Journal of Artificial Intelligence Research 80 (2024) 1311-1341 Submitted 05/2023; published 08/2024

Towards Trustworthy AI-Enabled Decision Support Systems: Validation of the Multisource AI Scorecard Table (MAST)

Pouria Salehi PSALEHI@ASU.EDU
Yang Ba YANGBA@ASU.EDU
Nayoung Kim NKIM48@ASU.EDU
Ahmadreza Mosallanezhad AMOSALLA@ASU.EDU
Anna Pan AROLSO10@ASU.EDU
Myke C. Cohen MYKE.COHEN@ASU.EDU
Yixuan Wang YWAN1290@ASU.EDU
Jieqiong Zhao JZHAO153@ASU.EDU
Shawaiz Bhatti SABHATT1@ASU.EDU
Arizona State University, USA

James Sung JAMES.SUNG@HQ.DHS.GOV
DHS Office of Intelligence and Analysis, USA

Erik Blasch ERIK.BLASCH.1@US.AF.MIL
Air Force Office of Scientific Research, USA

Michelle V. Mancenido MVMANCENIDO@ASU.EDU
Erin K. Chiou ECHIOU@ASU.EDU
Arizona State University, USA

Abstract

The Multisource AI Scorecard Table (MAST) is a checklist tool to inform the design and evaluation of trustworthy AI systems based on the U.S. Intelligence Community's analytic tradecraft standards. In this study, we investigate whether MAST can be used to differentiate between high and low trustworthy AI-enabled decision support systems (AI-DSSs). Evaluating trust in AI-DSSs poses challenges to researchers and practitioners. These challenges include identifying the components, capabilities, and potential of these systems, many of which are based on the complex deep learning algorithms that drive DSS performance and preclude complete manual inspection. Using MAST, we developed two interactive AI-DSS testbeds. One emulated an identity-verification task in security screening, and another emulated a text-summarization system to aid in an investigative task. Each testbed had one version designed to reach low MAST ratings, and another designed to reach high MAST ratings. We hypothesized that MAST ratings would be positively related to the trust ratings of these systems. A total of 177 subject-matter experts were recruited to interact with and evaluate these systems. Results generally show higher MAST ratings for the high-MAST compared to the low-MAST groups, and that measures of trust perception are highly correlated with the MAST ratings. We conclude that MAST can be a useful tool for designing and evaluating systems that will engender trust perceptions, including for AI-DSS that may be used to support visual screening or text summarization tasks. However, higher MAST ratings may not translate to higher joint performance, and the connection between MAST and appropriate trust or trustworthiness remains an open question.

1. Introduction

Decision-making in high-stakes domains may become increasingly dependent on advanced artificial intelligence (AI) (Phillips-Wren, 2012; Zhu et al., 2022), as a way for organizations to meet high service demands with limited human resources (Knop et al., 2022). However, alongside these advancements, there is growing concern about the trustworthiness of AI-enabled decision support systems (AI-DSSs), particularly where AI performance issues could result in catastrophic consequences (Cooke & Durso, 2007). In these high-stakes contexts, applications of AI technologies are typically designed with human specialists in-the-loop.

Human-in-the-loop systems often integrate human supervision with AI processes to help ensure that decisions are both contextually informed and data-driven (Parasuraman & Wickens, 2008). Trust plays a pivotal role in human-in-the-loop systems, because trust influences people's willingness to engage with those systems (Lee & See, 2004). For safety-critical and time-constrained tasks, trust becomes even more crucial when people are not able to consistently monitor or intervene with AI recommendations or actions. Consequently, designing for trustworthy AI-DSSs must draw a delicate balance between enabling a human supervisor's ability to intervene and AI actions.

Existing design frameworks and best practice guidelines for human-AI systems are generally broad-stroked in their recommendations (e.g., de Visser et al., 2020; Schaufeli et al., 2002). The challenge of translating these recommendations into implementable features of AI technologies has tested the overall practicality and impact of these frameworks on developing and built systems. More specific guidance that can be operationalized with minimal design-test-evaluation cycles to get to effectiveness, remains an ongoing pursuit for both researchers and practitioners.

To bridge this gap, the Multisource AI Scorecard Table (MAST) was developed as a structured checklist to aid in designing, testing, and, evaluating AI systems for trustworthiness (Blasch et al., 2020; Sung et al., 2019). MAST reflects Intelligence Community Directive (ICD) 203, which has nine tradecraft criteria for evaluating the quality of human intelligence reporting (ODNI, 2015). These criteria include sourcing, uncertainty, distinguishing, analysis of alternatives, customer relevance, logic, change, accuracy, and visualization. However, MAST extends these criteria to include aspects of data transformation, aggregation, labeling, data display, and contextual relevance that cover various phases of AI system life cycle from data collection to continuous monitoring (Blasch et al., 2019). The underlying premise is that by integrating these nine criteria into system design, AI outputs will be more transparent and trustworthy, thereby improving system utility and effectiveness with a human-in-the-loop. While MAST's usefulness has been demonstrated through several case studies in intelligence and reconnaissance tasks (Blasch et al., 2020; Sung et al., 2019), empirical studies dedicated to validating this tool are lacking.

To address this validation issue, we first apply the MAST framework to the design of two AI-DSS emulators. Facewise is an identity-verification system similar to those used in security screening and READIT (REporting Assistant for Defense and Intelligence Tasks) is a system for text-summarization and data visualization. We then investigate two key aspects, (1) the potential of the MAST to aid in the design and evaluation of human-AI systems that reflect human trust perceptions, and (2) the broader applicability of MAST in assessing the trustworthiness of AI-DSSs in safety critical environments, beyond its use in

intelligence or reconnaissance task contexts. This paper focuses on these two key aspects and the overall validation of MAST. A deeper dive on our design process using MAST is reported in a separate paper, for scope (Kim et al., 2024).

The results of this work offer valuable insights into the utility of MAST as a tool for the design and evaluation of AI systems, and also contributes to the current body of knowledge about trust in AI-enabled systems. Our findings suggest that integrating the nine MAST criteria into AI system design positively influences people's trust perceptions. Moreover, we find that MAST as a design tool is effective in improving trust perceptions, not only in systems designed for intelligence tasks but also in a broader range of AI-enabled applications. However, the study also uncovers potential limitations of MAST, suggesting areas for future research. An important finding, echoing previous findings in other research, shows that high trust perceptions and in this case high MAST scores also do not necessarily translate to higher human-AI system performance.

Overall, this study underscores the challenge of operationalizing universal criteria that can improve human-AI system performance and that can effectively incorporate trust concepts into human-AI system design. Despite these challenges, our findings support the potential of MAST as a viable tool for system design teams. It contributes to aligning researcher and practitioner norms, facilitates the documentation of essential transparency information, and can help engender trust perceptions of systems used in safety-critical tasks.

2. Background

The role of people as supervisors of imperfect automation has a long history (Bainbridge, 1983; Sheridan, 1975). In this supervisory structure, people are tasked with assessing and, if necessary, intervening in automated outputs. However, many AI-DSSs are designed for task environments in which people may not have the cognitive and physical resources to sufficiently understand, assess, and intervene with every recommendation (McGuirl & Sarter, 2006). This is especially a problem in high-stakes domains if people are expected to attend to every outcome produced by the AI-DSSs.

Limitations of human decision-making coupled with imperfect AI-DSSs have resulted in novel problems, some of which resulted in catastrophic outcomes. An infamous case is from the Iraq war, in which the Patriot missile DSS erroneously identified allied fighter jets as enemy aircraft. Operators of the missile system approved the DSS-recommended decision to attack the aircraft, causing the fratricide of American and British pilots. More recently, a series of wrongful arrests in the United States was traced to law enforcement reliance on facial recognition technologies that have considerable racial and gender biases (e.g., Hill, 2022; Hill and Mac, 2023). However, upon recognizing errors in AI recommendations, there is then the tendency to over-correct in rejecting future AI recommendations (e.g., automation aversion; Dietvorst et al., 2015), especially by experts (Snow, 2021).

People's tendency to overuse, misuse, or disuse DSS has long been linked to poorly calibrated perceptions of the DSS's trustworthiness with respect to its actual reliability (Parasuraman & Riley, 1997). As such, methodological frameworks, policy guidelines, and other tools for designing and evaluating DSS trustworthiness have proliferated alongside advancements in AI-DSS capabilities. These include but are not limited to, the Microsoft UX Design Principle (Microsoft, 1995), NISTIR 8330 by National Institute of Standards

and Technology (Stanton & Jensen, 2021), AI Fairness 360 Toolkit by IBM (Bellamy et al., 2019, and others), IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (Chatila & Havens, 2019), UXPA Guidelines for Trustworthy User Experiences (Kriskovic et al., 2017), or Ethical OS Toolkit (Lilley et al., 2020). Although these tools do not all explicitly focus on the concept of trust and trustworthiness, they share an underlying motivation that the design, development, and evaluation of technological systems that impact people and organizations require attention to human factors.

Despite having many frameworks and tools to guide design of trustworthy systems, designing for trust and evaluating trustworthiness in practice remains a challenge. There is a wide gap between theory and practice, partly because trust is an abstract construct with myriad related concepts. For example, designing trustworthy systems may also involve designing for transparency, individual differences, workload, situation awareness, and attending to factors like etiquette and anthropomorphism (Hoff & Bashir, 2015; Parasuraman & Miller, 2004). Another challenge is that expert communities with different priorities may define trustworthiness differently. For example, the Intelligence Community might value high-quality data as a foundation for analysis. Within the transportation security community, emphasizing high-quality decisions at the front lines of traveler safety might be more important than being able to inspect the data. Although it may be possible to outline a generalizable ideal of trustworthiness, a pragmatic view requires accepting trade offs.

To address this gap between concept and practice of designing and evaluating for AI system trustworthiness, the Multisource AI Scorecard Table (MAST; Blasch et al., 2020; Sung et al., 2019) was developed by the AI Team of the 2019 Public-Private Analytic Exchange Program, supported by the Office of the Director of National Intelligence and Department of Homeland Security. MAST describes nine criteria derived from analytic tradecraft standards ICD 203 to assess the trustworthiness of intelligence reporting, and additionally includes a four-level quantitative breakdown for each criterion. The idea is that MAST could serve as an easy-to-use checklist for designing trustworthy AI-enabled systems, and for evaluating trustworthiness after system development. Although the principles behind MAST would seem more suitable for intelligence tasks given its focus on information quality and integrity, it is possible that these criteria may be applied to other human-in-the-loop systems used for information-processing and other human decision-making tasks. For example, AI-enabled systems in computer vision, natural language processing, and medical diagnostic tasks may all be rated according to the MAST criteria, e.g., rating the system's sourcing (e.g., credibility of training data), or its ability to describe and propose alternative recommendations. Medical professionals and their patients may be more willing to trust an AI-derived diagnosis and treatment plan if the system was developed to include the MAST criteria of uncertainty, analysis of alternatives, and customer relevance.

