tamper) seal and by approximately 20-25% when VAL did not detect tamper on tampered seals. This drop in performance indicates that participants were influenced by VAL's recommendations. This interpretation is further supported by substantially slower response times (2-3 seconds slower) for false positive trials. Participants spent more time searching for signs of tamper on a normal seal when VAL incorrectly indicated tamper.

Slower response times for false positives might be alleviated by providing explanations such as location information for the detected tamper. Verbal descriptions of the location of detected tamper are similar to the visual bounding boxes commonly displayed by machine learning algorithms during visual target detection tasks. However, this additional explainability information can also reduce human accuracy performance due to overconfidence in the model's decision (e.g., as with bounding boxes in Cunningham, Drew, & Wolfe, 2017). Follow-on work providing additional explainability information for VAL's recommendation will help to inform the potential costs and benefits of including supplementary explainability information in the VUI.

Participants missed VAL's incorrect reading of the seal ID on 62% of the transposed seal ID trials. This outcome could indicate that participants were not carefully listening to VAL's reading of the seal ID and/or they were focused more on the tamper detection task than on VAL's analysis. While re-runs were uncommon, participants who re-ran VAL when she provided an incorrect seal ID were more likely to make the correct decision than those who did not re-run VAL. These participants might have used the re-run feature to check VAL's incorrect seal ID. In our targeted application of IAEA nuclear safeguards, inspectors must perform multiple tasks while examining seals—including confirming seal IDs, checking for signs of tamper, and checking against a list of seals pre-selected for replacement or off-site examination. Similar multitasking needs will hold for other global security applications. Voice assistants must be able to support all critical tasks—and at the very least, not undermine secondary tasks such as seal identification.

While the accuracy and response time metrics give us insight into the influence VAL had on participants' behavior and their compliance with her recommendations, the post-task questionnaire allows us to examine participants' subjective ratings of trust in the VUI more directly. Across both experiments, trust and reliability ratings were highly correlated, indicating that participants thought of the two concepts similarly for this study. The TOAST System Performance subcomponent also tended to correlate with trust and reliability ratings, whereas the TOAST System Understandability subcomponent did not. This pattern is consistent with past work showing that System Performance is more important for trust than System Understanding (Wojton, Porter, Lane, Bieber, & Madhavan, 2020). Participants tended to give numerically higher ratings for VAL's System Understandability than System Performance, indicating that understanding how VAL worked was less of an issue than satisfaction with her performance. They also tended to underestimate VAL's reliability in the first experiment where seal IDs were always correct. Participants' reliability ratings were better calibrated for the second experiment where seal IDs were sometimes incorrect. It is unclear whether the improved calibration in the second experiment was driven by participants noticing the incorrect seal IDs or by VAL simply having a lower reliability in the second experiment that better matched participants' expectations. The relationship between accuracy performance in the seal examination task and subjective trust ratings was not stable across the two experiments—trust and reliability scores were positively correlated with accuracy in the first experiment, but only system understandability was positively correlated with accuracy in the second experiment. The exact mechanism driving that change is unclear, but it highlights the importance of considering the complexities of the targeted application when measuring trust in a VUI.

International nuclear safeguards inspectors must have appropriately calibrated trust in their tools, including future digital assistants. Tools that are trusted too little lose their utility; while tools that are trusted too much can lead to complacency. Either of these scenarios could result in high-consequence errors. We hypothesize that factors such as explainability, confidence, granularity, and/or provenance information can influence users' trust in a VUI decision aid. While these studies focus on nuclear safeguards applications, the findings also have implications for trust in cognitive assistants and/or VUIs in personal, industry, and government settings (National Science Foundation, 2018; National Academies of Sciences, Engineering, and Medicine, 2022; Pearl, 2016). This work also ties back to VUI considerations in other high-consequence settings such as emergency response teams (Preum, et al., 2018) and pilots (Estes, et al., 2018), building on and expanding important factors for VUI usability and trust such as environmental noise and perceived distraction.

These experiments are the first in a line of research examining trust in VUI assistants for safeguards applications; they provide a proof of concept and baseline to test against in the future. Here, we found that the VUI influenced participants' decision making and

that participants' perceptions of the VUI's reliability were not always well calibrated with its underlying performance. In future work, we will examine how varying tamper location explanations and confidence information impact trust and performance. We will also expand the tasking to include nuclear material measurements and replicate key findings with nuclear safeguards professionals. The results will inform best practices and guidelines for VUIs and high-consequence domains in general, but particularly for future nuclear safeguards voice assistants.

6. Acknowledgements

We would like to thank Haley Norris for her help in designing the seals and pattern layers. This research was funded by the National Nuclear Security Administration's Office of Defense Nuclear Nonproliferation. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

7. References

Anton, E., Oesterreich, T. D., Schuir, J., & Teuteberg, J. (2022). Painting A Holistic Picture of Trust in and Adoption of Conversational Agents: A Meta-Analytic Structural Equation Modeling Approach. *Proceedings of the 55th Hawaii International Conference on System Sciences*, 5871-5880.