In academic literature, several instruments have been developed to measure trust in automation, including instances of AI-enabled automation (Alsaid et al., 2023; Kohn et al., 2021). Many of them have been widely adopted, others have been independently validated. However, these instruments were mainly designed for research or technology evaluation purposes from the perspective of the operators, rather than for system design or technology development. Although these instruments could be considered relatively robust when used appropriately, they suffer from similar limitations as the design frameworks and tools described previously. There remain wide translation gaps, and highly variable interpretation

from principles to practice, given the hundreds of under-specified conditions and decisions that practitioners face. For example, underlying many of these instruments is a nuanced presumption that assessing domain experts' trust in a particular technology, after they have experienced using it, could be some indication of the technology's trustworthiness. This presumed connection between trust and trustworthiness is then flattened in some practitioner circles, where high trust perceptions are equated with high technology trustworthiness, despite most trust experts being careful not to conflate the two.

To situate the MAST tool in the context of current trust scholarship, our primary objective is to assess the construct validity of MAST relative to human trust. Construct validity is the degree to which an instrument measures the construct it was designed to measure (Cronbach & Meehl, 1955). Approaches for evaluating construct validity include multivariate analytical tools, such as Factor Analysis (Raykov & Marcoulides, 2008; Tabachnick et al., 2013), Principal Components Analysis (PCA; Bandalos, 2018), and Structural Equation Modeling (Kline, 2023). The goal of using multivariate analysis in construct validation is to capture, explain, and measure the amount of variation among items for a construct and to associate these with previously validated constructs (Chancey et al., 2017; Jian et al., 2000). This study aimed to validate MAST as an instrument for assessing trust by investigating how MAST items are associated with validated trust questionnaires.

3. General Method

To validate MAST in different contexts, we designed two AI-DSS testbeds, "Facewise" for identity-verification in a security screening task and, "READIT" for text-summarization and data visualization in an investigative reporting task to support intelligence analysis. General descriptions of these testbeds are described next. For details on designing Facewise and READIT using MAST please refer to our design process paper (Kim et al., 2024).

3.1 Testbeds: Facewise and READIT Platforms

Facewise is a simulated 1-to-1 identity-verification system that uses a pre-trained convolutional neural network, further fine-tuned for face recognition using Cross-entropy loss. The system compares an ID photo with an encounter photo and outputs a decision on whether they represent the same identity (match) or different identities (mismatch). For this study, 80 pairs of face images with known ground truth were hand-selected from various facial datasets including the Iranian emotional face database (Heydari et al., 2023), MorphDB (Ricanek & Tesafaye, 2006), VGG (Parkhi et al., 2015), HUMBI (Yu et al., 2020) presented in randomized order. READIT, which stands for the REporting Assistant for Defense and Intelligence Tasks, is an emulated natural language processing system that visualizes, summarizes, and categorizes documents of limited length (news articles, reports, etc.) to expedite intelligence gathering and reporting. READIT uses BERT (Devlin et al., 2019) to generate outputs. To enhance its usefulness and usability, some manual modifications were also applied. The task and dataset for READIT is based on the 2011 IEEE Visual Analytics Science and Technology (VAST) Challenge (SEMVAST Project, 2011).

Both AI-DSS testbeds and use cases were developed based on information gathered from field visits and bi-monthly consultations with operational stakeholders (i.e., national security researchers, practitioners, and analysts). The use case for READIT was selected to

assess MAST within the text-summarization contexts that MAST was originally designed and evaluated for (Blasch et al., 2020). The use case for Facewise was selected to test the validity of applying MAST to a different type of AI capability, in a different type of task environment, while staying within a national security context (i.e., high-stakes domain). The use case for Facewise also represents an increasingly common one in airport security checkpoints, with systems like the CAT-C or CAT-2 (Lim & Cantor, 2021).

Both Facewise and READIT were developed using cloud-based services and a client-server model for participant-AI interaction. For Facewise, we leveraged Amazon Web Services (AWS) and Google Cloud Platform (GCP) for efficient storage and use of resources. We built the client part of the platform with HTML5 and JavaScript. We collected responses from participants on the client side and sent them to GCP through Python3 and Flask library to save them in the database. Similarly, READIT consisted of a JavaScript based client that enables participant-AI interaction, and the server was built using Python3 and Flask library, hosted on GCP. Data visualizations on READIT were implemented using D3.js, a popular open-source JavaScript library for creating custom interactive data visualizations. The implementation code for READIT and Facewise is available at https://github.com/nayoungkim94/PADTHAI-MM.

3.2 Constructs and Measures

In both platforms, system features were manipulated to compose High-MAST and Low-MAST versions with eight outcome variables: MAST criteria ratings; perceptions of risk, benefit, trust, credibility, engagement, and usability; along with scenario-specific performance metrics.

Versions of platforms: High-MAST and Low-MAST. System features refer to the available features that a DSS can provide to its operators. Based on the MAST criteria, two levels of features for each platform were created: High-MAST and Low-MAST. High-MAST features were designed to reach high MAST criteria ratings through a set of rich features that ideally support high task performance. Low-MAST features were designed to reach low MAST criteria ratings, but with a minimum set of features to enable task completion. Both High-MAST and Low-MAST versions were designed to be as equal as possible in terms of engagement and usability to avoid these manifesting as confounding factors. Appendices A and B delineate the MAST criteria and detailed feature descriptions for Facewise and READIT, respectively.

Variables of interest: MAST criteria, risk, benefit, trust, credibility, performance, engagement, and usability. Each DSS was evaluated based on the MAST criteria descriptions and using a Likert-like scale of 1 to 4, with 1 being poor and 4 being excellent. Each MAST criterion was shown in a question format and accompanied by a corresponding feature description in the DSS. The MAST-total score was created by adding up the 9 criteria with a range of 9 to 36. Participant perception of risk and benefit was measured through two items derived from (Weber et al., 2002). Risk was included due to the well-known relationship between trust and risk (Lee & See, 2004) and perceived benefit was included to check whether participants felt that using the DSS was beneficial for the task they were asked to complete. To measure trust, we used two common questionnaires. One is a previously validated, 12-item instrument known to measure general trust perceptions of automation

(Jian et al., 2000; Spain et al., 2008). The second 15-item instrument measures trust by querying about specific types of information known to affect trust – purpose, process, and performance (Chancey et al., 2017). Because MAST is largely focus on the presentation, availability, types, and quality of information presented by the AI system, we also included a measurement for message credibility (i.e., excluding source credibility), adopting a 3-item survey (Appelman & Sundar, 2016).

We measured performance based on average task completion time and a scenario-specific performance metric. For Facewise, this scenario-specific metric was accuracy of the identity-verification task that took roughly 30 minutes to complete. For READIT, it was the score of a written report completed within roughly 60 minutes. For READIT, we asked participants to identify any present terrorist threat based on past news about a fictitious city named "Vastopolis" and to write a 250-word report detailing the type of terrorist activity and name of the group behind it. To ensure that our implementation of different system features across the two testbeds would not cause major differences in perceived system usability and task engagement (potentially affecting perceived trust, risk, and benefit), we also measured participants' perceived usability and engagement. Usability was assessed with a widely-used 10-item questionnaire known as the System Usability Scale (Brooke, 1996) and engagement was assessed with a 17-item questionnaire (Schaufeli et al., 2002). Appendix C presents the scale, example items, number of items for each variable, and their Cronbach's alpha. Apart from the scenario-specific performance metrics, the other dependent variables and covariates were identical for Facewise and READIT.

Because our study participants were subject-matter experts, we manipulated task difficulty to ensure sufficient task engagement. For Facewise, we did this by selecting at least 40 pairs of difficult images largely from a sibling database (Parkhi et al., 2015). Difficulty was determined by assessing a general population sample in a pilot study, wherein study participants were more prone to incorrectly answer challenging pairs on average. The algorithm demonstrated a 95% accuracy rate across test data during model training (Coşkun et al., 2017). However, when tested with the challenging database, the performance dropped to approximately 56%. Participants in this study were not given any information about the algorithm's performance or task difficulty in advance, but they were alerted to the fact that part of their task was to ensure that the correct decision was made with an algorithm that was potentially fallible.

To make the READIT task more difficult, we included 'red herring' documents. Several of these documents were related and collectively formed narratives that presented multiple plausible causes for the terrorist threat scenario, ideally causing sufficiently engaged participants to consider several highly plausible conclusions for their final report. We devised a rubric from 1 to 5 to evaluate their written reports based on their alignment with the VAST challenge's ground truth. A score of 1 or 2 indicated unsatisfactory to less satisfactory content, primarily comprising red herrings or non-ground truth clusters in the final report. A score of 3 represented satisfactory content, with more ground truth clusters than red herrings. A score of 4 signified mostly satisfactory content, mainly consisting of ground truth clusters, while a score of 5 indicated excellent content, comprising only ground truth clusters. Reports were color-coded for clarity, with red indicating red herrings and bold highlighting ground truth clusters. Two researchers independently assessed reports, achieving a 73.91% inter-rater reliability. In case of discrepancies, the lower performance score was applied.

3.3 Procedure

Figure 1 illustrates the general procedure for Facewise and READIT. We first created a virtual hub using the web-based software platform Qualtrics (Qualtrics, 2020) for participants to access the study. After random assignment to one of the two conditions (High-MAST or Low-MAST), participants were asked to input their given participant IDs on the first page of Qualtrics. Next, informed consent approved by our Institutional Review Board (IRB) was obtained. Then, participants were given a description of their task scenarios. To facilitate engagement and a sense of risk in the study scenario, participants in all conditions were told that they were being tasked to complete an important assignment, and that a previous agent assigned to their task had failed, was put on probation, and subsequently demoted. After reading the task scenario, participants then watched a short recorded video demonstration of the interface and features, and were asked to respond to quiz questions about the video. In the video, all DSS versions were presented as technology aids that exist to supplement the participant's own abilities. Afterward, participants performed the study task. Lastly, participants were asked to evaluate the system and their experience by responding to questionnaires including the MAST criteria, risk, benefit, trust, credibility, engagement, and usability. Given our targeted population of subject-matter experts in national security, limited optional demographic information was collected at the end to assess the representativeness of our sample population.

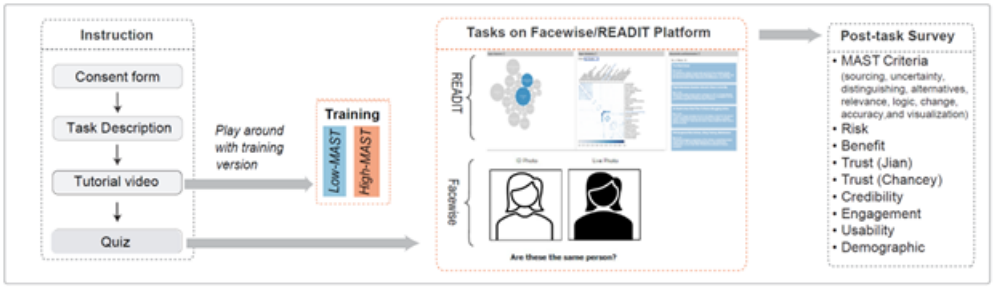


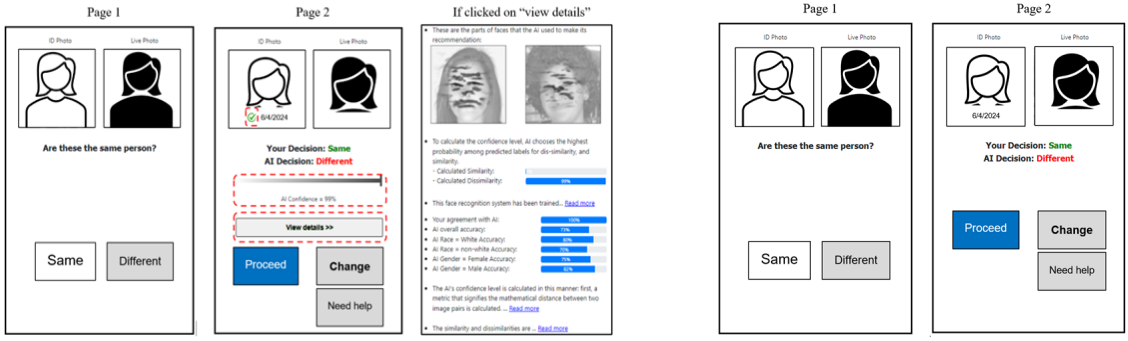
Figure 1: Study Procedure for Facewise and READIT.

3.4 Data Analysis

Data analysis was accomplished in JMP (SAS Institute Inc., 2023) and R using "dplyr" (Wickham et al., 2019), "psych" (Revelle, 2024), "Rmisc" (Hope, 2022), and "compareGroups" (Subirana et al., 2014). Figures were created by "ggmap" (Kahle & Wickham, 2013) and "gridExtra" (Auguie & Antonov, 2017). To confirm associations between the MAST items, trust items, and other validated metrics, we performed the analysis in three steps. First, to compare the different levels of Facewise and READIT, Analysis of Variance (ANOVA) was used. Secondly, Simple Linear Regression (SLR) was used to further investigate the strength and directionality between MAST and other survey measures. Finally, Multivariate analysis via Principal Components Analysis (PCA) was then performed on the perceptual metrics for dimension reduction. We regressed the MAST ratings with the principal component scores. PCA was employed to find coherent and appropriate structures in the perceptual metrics within the first few principal components (Bandalos, 2018).

4. Experiment 1: Facewise Scenario

Participants in the Facewise experiment were told that they were airport security officers tasked to screen passengers by checking their identification materials with the assistance of Facewise, and they had roughly 30 minutes to complete a series of identification verification tasks which is roughly the length of an officer's shift in the document checker position (Greene et al., 2014). Figure 2 outlines the similarities and differences between the two levels of Facewise, High-MAST and Low-MAST. For both levels, Page 1 asks for an initial judgment of human operators. We adopted this structure based on previous work, which we found would increase accuracy (Salehi et al., 2021). In Page 1 (Figure 2), the left image with an off-white background presents the ID photo and the right image with an airport background provides a "live" photo, supposedly taken at the airport. For both levels, these images were cropped and zoomed in for Page 2, which is where most of the differences between High-MAST and Low-MAST appear. Three red dotted lines highlight these differences including the Crossmark/Checkmarks, AI confidence, and a "View Details" button. For more details regarding the AI-DSS features and how they map to each of the MAST criteria, please refer to Appendix A.



(a) Facewise High-MAST (b) Facewise Low-MAST

Figure 2: Comparing Facewise High-MAST (left) and Facewise Low-MAST (right). Icons in this figure replace actual photos used in the study. Red dashes highlight the differences between Low and High platforms. The High-MAST version has more informational features than the Low-MAST version including the ID expiration check, AI confidence level, and "View Details" page.

4.1 Facewise Participants

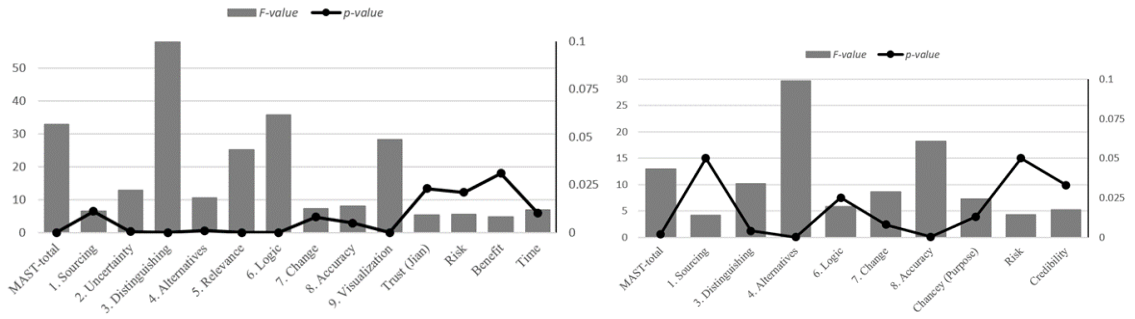
A total of 152 subject-matter experts, U.S. Transportation Security Officers (TSOs), were recruited from three major U.S. airports in Arizona, Nevada, and California, split across 11 days of data collection at the participating airports. Six participants were removed due to very high response time or very low accuracy, resulting in 73 participants each for the High-MAST and Low-MAST conditions. On average, participants spent 76 minutes to complete the entire study. Because participants were federal employees, we were not permitted to provide compensation despite their participation being voluntary. Only light refreshments were provided in appreciation of their participation. Table 5 in Appendix F reports the available participant demographics across the Facewise conditions. Race, ethnicity, and

gender items were not collected due to expressed concerns by some of our collaborative partners, given our limited population of subject-matter experts.

4.2 Facewise Results and Discussion

Table 1 reports descriptive statistics including mean (M) and standard deviations (SD) for the study variables. Results of F statistics in Figure 3 (a) show that participants in the High-MAST group rated Facewise significantly higher across all nine MAST criteria. The High-MAST group found Facewise more trustworthy according to their Jian et al. (2000) scores, less risky, and more beneficial than the Low-MAST group. No significant difference in credibility ratings was found between the two conditions, possibly due to similar system errors experienced in both levels. While the High-MAST group spent significantly more time on the task than the Low-MAST group, they made slightly more accurate decisions than Low-MAST, but this difference was not significant. No significant differences in engagement and usability were found between the High-MAST group and Low-MAST group, meaning we were able to achieve relatively equal engagement and perceived usability for both levels.

Further regression analysis in Figure 4 shows that MAST-total were positively associated with trust; people who rated trust highly also tended to rate MAST highly. Increasing the MAST-total by 1 would increase the Jian et al. (2000) score by 0.1 ($F(1, 144) = 64.94$, $p < .001$, $\beta = 0.1$, $R2 = 0.31$) and the Chancey et al. (2017) score by 0.12 ($F(1, 144) = 87.83$, $p < .001$, $\beta = 0.12$, $R2 = 0.37$). In addition, a positive relationship between MAST and credibility was found; increasing the MAST by 1 would increase credibility by 0.11 ($F(1, 144) = 62.96$, $p < .001$, $\beta = 0.11$, $R2 = 0.30$). Furthermore, this study found that there was a negative correlation between MAST and risk; increasing the MAST by 1 would decrease the risk by 0.067 ($F(1, 144) = 24.32$, $p < .001$, $\beta = -0.067$, $R2 = 0.14$). Finally, there was a positive relationship between the MAST and benefit; increasing the MAST by 1 would increase the benefit by 0.091 ($F(1, 144) = 71.89$, $p < .001$, $\beta = 0.091$, $R2 = 0.33$).



(a) Facewise (b) READIT

Figure 3: The F-test results and corresponding p-values for significant variables in (a) Facewise and (b) READIT are displayed as grey bars and black dots, respectively.

To further validate the association between MAST and other study variables, we needed to run multiple regression analysis. However, because trust, risk, benefit, and credibility were highly correlated, it was inappropriate to run multiple regression analyses. Therefore, we applied Principal Component Analysis (PCA) to reduce the dimensionality within our dataset. The result of PCA in Figure 5 shows that the first two principal components explain

	Facewise			READIT		
	High	**Low**	p	**High**	**Low**	p
MAST-total	26.5 (5.73)	21.0 (5.77)	$< 0.001*$	27.8 (5.38)	19.9 (5.14)	0.002*
1. Sourcing	2.85 (0.84)	2.48 (0.90)	0.011*	3.27 (0.47)	2.50 (1.17)	0.05*
2. Uncertainty	2.85 (0.83)	2.32 (0.97)	$< 0.001*$	2.91 (0.70)	2.42 (0.79)	0.129
3. Distinguishing	3.19 (0.78)	2.15 (0.88)	$< 0.001*$	3.27 (0.65)	2.25 (0.87)	0.004*
4. Alternatives	2.79 (0.82)	2.33 (0.91)	$< 0.001*$	2.91 (0.94)	1.25 (0.45)	0.001*
5. Relevance	3.00 (0.69)	2.34 (0.89)	$< 0.001*$	3.18 (0.75)	3.17 (0.72)	0.961
6. Logic	2.97 (0.87)	2.07 (0.96)	$< 0.001*$	3.18 (0.87)	2.25 (0.97)	0.024*
7. Change	2.88 (0.83)	2.51 (0.82)	0.008*	2.82 (0.75)	1.75 (0.97)	0.007*
8. Accuracy	2.82 (0.87)	2.42 (0.82)	0.005*	3.18 (0.75)	1.92 (0.67)	$< 0.001*$
9. Visualization	3.14 (0.75)	2.42 (0.86)	$< 0.001*$	3.09 (0.94)	2.42 (0.90)	0.095
Trust (Jian)	4.62 (1.12)	4.18 (1.15)	0.023*	5.11 (0.95)	4.42 (1.08)	0.119
Trust (Chancey)	4.26 (1.28)	4.00 (1.16)	0.188	4.57 (1.25)	3.73 (1.35)	0.135
Chancey (Performance)	4.24 (1.42)	3.95 (1.28)	0.188	4.85 (1.31)	3.93 (1.47)	0.126
Chancey (Process)	4.68 (1.39)	4.52 (1.36)	0.472	4.76 (1.56)	4.48 (1.53)	0.669
Chancey (Purpose)	3.87 (1.28)	3.53 (1.16)	0.093	4.09 (1.15)	2.77 (1.20)	0.013*
Risk	2.67 (1.11)	3.10 (1.09)	0.021*	2.55 (0.93)	3.33 (0.89)	0.05*
Benefit	3.37 (0.99)	3.01 (0.98)	0.031*	3.45 (0.93)	3.00 (1.13)	0.303
Credibility	4.31 (1.24)	4.23 (1.29)	0.712	5.39 (0.96)	4.44 (1.03)	0.033*
Average response time (seconds)	13.3 (4.87)	11.3 (4.46)	0.010*	274 (90.7)	214 (93.6)	0.137
Performance (in Platforms)	0.77 (0.08)	0.75 (0.07)	0.097 (accuracy)	2.82 (1.94)	3.33 (1.67)	0.505 (report)
Engagement	3.96 (1.07)	3.99 (1.10)	0.874	4.37 (0.78)	4.52 (1.28)	0.736
Usability	3.58 (0.70)	3.65 (0.60)	0.494	3.49 (0.94)	3.90 (0.67)	0.248

Table 1: Means, Standard Deviation (in parentheses), and p-values of study variables for High-MAST and Low-MAST groups across Facewise and READIT platforms. Asterisk(*) emphasizes the significant differences.