Bell, S., Benaloh, J., Byrne, M. D., DeBeauvoir, D., Eakin, B., Kortum, P., . . . Winn, M. (2013). STAR-Vote: A secure, transparent, auditable, and reliable voting system. *USENIX Journal of Election and Technology Systems*, 18-37.

Chien, S.-Y., Lewis, M., Semnani-Azad, Z., & Sycara, K. (2014). An empirical model of cultural factors on trust in automation. *Proceedings of the human factors and ergonomics society annual meeting. 58(1)*, pp. 859-863. Los Angeles, CA: SAGE Publications.

Cunningham, C. A., Drew, T., & Wolfe, J. M. (2017). Analog computer-aided detection (CAD) information can be more effective than binary marks. *Attention, Perception, & Psychophysics, 79*, 679-690.

Deibel, D., & Evanhoe, R. (2021). *Conversations with things: UX design for chat and voice.* Rosenfeld.

Divis, K., Howell, B., Matzen, L., Stites, M., & Gastelum, Z. (2022). The cognitive effects of machine learning aid in domain-specific and domain-general tasks. *Proceedings of the 55th Annual Hawaii International Conference on System Science*, (pp. 1-8).

Estes, S., Helleberg, J., Long, K., Menzenski, J., Myles, C., Pollack, M., . . . Stein, J. (2018). Principles for minimizing cognitive assistance distraction in the cockpit. *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*, 1-6. doi:10.1109/DASC.2018.8569802

Franke, T., Attig, C., & Wessel, D. (2018). A personal resource for technology interaction: Development and validation of the Affinity for Technology Interaction (ATI) scale. *International Journal of Human-Computer Interaction, 35*(6), 456-467.

Gastelum, Z. N., Matzen, L. E., Divis, K. M., & Howell, B. (2022). Cognitive impacts of computer vision-based decision support for international nuclear safeguards-relevant visual analysis tasks. *Proceedings of the IAEA Symposium on International Safeguards.*

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517-527.

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2021). Evolving trust in robots: specification through sequential and comparative meta-analyses. *Human Factors, 63*(7), 1196-1229.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407-434.

Jung, D., Dorner, V., Weindhardt, C., & Pusmaz, H. (2018). Designing a robo-advisor for risk-averse, low-budget consumers. *Electron. Mark., 28*(3), 367-380.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2021). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*. doi:10.1177/00187208211013988

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology, 12*.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*, 153-184.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80.

Maier, T. (2021). *An exploration of cognitive assistants and their challenges*. [Doctoral dissertation, The Pennsylvania State University].

Matzen, L. E., Stites, M. C., Howell, B. C., & Gastelum, Z. N. (2021). Different visualizations of machine learning outputs influence the spped and accuracy of user evaluations". *IEEE InfoVis x Vision Science Workshop.*

McGuinness, L. D., Glass, A., Wolverton, M., & da Silva, P. P. (2007). Explaining task processing in cognitive assistants that learn. *Proceedings of AAAI Spring Symposium: Interaction Challenges for Intelligent Assistants*, (pp. 80-87).

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors, 50*, 194-210.

Murad, C., Cowan, B. R., Munteanu, C., & Clark, L. (2019). Revolution or evolution? Speech interaction and HCI design guidelines. *IEEE Pervasive Computing*, (pp. 33-45).

National Academies of Sciences, Engineering, and Medicine. (2022). *Human-AI Teaming: State-of-the-Art and Research Needs.* Washington, DC: The National Academies Press. doi:https://doi.org/10.17226/26355

National Science Foundation. (2018). *Intelligent Cognitive Assistants.* SRC. Retrieved from https://www.src.org/program/ica/research-needs/ica-research-needs-20180207.pdf

Pearl, C. (2016). *Designing voice user interfaces: Principles of conversational experiences.* O'Reilly.

Platz, C. (2020). *Designing beyond devices: Creating multimodal, cross-device experiences.* Rosenfeld.

Preum, S. M., Shu, S., Ting, J., Lin, V., Williams, R., Stankovic, J., & Alemzadeh, H. (2018). Towards a cognitive assistant system for emergency response. *ACM/IEEE International Conference on Cyber-Physical Systems*, (pp. 347-348).

Research & Markets. (2023). *Voice User Interface Global Market Report 2023.* Retrieved from www.researchandmarkets.com/reports/5783085/

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377-400.

Sheridan, T. B. (2019). Individual differences in attributes of trust in automation: measurement and application to system design. *Frontiers in Psychology, 10*, 1117.

Siddike, M. A., & Kohda, Y. (2019). Trust in cognitive assistants: A theoretical framework. *International Journal of Applied Industrial Engineering, 6*(1), 60-71.

Smartt, H., Gastelum, Z., Rutkowski, J., Peter-Stein, N., & Shoman, N. (2021). Hey Inspecta! *Proceedings of the INMM & ESARDA Joint Virtual Annual Meeting.*

Wojton, H. M., Porter, D., Lane, S. T., Bieber, C., & Madhavan, P. (2020). Initial validation of the trust of automated systems test (TOAST). *Journal of Social Psychology*, 1-16.