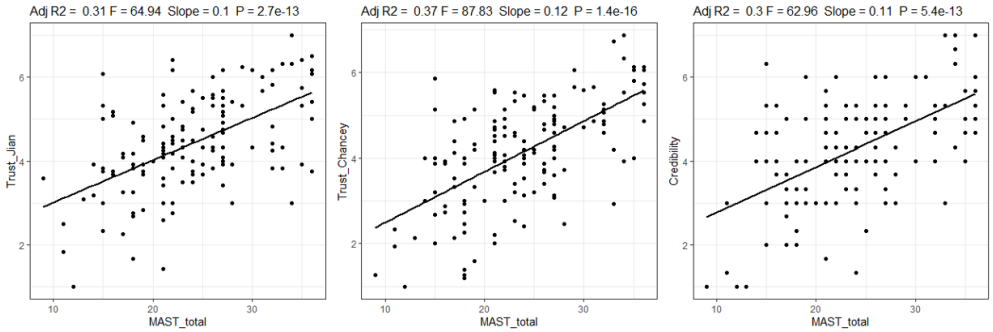
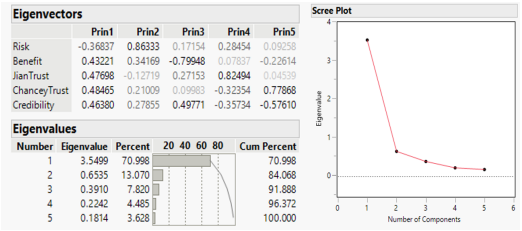
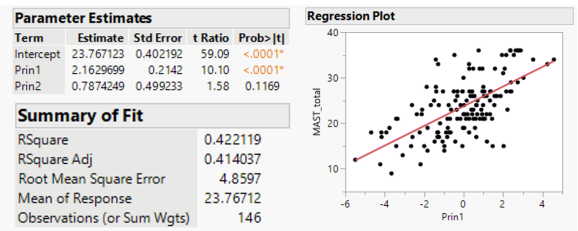


Figure 4: Least Squares Regression plots for Facewise.

84.06% of variation within the dataset. The first principal component can be perceived as an overall average of trust, risk, benefit, and credibility, while the second principal component is mainly related to negative perceptions about risk. These two principal components were used as new variables for a linear regression analysis with MAST performed for each level, High- and Low-MAST. We found that MAST-total was highly associated with the first principal components ($F(1, 144) = 100.92$, $p < .001$, $\beta = 0.19$, $R2 = 0.41$).



(a) PCA

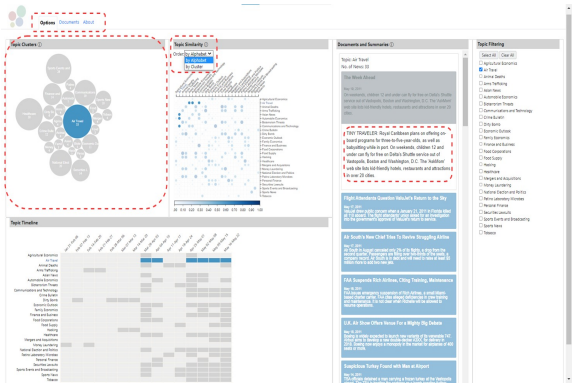


(b) Linear Regression

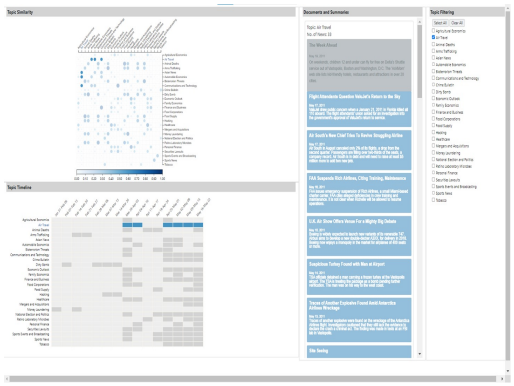
Figure 5: PCA (left) and Linear regression results(right) for Facewise.

5. Experiment 2: READIT Scenario

READIT participants were told they were intelligence analysts in a fictional major city in the United States who were tasked with monitoring the news for any ongoing threats to public safety. READIT participants were given a specific assignment to use READIT to quickly locate and search through relevant news articles and uncover an ongoing terrorist activity that had gone unnoticed for the previous five months. Figure 6 illustrates the similarities and differences between the High-MAST and Low-MAST levels of READIT. Four red dotted lines highlight the differences including the availability of "documents" and "about" tabs (Appendix D), the "topic clusters" bubble graph, the sorting option by cluster relationship strength, and the complete news pieces. For more details regarding the AI-DSS features and how they map to each of the MAST criteria, please refer to Appendix B.



(a) High-MAST READIT



(b) Low-MAST READIT

Figure 6: High-MAST (left) and Low-MAST READIT (right). Compared with the Low-MAST READIT, High-MAST READIT has more interactive features (Topic Clusters, Topic Similarity, original documents, and clickable timelines) to demonstrate the MAST criterion.

5.1 READIT Participants

A total of 25 Intelligence Analysts (IAs) from the U.S. Department of Homeland Security (DHS) were recruited to complete our study, administered through Microsoft Teams or Zoom, over a period of 19 days. Two participants were unable to complete the study due to unexpected scheduling conflicts. The resulting High-MAST and Low-MAST versions of

READIT were tested with a sample of 11 and 12 IAs, respectively. On average, participants spent 75 minutes to complete the study, including onboarding and responses to questionnaire items. We were not permitted to compensate participants monetarily because they were federal employees. However because participants were self-selected volunteers who responded to our recruitment script and were willing to spend time completing our study, we assumed they were sufficiently motivated to complete this study to the best of their ability. Table 6 in Appendix F reports the participant demographics per condition.

5.2 READIT Results and Discussion

Table 1 reports descriptive statistics (M and SD) for the study variables. Results in Figure 3 (b) show that the High-MAST group rated READIT higher on the MAST checklist than the Low-MAST group, and this was significantly different for six out of nine MAST criteria (i.e., except for uncertainty, relevance, and visualization). Trust ratings were also generally higher for those in the High-MAST group; however, the difference was only significant for the "purpose" dimension of Chancey et al. (2017). Moreover, the High-MAST group compared to the Low-MAST group found READIT less risky to use and more credible. No significant differences in performance were found between the High-MAST and Low-MAST groups in terms of average response time or on their 250-word report. However, descriptively, the Low-MAST group spent less time completing the task and had higher performance scores than the High-MAST group. The study found no notable variances in engagement and usability ratings between the High-MAST and Low-MAST groups. This supports our aim to maintain similar levels of engagement and usability across different READIT versions.

Further Regression Analysis in Figure 7 showed that MAST ratings are positively associated with trust ratings. There is a positive relationship between MAST and trust scores; increasing MAST by 1 increases the Jian et al. (2000) score by 0.13 ($F(1, 21) = 32$, $p < .001$, $\beta = 0.13$, $R2 = 0.58$) and increases the Chancey et al. (2017) score by 0.16 ($F(1, 21) = 35.29$, $p < .001$, $\beta = 0.16$, $R2 = 0.61$). A positive relationship was also found between MAST and credibility scores; increasing the MAST by 1 would increase credibility by 0.14 ($F(1, 21) = 52.46$, $p < .001$, $\beta = 0.14$, $R2 = 0.70$). This study also found that there was a negative relationship between MAST and risk; increasing the MAST by 1 would decrease perceived risk by 4.9 ($F(1, 21) = 24.89$, $p < .001$, $\beta = -4.9$, $R2 = 0.52$). In addition, there was a positive relationship between MAST and perceived benefit; increasing the MAST by 1 would increase benefit by 0.11 ($F(1, 21) = 17.19$, $p < .001$, $\beta = 0.11$, $R2 = 0.42$).

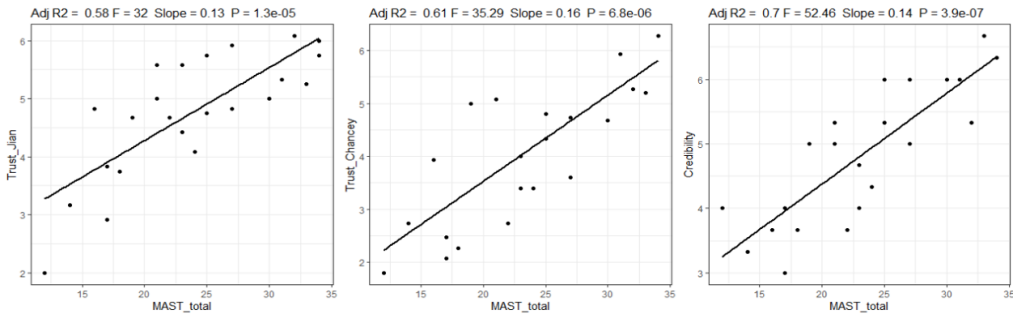
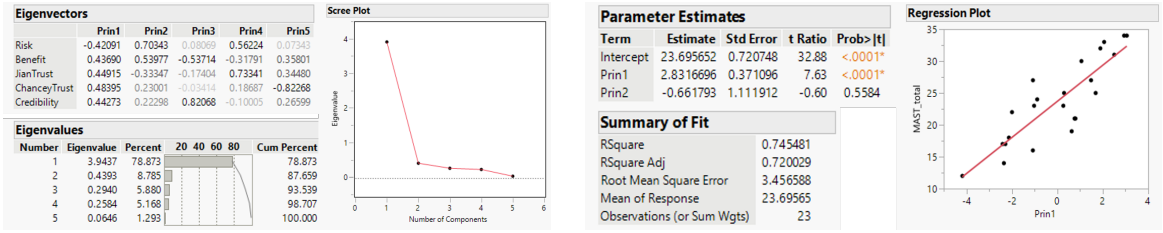


Figure 7: Least Squares Regression plots for READIT.

Because trust, risk, benefit, and credibility were highly correlated for READIT, we could not run multiple regression and instead used PCA. The PCA results in Figure 8 show that the first two principal components can explain 87.65% of variation within the dataset. The first principal component can be interpreted as an overall average of trust, risk, benefit, and credibility. However, the second principal component was primarily related to negative perceptions about risk. For all observations, two PC scores were calculated using each principal component and these scores served as the regressors for further analysis. We found that averaging across all MAST criteria to produce a MAST-total score can significantly predict the first principal components ($F(1, 144) = 60.07$, $p < .001$, $\beta = 0.26$, $R2 = 0.74$).



(a) PCA (b) Linear Regression

Figure 8: PCA (left) and Linear regression (right) results for READIT.

6. General Discussion

In this study, we recruited subject-matter experts to interact with an AI-DSS in their field, either Facewise or READIT platforms, and tested two levels of each platform, a High-MAST version or a Low-MAST version. In this section, we discuss our findings with respect to experts' ratings of these systems and their performance metrics. Then, we elaborate on our analysis of the MAST items and other perceptual measures, and conclude with some caveats regarding our study approach and findings.

Overall MAST ratings. The application of MAST to both platforms resulted in notable differences in MAST ratings between High- and Low-MAST conditions. Under High-MAST conditions, Facewise achieved higher scores across all criteria (9/9), while READIT achieved higher scores on 6 out of the 9 criteria. This difference between Facewise and READIT indicates that the type of system and use context matters when applying the MAST checklist. For an image processing and signal detection type system like Facewise, using MAST to evaluate elements like accuracy, source reliability, and user interface clarity may be more straightforward, as reflected in the consistently higher ratings across all criteria in the High-MAST condition. In contrast, READIT as a text-summarization system is riddled with the complexities of natural language processing model outputs and the semiotics of text interpretation. For example, our team was particularly challenged in designing appropriate visualizations for model outputs and explanations that could differentiate between High-MAST and Low-MAST READIT systems. In the end, the MAST ratings for the *Visualization* criterion were not significantly different. Moreover, the lack of statistical significance between the *Uncertainty* and *Customer Relevance* criteria may also point to High-MAST design features that had marginal to no impact. In the Low-MAST system, uncertainty measures

such as *term frequency–inverse document frequency* (*tf-idf)* and cosine similarity scores (see Appendix B) were omitted, unlike in the High-MAST system where they were included. Our results could suggest that these uncertainty scores were either ineffective in conveying uncertainty or were implemented in a manner that limited their usefulness. The same observation applies to the feature of filtering by location and topic, which was incorporated specifically to meet the *Customer Relevance* criterion, and was a differentiating feature between the two system versions. The rating similarities between the systems could indicate that these features did not significantly influence users' perceptions of *Customer Relevance*.

Trust perceptions. Evaluation of the Facewise systems revealed significantly different trust ratings on the Jian et al. (2000) questionnaire, with higher ratings observed for the High-MAST condition. In contrast, the higher trust ratings observed in the High-MAST READIT system compared to its Low-MAST counterpart were not statistically different. This discrepancy may have stemmed from the smaller participant pool evaluating the READIT system, leading to increased standard errors in the observed differences. It was a challenge to recruit participants for READIT, given the relative inaccessibility of remotely-recruited working intelligence analysts relative to the on-site recruited Transportation Security Officers for Facewise, and the general challenge of recruiting subject-matter experts to volunteer their participation in research studies. Additionally, for the trust constructs in the Chancey et al. (2017) questionnaire, the Facewise systems did not exhibit significant differences, despite the Jian et al. (2000) ratings demonstrating otherwise. We speculate that this might be due to the nature of the Jian et al. (2000) items being valenced both negatively and positively whereas for the Chancey et al. (2017) questionnaire, the items are all positively valenced. Prior research has shown that item valence can affect responses in trust measures (Gutzwiller et al., 2019), and that negatively valenced trust items may result in more variable responses compared to positively valenced trust items (Schroeder et al., 2021). More research is needed to investigate why these two instruments measuring the same construct could result in different responses (e.g., Long et al., 2020).

Despite being evaluated by a considerably smaller group of participants, the READIT systems demonstrated a notable difference in the Chancey et al. (2017) trust measure on the Purpose dimension. No other differences were significant in other trust measurements. It is important to note that the "purpose" items in the Chancey et al. (2017) questionnaire are designed to assess participants' belief in the READIT system's ability to assist them in their tasks or missions, even amid uncertainties or perceived errors. This difference in ratings could therefore be linked to the inherent challenges and semiotic nature of text-summarization tasks, as opposed to the more straightforward outcome of signal detection (e.g., face recognition) type tasks. The higher rating for the "purpose" dimension in the High-MAST READIT group suggests that people value the system and its associated features to help summarize documents.

Other perception metrics. Comparative analysis of other perception metrics revealed that High-MAST versions of both platforms generally led to reduced perceived risk and higher perceived benefit and credibility, aligning with our initial hypotheses. However, for READIT the "benefit" ratings, and for Facewise, "credibility" ratings were not significantly different between versions. Notably, these items exhibited greater variability, indicating

divergent perceptions of Facewise's credibility, likely influenced by the model's errors in tasks that may have been easy for our expert participants. In READIT's Low-MAST group, the "benefit" ratings varied more, suggesting that some of the analyst participants still found the system useful with respect to the imagined increased task load that would have come from manually inspecting the raw document data on their own. In terms of "engagement" and "usability", we observed no significant differences between the High- and Low-MAST versions for both platforms. This uniformity in perceived engagement and usability ratings indicates a consistent user experience, and minimal impact of perceived usability or ease of use differences on the evaluation of the systems.

Performance measures. Neither platform showed significant differences in system performance between the High- and Low-MAST versions, whether in face-matching accuracy for Facewise or the report score for READIT. This aligns with prior research suggesting that AI transparency or trustworthiness *alone* does not necessarily result in improved human performance (Palanski & Yammarino, 2011; Schelble et al., 2023). The absence of observed differences in our study could also be attributed to factors such as limited variation in the image database for Facewise, and the use of under-optimized AI algorithms common to both versions. In the case of READIT, despite clarification in the participant onboarding video, there might still have been some confusion among participants on how to interpret the bubble graph topic clusters. We discovered this issue during pilot tests with non-expert participants, who mistook the size of the bubble graph topic clusters for importance rather than topic frequency in the anomaly detection task. We attempted to correct for this in our onboarding video by highlighting how to interpret the bubble graph, but ultimately we did not test to confirm their understanding or use of the bubble graph, which was one of the more salient differences between the High- and Low-MAST versions of READIT. Designing the READIT interface was challenging due to project time constraints and the need to balance the presentation of detailed information with navigational ease on a single browser page, without overly guiding participants to the correct answers. Future research could better refine these AI-DSS testbeds, improve AI performance, task variation, and optimizing the level of information detail in the interfaces.

Association between MAST and perception metrics. Principal Component Analysis (PCA) was conducted to assess whether MAST could accurately capture key constructs from well-established human perception metrics. These metrics, which show significant marginal associations with MAST ratings, include trust measures from Jian et al. (2000) and Chancey et al. (2017), and measures of perceived benefit, credibility, and risk. The primary goal of PCA in this context was to produce comprehensive summaries that capture the majority of variation within these metrics. The analysis showed that for both platforms, the first two eigenvalues accounted for over 80% of the total, suggesting that the first two principal components are sufficient in explaining the majority of variation in the data. In both platforms, the first principal component (PC) uniformly displayed positive loadings for all metrics except risk. This consistent pattern across both platforms indicates that a uniformly weighted average of these metrics, negatively weighted for risk, effectively captures the essential constructs of participant perceptions in the evaluated technologies. Conversely, the second PC showed significant positive loadings exclusively for risk and benefit. This pattern

suggests that participants tend to conduct a risk-benefit analysis, with the pronounced loading on risk indicating a stronger focus on risk assessment when evaluating AI systems.

The regression analysis of the first two PC scores against the MAST ratings revealed significant associations with the scores of the first PC, but not with those of the second. This finding indicates that the MAST ratings predominantly align with the factors captured by the first PC. Because the first PC primarily reflects a uniformly weighted combination of the perception metrics, with an inverse weighting for risk, it can be inferred that MAST ratings are similar to an averaged rating of these metrics. This suggests that MAST effectively captures a broad spectrum of perceptions measured in our study, particularly trust, benefit, and credibility, while inversely accounting for risk. However, the lack of association with the second PC, which focuses more on risk-benefit analysis, implies that MAST may not fully capture the nuances of how participants weighed risks against benefits when evaluating Facewise and READIT. These potential nuances would further speak to the challenge of soliciting input on some of the MAST criteria, input that may vary widely depending on the experience level and perspectives of respondents (Ananny & Crawford, 2018).

Study limitations. The AI-DSSs in this study were intentionally designed to align with either High-MAST or Low-MAST ratings. This methodology might invite criticism because MAST, which required validation, was also employed in designing the experimental manipulations. However, we assert the validity of this approach based on the independence of the raters (i.e., recruited study participants). Study participants who assessed the platforms were not engaged in either the design process or the development of MAST, ensuring that their ratings were not self-serving biased. Further, intentionally aligning the designed features with the MAST checklist was necessary for internal validation of the tool. This designed distinction allowed us to assess whether MAST, as an evaluative tool, could effectively differentiate between technologies with varying MAST alignment levels. Establishing internal validity serves as a foundation toward external validation in collaboration with the broader research community. Additionally, this study sets a benchmark for future applications of MAST as a tool across various technology applications and task contexts.

Although MAST ratings were highly associated with trust, our results do not factor in whether trust or distrust levels were calibrated with system performance. Such an analysis may be possible for Facewise, in which system reliability can be precisely gauged using signal detection metrics. However, trust and distrust calibration is difficult to define for the READIT platform because it does not offer direct recommendations or answers that could be rated as easily. Future studies should consider these different forms of decision support, and how those different forms can affect trust responses (Chiou & Lee, 2023). Finally, although the signals were strong for READIT, we could not reach our desired sample size within our project timeline, due to the challenge of recruiting intelligence analysts. Lastly, this study was focused on the use and validation of MAST specifically; a comprehensive review and comparison of MAST against other similar frameworks would be a valuable exercise, but beyond the scope of this project. Other literature has reviewed similar tools for trust assessment (Alsaid et al., 2023; Kohn et al., 2021), and MAST might be used in conjunction with some of these other tools alongside a work-centered field-based approach (Roth et al., 2021) to achieve a more comprehensively designed and functionally trustworthy system.

7. Conclusion and Future Directions

The primary objective of this study was to establish the utility of the Multisource AI Scorecard Table (MAST) for evaluating the trustworthiness of AI-enabled decision support systems (AI-DSSs). This resulted in an interesting opportunity to evaluate whether the tradecraft standards behind MAST are related to the existing tools developed by the scientific community of trust researchers. We employed a three-step analysis method (ANOVA, SLR, and PCA) to investigate possible connections between MAST and trust, and based on that we conclude there are strong associations between MAST ratings and trust perceptions. Furthermore, by testing both Facewise and READIT with respective subject-matter experts, we also demonstrated the utility of MAST across high-stakes domains, showing that these patterns of associations persist across two different AI applications.

Compared to other frameworks for designing or evaluating AI-DSSs, a benefit of MAST is that it is derived from, and operationalized by, a practitioner community (Blasch et al., 2020). Thus, the underlying principles of MAST are more likely to be "customer relevant" and accepted by the Intelligence Community, while aligning with an empirical and scholarly understanding of trust and credibility that we report here. However, just as high quality analytical reporting may not result in good decision making, it is important to note that high MAST ratings do not necessarily result in improved performance of a human-AI decision system, given the variety of factors that can contribute to this performance, including factors in the task environment, cognitive workload (Sargent et al., 2023), available system features, and task difficulty. Furthermore, it is still possible that high *intended* MAST ratings by a design team may not result in higher perceptual ratings by evaluators. Additional testing should be done with other AI systems, including key factors in the organizational and task environment (Chiou & Lee, 2023), and more formal risk analyses. In-depth exploration of the behavioral data captured during task performance may also shed light on the gap between trust perceptions and trustworthiness.

Acknowledgments

This material is based on work supported by the U. S. Department of Homeland Security under Grant Award Number 17STQAC00001-05-00. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Homeland Security. PS contributed the initial complete draft of the paper. PS, YB, MM contributed to data analysis. PS, NK, MC, YB, SB contributed written sections and response to reviewers. PS, AP, AM, NK, YW, JZ, and YB contributed to design and development of Facewise and READIT. PS, AP, and MC contributed other study materials. PS, AP, MC, YB, NK, AM contributed to data collection. MM, JS, EB contributed to study conception, design, interpretation, and participant recruitment. EC contributed to all aspects of the study. All authors reviewed and approved the final version of this paper.

Appendix A. The Nine MAST Criteria for Facewise

Item Question(s) and Feature Description(s)
Sourcing: How well can the system identify underlying sources and methodologies upon which results are based? **@High-MAST:** The "View Details" page provides the name of image sources and demographical information about the people whose image data were used to train the AI, such as their race and gender. **@Low-MAST:** The system interface does not include the name of image sources and demographical information about the people whose image data were used to train the AI system, such as their race and gender.
Uncertainty: How well can the system indicate and explain the basis for the uncertainties associated with derived results? **@High-MAST:** For each case, the system will display a certainty score from 0%-100% to indicate its confidence about its recommended decision. The system also gives an alert if the uncertainty is too high when you click the "Final Decision" button, depending on your decision. Details about how the system calculates the certainty score are available by clicking on the "More Details" button under every decision. The AI's confidence level is calculated in this manner: first, a metric that signifies the mathematical distance between two image pairs is calculated. Then, the difference between the mathematical distance and a pre-determined (computed during the training and validation stages) threshold is calculated. Finally, the difference is normalized by a factor and the confidence level is calculated using probability measures associated with the standard normal distribution. Thus, the AI's confidence is an indication, based on the predetermined threshold. Confidence levels closer to 100% indicate higher confidence. **@Low-MAST:** For each pair of images, the system only recommends a binary decision (same or different) and does not indicate its confidence in the decision.
Distinguishing: How well can the system clearly distinguish derived results and underlying data? **@High-MAST:** The system can distinguish whether a presented ID is invalid or expired, or if the ID photo may have been digitally altered. An alert message will be automatically shown in these cases by the system. Details about how the system identifies these features in the ID photo are available by hovering over the Crossmark or checkmark icon next to the ID expiration date. **@Low-MAST:** The system cannot distinguish whether a presented ID is invalid or expired, or if the ID photo may have been digitally altered.
Analysis of Alternatives: How well can the system identify and assess plausible alternative results? **@High-MAST:** In the "View Details" page, the system provides dissimilarity and similarity probabilities as alternatives for each pair. The similarity and dissimilarity numbers are directly derived from the AI's confidence level. The higher of the two probabilities is selected to represent the AI's confidence level. The calculation of the similarity and dissimilarity probabilities assumes that the threshold is distributed as standard normal, and that the scaled differences are realizations of a noise-generating process. Both probabilities are calculated using the scaled difference between the distance metric and the threshold. **@Low-MAST:** For each pair of images, the system only gives a decision and does not indicate its confidence in the current decision based on the training and validation stages, nor on probability measures of alternatives associated with the standard normal distribution.
Customer Relevance: How well can the system provide information and insight to users? **@High-MAST:** Besides providing the binary decision of same or different, the confidence level, and ID validation on the main page, the system provides additional details through a "More Details"

button. This includes information and explanations about similarity, dissimilarity, confidence level, and sources of training for AI. To present the information more efficiently, the system will minimize explanations that have already been shown. Conditional alerts when the system's certainty level is low and alerts about individuals who may need additional screening per the protocol are also included as part of the system with the information displayed as detected.
@Low-MAST: Besides providing the binary decision of the same or different and ID expiration date on the main page, the system does not provide any additional details or any conditional alerts.

Logic: How well can the system help the user understand how it derived its results?
@High-MAST: The system bases its final decision by choosing the larger of similarity and dissimilarity probabilities. "More Details" button also provides an explanation and interpretation of how a prediction or classification is made. Conditional alerts when the system's certainty level is low, and alerts about individuals who may need additional screening per the protocol are also included. To detect the authenticity of an ID photo, a second model was trained, tested, and validated on proprietary datasets of anomalous and non-anomalous travel documents, digitally altered and original images. A separate model further performs character recognition to analyze expiration dates on travel documents.
@Low-MAST: The system does not give any information on how its recommendation is determined. It also does not provide any conditional alerts or any information about the authenticity or validity of the ID photo image.

Change: How well can the system help the user understand how derived results on a topic are consistent with or represent a change from previous analysis of the same or similar topic?
@High-MAST: As you interact with the system, by clicking "more details" you will see a report about your agreement with the system, which indicates how often the system has been uncertain about your final decisions.
@Low-MAST: As you interact with the system, the system does not indicate how often it has been uncertain about your final decisions.

Accuracy: How well can the system make the most accurate judgments and assessments possible, based on the information available and known information gaps?
@High-MAST: For each pair of images, the system will display a certainty score from 0%-100% to indicate its confidence about its recommended decision. System's performance according to the training data and more details about how the system calculates the certainty score are available by clicking on the "More Details" button under every decision.
@Low-MAST: For each pair of images, the system only gives a binary decision and does not indicate its confidence in the decision, the system's performance according to the training data, or more details about how the system made the decision.

Visualization: How well can the system incorporate visual information if it will clarify an analytic message and complement or enhance the presentation of data and analysis? Is visual information clear and pertinent to the product's subject?
@High-MAST: The system automatically shows you an enlarged version of a traveler's ID photo and their photo taken at the security checkpoint. These images will be shown side by side. Distinguishing features that played a big role in determining the recommended decision will also be highlighted by clicking the "View Details" button.
@Low-MAST: The system only shows you an enlarged version of a traveler's ID photo and their photo taken at the security checkpoint without any additional visualized explanation about the recommended decision.

Table 2: MAST criteria (Blasch et al., 2020) and Facewise feature descriptions for High-MAST and Low-MAST.

Appendix B: The Nine MAST Criteria for READIT

Item Question(s) and Feature Description(s)
Sourcing: How well can the system identify underlying sources and methodologies upon which results are based? **@High-MAST:** In the documents page, you can see descriptive information about the data used to gather the clusters including basic information and detailed descriptions of the sources. The datasheet for READIT includes information on the clustering model, models for summarization, training data, possible biases, pre-processing of data, and quality of the data used in training to derive results. In the main dashboard view, you can view the data used to derive the cluster either by hovering or clicking on it including the cluster title, number of documents, top terms, and representative documents. The representative documents can be viewed as a summary or raw version. **@Low-MAST:** For any given cluster in the main dashboard view, you can view more details about it by clicking on it. The title of the cluster, number of documents, and summaries of the documents will be displayed in the documents and summaries pane. Only the derived results are shown, not the underlying sources and data used to derive the clusters or summaries.
Uncertainty: How well can the system indicate and explain the basis for the uncertainties associated with derived results? **@High-MAST:** READIT indicates levels of uncertainty with derived results in two ways, as described in the datasheet. First, READIT includes keywords per cluster to show how documents in clusters are related to each other. Keywords are displayed with a term frequency–inverse document frequency (tf-idf) score which measures the certainty the word fits with the cluster. Second, READIT includes similarity scores to assess the similarity between clusters. This score is calculated using cosine similarity to show the certainty that clusters are related to each other. **@Low-MAST:** In the topic similarity visualization, the relationship between two topics is colored from white to dark blue with dark blue indicating a higher certainty the two topics are related. These relationships are not labeled with numbers, neither is it explained how this similarity is calculated.
Distinguishing: How well can the system clearly distinguish derived results and underlying data? **@High-MAST:** For any cluster you can view more details about the data used to derive the cluster either by hovering or clicking on it. The datasheet includes information on the clustering model, models for summarization, training data, underlying assumptions for choice of training data, quality of the data used in training to derive results, possible biases, pre-processing of data, recommended uses and users, and restrictions on use. The datasheet was created with domain expert input. **@Low-MAST:** In clusters, you can view more details about that cluster. The title and summary of representative documents will appear. The raw data used to derive the title and summaries is not displayed. There is no datasheet with information on how these titles or summaries are calculated.
Analysis of Alternatives: How well can the system identify and assess plausible alternative results? **@High-MAST:** In the topic similarity, users initially view the visualization where the topics are ordered alphabetically. By factoring in the similarity score and uncertainties, READIT can reorder the view in this visualization such that highly related topics will appear together to present an alternative view. **@Low-MAST:** READIT is not able to show alternative results when uncertainties in the data warrant them. There is no way to reorder visualizations based on any criteria.
Customer Relevance: How well can the system provide information and insight to users? **@High-MAST:** READIT synthesizes large corpora of documents and produces clusters of similar documents. The topic similarity visualization shows which clusters are most highly related to each other. Users can examine the clusters and their relationships in the topic similarity view for trends for follow-up work. READIT is also able to suggest locations to filter by if the documents contain

multiple locations. Users can also filter all visualizations by topic. There is a topic filtering pane where users can check all, or some topics and the corresponding selected topics will be highlighted in the visualizations.
@Low-MAST: READIT synthesizes documents and produces clusters of similar documents. The topic similarity visualization shows which clusters are most highly related to each other. Users can examine the clusters and their relationships in the topic similarity view for trends for follow-up work.

Logic: How well can the system help the user understand how it derived its results?
@High-MAST: For any cluster you can view more details about the data used to derive the cluster either by hovering or clicking on it. The datasheet includes information on pre-processing of data. READIT includes an option to filter results by location, if location information is detected in the document. To give location options, READIT must consider the location information in the context of the document, and other assumptions about the embedding of the location in the document.
@Low-MAST: In clusters, you can view the title and representative documents in summary form. The titles and summaries are understandable to users. Information on how clusters, titles, and summaries are formed is not included. There is also no information on the pre-processing of data.

Change: How well can the system help the user understand how derived results on a topic are consistent with or represent a change from previous analysis of the same or similar topic?
@High-MAST: In the documents page, READIT includes information on similar searches from other agencies. Similar searches may be based on the average length of the document, number of documents, or number of clusters generated.
@Low-MAST: READIT does not note changes from previous analyses or similar analyses. It also cannot compare current results with those of other agencies which had similar results.

Accuracy: How well can the system make the most accurate judgments and assessments possible, based on the information available and known information gaps?
@High-MAST: The READIT datasheet includes information on system verification and validation methodology, and results from the training data where the system achieved sufficiently high accuracy. To assess the accuracy of READIT, users can view the full documents used in each cluster and compare them against the top terms to independently determine whether the documents match the top terms. Likewise, users can view a summary of the document and compare it against the full version of the document in the documents and summaries view to see if the summary is accurate.
@Low-MAST: READIT does not include information on system verification, validation methodology, or the training of the system. Since underlying sourcing information and raw data are not included in the system, it is difficult to assess whether the topics and summaries are accurate.

Visualization: How well can the system incorporate visual information if it will clarify an analytic message and complement or enhance the presentation of data and analysis? Is visual information clear and pertinent to the product's subject?
@High-MAST: READIT uses three main visualizations to enhance users' understanding of the clusters. First, in the topic overview visualization, clusters are displayed as bubbles where the size of the bubbles can indicate anomalies. Next, READIT also creates and displays a topic similarity visualization to help understand the connections between clusters. Lastly, there is a timeline view in READIT to display clusters on a timeline (if documents contain date information). All visualizations are simple and labeled properly. Users can view more details about the visualizations by clicking on them or hovering over them or filtering all visualizations by cluster using the filtering option.
@Low-MAST: READIT uses two visualizations. The similarity matrix shows the similarity scores between topics. Darker colors indicate more similarity but score values are not shown. The timeline shows the clusters on the timeline. Visualizations contain no interactivity and users are not able to click or hover on items to view more details about the visualizations.

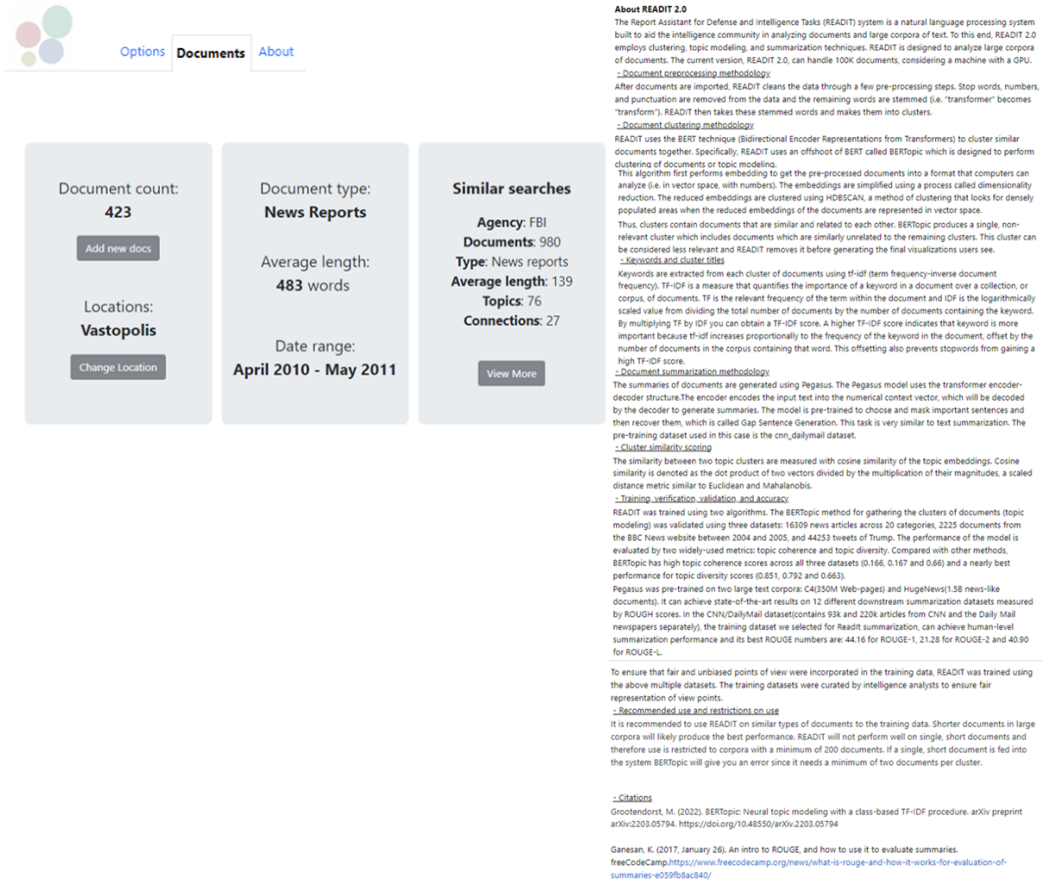
Table 3: MAST criteria (Blasch et al., 2020) and READIT feature descriptions for High-MAST and Low-MAST.

Appendix C: Study Questionnaires

Variables	Reference	Example item(s)	Number of items/ Reverse items	Scale	Facewise Cronbach's Alpha	READIT Cronbach's Alpha
MAST-total	(Blasch et al., 2020)	Sourcing, uncertainty, distinguishing, analysis of alternatives, customer relevance, logic, change, accuracy, and visualization	9/0	9 - 36	.91	.91
Risk	(Weber et al., 2002)	Please indicate how risky you perceive it is to use this system for completing your task well.	1/0	1 - 5	-	-
Benefit	(Weber et al., 2002)	Please indicate how beneficial you perceive it is to use this system for completing your task well.	1/0	1 - 5	-	-
Trust (Jian)	(Jian et al., 2000)	"I can trust the system."; "The system looks deceptive."	12/5	1 - 7	0.90	0.92
Trust (Chancey)	(Chancey et al., 2017)	"I understand how the system will help me perform well. "; "The information the system provides reliably helps me perform well.	15/0	1 - 7	0.96	0.96
Credibility	(Appelman & Sundar, 2016)	"How accurate do the results of the system appear to be?"; "How believable do the results of the system appear to be?"	3/0	1 - 7	0.92	0.92
Engagement	(Schaufeli et al., 2002)	"I was immersed in this research task."; "To me, this research task was challenging."	17/0	1 - 7	0.91	0.93
Usability (SUS)	(Brooke, 1996)	"I felt very confident using the system."; "I thought the system was easy to use."	10/5	1 - 5	0.80	0.88
Task performance	-	Average response time and Accuracy for Facewise and final report gradings for READIT	2/0	0	-	-

Table 4: Dependent and Control Variables.

Appendix D: READIT Documents and About Tabs for the High-MAST Version



(a) Documents tab

(b) About tab

Figure 9: Documents and About tabs in High-MAST READIT platform.

Appendix E: 95% Confidence Interval Figures

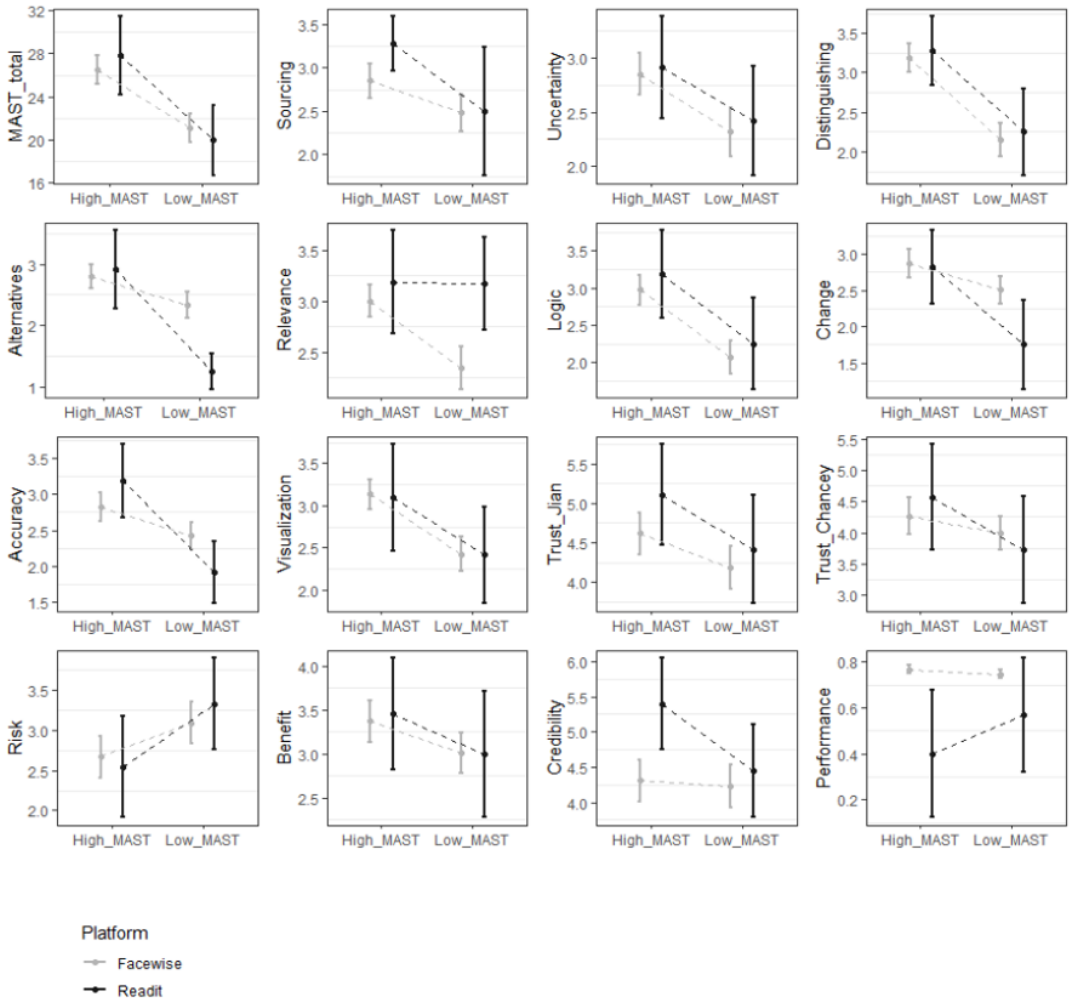


Figure 10: Means with 95% Confidence Intervals for Facewise and READIT across different levels of Low-MAST and High-MAST. We used Facewise for Facewise and Readit for READIT.

Appendix F: Participant demographics for Facewise and READIT

	High-MAST ($n = 73$)	**Low-MAST** ($n = 73$)
Years of experience as a TSO	55% 3 years or less 24% 10 or more years	50% 3 years or less 28% 10 or more years
Highest degree	69% 2-year college or less 26% 4-year college	74% 2-year college or less 21% 4-year college
Volunteer hours in the past 3 months	62% 0 hours	71% 0 hours
Computer habit	58% daily	65% daily
Gaming habit	18% daily 26% never 26% never	30% daily 17% never
Screen hours before study	Mean: 2 hrs. Median: 1.2 hrs.	Mean: 2.2 hrs. Median: 2 hrs.

Table 5: Participant demographics across High-MAST and Low-MAST for Facewise.

	High-MAST ($n = 11$)	**Low-MAST** ($n = 12$)
Age	36% 30 years or less 18% 31-39 years 46% 40 or more years	34% 30 years or less 33% 31-39 years 33% 40 or more years
Gender	73% man 27% woman	50% man 50% woman
Race	82% white	83% white
Years of experience as an IA	18% 2 years or less 27% 3-5 years 55% 6 years or more	25% 2 years or less 17% 3-5 years 58% 6 years or more
Experience with AI-DSS	46% no prior experience	33% no prior experience
Highest degree	27% 4-year college 73% master's	17% 4-year college 66% master's 17% doctorate
Experience with VAST challenge	100% no	100% no
Experience with clustering tools	55% no	33% no
Screen hours before study	Mean: 5.5 hrs. Median: 6 hrs.	Mean: 5 hrs. Median: 5 hrs.

Table 6: Participant demographics across High-MAST and Low-MAST for READIT.

References

Alsaid, A., Li, M., Chiou, E. K., & Lee, J. D. (2023). Measuring trust: A text analysis approach to compare, contrast, and select trust questionnaires. *Frontiers in Psychology*, *14*, 1192020. https://doi.org/10/gs53sq

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism and Mass Communication Quarterly*, *93*(1), 59–79. https://doi.org/10.1177/1077699015606057

Auguie, B., & Antonov, A. (2017). Gridextra: Miscellaneous functions for "grid" graphics. *R package version 2.3*, *2*(1). https://CRAN.R-project.org/package=gridExtra

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications. https://books.google.com/books?id=caxCDwAAQBAJ

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1–4:15.

Blasch, E., Sung, J., & Nguyen, T. (2020). Multisource AI scorecard table for system evaluation [Presented at AAAI FSS-20: Artificial Intelligence in Government and Public Sector, Washington, DC, USA, November 11-12, 2020].

Blasch, E., Sung, J., Nguyen, T., Daniel, C. P., & Mason, A. P. (2019). Artificial intelligence strategies for national security and safety standards [Presented at AAAI FSS-19: Artificial Intelligence in Government and Public Sector, Arlington, Virginia, USA, November 7-9, 2019].

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In *Usability evaluation in industry* (pp. 207–212). CRC Press. https://doi.org/10.1201/9781498710411-35

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. (2017). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors*, *59*(3), 333–345. https://doi.org/10.1177/0018720816682648

Chatila, R., & Havens, J. C. (2019). The ieee global initiative on ethics of autonomous and intelligent systems. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and well-being* (pp. 11–16). Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_2

Chiou, E. K., & Lee, J. D. (2023). Trusting automation: Designing for responsivity and resilience. *Human Factors*, *65*(1), 137–165. https://doi.org/10.1177/00187208211009995

Cooke, N. J., & Durso, F. (2007). *Stories of modern technology failures and cognitive engineering successes*. CRC Press.

Coşkun, M., Uçar, A., Yildirim, O., & Demir, Y. (2017). Face recognition based on convolutional neural network. *2017 International Conference on Modern Electrical and Energy Systems (MEES)*, 376–379. https://doi.org/10.1109/MEES.2017.8248937

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Greene, F., Kudrick, B., & Muse, K. (2014). Human factors engineering at the transportation security administration. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 2255–2259. https://doi.org/10/ghzbt4

Gutzwiller, R. S., Chiou, E. K., Craig, S. D., Lewis, C. M., Lematta, G. J., & Hsiung, C.-P. (2019). Positive bias in the 'trust in automated systems survey'? an examination of the Jian et al. (2000) scale. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 217–221. https://doi.org/10.1177/1071181319631201

Heydari, F., Sheybani, S., & Yoonessi, A. (2023). Iranian emotional face database: Acquisition and validation of a stimulus set of basic facial expressions. *Behavior Research Methods*, *55*(1), 143–150. https://doi.org/10.3758/s13428-022-01812-9

Hill, K. (2022). Wrongfully accused by an algorithm. In K. Martin (Ed.), *Ethics of data and analytics, concepts and cases* (pp. 138–142). Auerbach Publications. https://doi.org/10.1201/9781003278290

Hill, K., & Mac, R. (2023). Thousands of dollars for something I didn't do. *The New York Times*. https://www.nytimes.com/2023/03/31/technology/facial-recognition-false-arrests.html

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

Hope, R. M. (2022). *Rmisc: Ryan miscellaneous* [R package version 1.5.1]. https://cran.r-project.org/package=Rmisc

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Kahle, D., & Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *The R Journal*, *5*(1), 144–161. https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

Kim, N., Cohen, M. C., Ba, Y., Pan, A., Bhatti, S., Salehi, P., Sung, J., Blasch, E., Mancenido, M. V., & Chiou, E. K. (2024). PADTHAI-MM: A principled approach for designing trustable, human-centered AI systems using the MAST methodology. https://doi.org/10.48550/arXiv.2401.13850

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford Publications.

Knop, M., Weber, S., Mueller, M., & Niehaves, B. (2022). Human factors and technological characteristics influencing the interaction of medical professionals with artificial intelligence–enabled clinical decision support systems: Literature review. *JMIR Human Factors*, *9*(1), 1–12. https://doi.org/10.2196/28639

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*. https://doi.org/10/gnrxj9

Kriskovic, M., Dutta, S., & Brewer, J. (2017). From dark patterns to angel patterns: Creating trustworthy user experience. *User Experience, The Magazine of the UXPA*, *16*(5). https://uxpamagazine.org/from-dark-patterns-to-angel-patterns

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

Lilley, M., Currie, A., Pyper, A., & Attwood, S. (2020). Using the ethical OS toolkit to mitigate the risk of unintended consequences. In C. Stephanidis, M. Antona, & S. Ntoa (Eds.), *HCI international 2020* (pp. 77–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-60700-5_10

Lim, J., & Cantor, J. R. (2021). *Privacy impact assessment for the travel document checker automation using facial recognition* (tech. rep. DHS/TSA/PIA-046(c), PIA-046(c)). Department of Homeland Security. https://www.dhs.gov/publication/dhstsapia-046-travel-document-checker-automation-using-facial-recognition

Long, S. K., Sato, T., Millner, N., Loranger, R., Mirabelli, J., Xu, V., & Yamani, Y. (2020). Empirically and theoretically driven scales on automation trust: A multi-level confirmatory factor analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 1829–1832. https://doi.org/10/grxqg7

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, *48*(4), 656–665. https://doi.org/10.1518/001872006779166334

Microsoft. (1995). *The windows interface guidelines—a guide for designing software*. https://books.google.com/books?id=G8peHj787EUC

ODNI. (2015). Intelligence Community Directive 203. https://www.dni.gov/files/documents/ICD/ICD-203_TA_Analytic_Standards_21_Dec_2022.pdf

Palanski, M. E., & Yammarino, F. J. (2011). Impact of behavioral integrity on follower job performance: A three-study examination. *The Leadership Quarterly*, *22*(4), 765–786. https://doi.org/10.1016/j.leaqua.2011.05.014

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51–55. https://doi.org/10.1145/975817.975844

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors*, *50*(3), 511–520. https://doi.org/10.1518/001872008X312198

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In X. Xie, M. W. Jones, & G. K. L. Tam (Eds.), *Proceedings of the british machine vision*

conference (bmvc) (pp. 41.1–41.12). British Machine Vision Association. https://bmva-archive.org.uk/bmvc/2015/toc.html

Phillips-Wren, G. (2012). AI tools in decision making support systems: A review. *International Journal on Artificial Intelligence Tools*, *21*(02), 1–13. https://doi.org/10.1142/S0218213012400052

Qualtrics. (2020). *Qualtrics* [https://www.qualtrics.com]. Provo, UT.

Raykov, T., & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis*. Routledge. https://doi.org/10.4324/9780203809532

Revelle, W. R. (2024). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 2.4.3]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

Ricanek, K., & Tesafaye, T. (2006). Morph: A longitudinal image database of normal adult age-progression. *7th international conference on automatic face and gesture recognition (FGR06)*, 341–345. https://doi.org/10.1109/FGR.2006.78

Roth, E. M., Bisantz, A. M., Wang, X., Kim, T., & Hettinger, A. Z. (2021). A work-centered approach to system user-evaluation. *Journal of Cognitive Engineering and Decision Making*. https://doi.org/10/gmv8qz

Salehi, P., Chiou, E. K., Mancenido, M., Mosallanezhad, A., Cohen, M. C., & Shah, A. (2021). Decision deferral in a human-AI joint face-matching task: Effects on human performance and trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 638–642. https://doi.org/10.1177/1071181321651157

Sargent, R., Walters, B., & Wickens, C. (2023). Meta-analysis qualifying and quantifying the benefits of automation transparency to enhance models of human performance. In M. Kurosu & A. Hashizume (Eds.), *Human-computer interaction* (pp. 243–261). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35596-7_16

SAS Institute Inc. (2023). *Jmp®, version 16*. Cary, NC.

Schaufeli, W. B., Salanova, M., Gonzalez-Roma, V., & Bakker, A. B. (2002). The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. *Journal of Happiness Studies*, *3*, 71–92. https://doi.org/10.1023/A:1015630930326

Schelble, B., Lancaster, C., Duan, W., Mallick, R., McNeese, N., & Lopez, J. (2023). The effect of ai teammate ethicality on trust outcomes and individual performance in human-ai teams. *Proceedings of the 56th Hawaii International Conference on System Sciences*, 322–331. https://doi.org/10.24251/HICSS.2023.040

Schroeder, N. L., Chiou, E. K., & Craig, S. D. (2021). Trust influences perceptions of virtual humans, but not necessarily learning. *Computers & Education*, *160*, 104039–1:15. https://doi.org/10/ghzb34

SEMVAST Project. (2011). IEEE VAST Challenge MC3 - Investigation into Terrorist Activity. https://visualdata.wustl.edu/varepository/benchmarks.php#VAST2011

Sheridan, T. B. (1975). Considerations in modeling the human supervisory controller. *IFAC Proceedings Volumes*, *8*(1, Part 3), 223–228. https://doi.org/10/gfkwdv

Snow, T. (2021). From satisficing to artificing: The evolution of administrative decision-making in the age of the algorithm. *Data & Policy*, *3*, e3–1:19. https://doi.org/10.1017/dap.2020.25

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an empirically developed scale for system trust: Take two. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(19), 1335–1339. https://doi.org/10.1177/154193120805201907

Stanton, B., & Jensen, T. (2021). *Trust and artificial intelligence* (tech. rep.). NIST Interagency/Internal Report (NISTIR) - 8332. https://www.nist.gov/publications/trust-and-artificial-intelligence

Subirana, I., Sanz, H., & Vila, J. (2014). Building bivariate tables: The compareGroups package for R. *Journal of Statistical Software*, *57*(12), 1–16. https://doi.org/10.18637/jss.v057.i12

Sung, J., Nguyen, T., Blasch, E., Daniel, C., Karl, G., & Mason, A. (2019). *AI Phase II National Security Standards For Artificial Intelligence: MAST Checklist* (tech. rep.). 2019 Public-Private Analytic Exchange Program (AEP).

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics* (Vol. 6). Pearson.

Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263–290. https://doi.org/10.1002/bdm.414

Wickham, H., François, R., Henry, L., & Muller, K. (2019). *Dplyr: A grammar of data manipulation* [R package version 0.8. 1].

Yu, Z., Yoon, J. S., Lee, I. K., Venkatesh, P., Park, J., Yu, J., & Park, H. S. (2020). Humbi: A large multiview dataset of human body expressions. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2987–2997. https://doi.org/10.1109/CVPR42600.2020.00306

Zhu, S., Gilbert, M., Chetty, I., & Siddiqui, F. (2022). The 2021 landscape of FDA-approved artificial intelligence and machine learning-enabled medical devices: An analysis of the characteristics and intended use. *International Journal of Medical Informatics*, *165*, 104828–1:7. https://doi.org/10.1016/j.ijmedinf.2022.104828