Wires Crossed in A Digital World: How to Prevent Misalignments in Human and AI Decision Making

Maria Chaparro Osman, Summer Rebensky, Audrey Reinert, Valarie Yerdon, Chris Jenkins, Jianna Logue, Charles Jusko, Gabe Ganberg
Aptima, Inc.
Woburn, MA
mosman@aptima.com, srebensky@aptima.com, areinert@aptima.com, vayerdon@gmail.com, cjenkins@aptima.com, jlogue@aptima.com, cjusko@aptima.com, ganberg@aptima.com

ABSTRACT

A boom in Artificial Intelligence (AI) technology and presence has made AI virtually omnipresent across domains. However, an important aspect of AI adoption is the level of trust and perceived competency of the system by the human (Hancock et al., 2011). For example, when done seamlessly, search engines provide users with results that are germane to their needs, historical interests, and are more naturalistic in interaction; ultimately increasing their perception of the systems competence (Low et al., 2021). Therefore, it is of high importance to ensure that systems are designed in a way that promotes users' trust while providing them with the support that they need as we look to integrate decision-support AI into intelligence, mission planning, and JADC2 applications. This paper presents the design of a novel system intersecting human factors, cognitive modeling, and recommendation AI to explore approaches for collaborative human-AI teaming. Under this effort, a web-based decision support AI provided recommendations for publicly available articles to answer an intelligence analysis Priority Intelligence Requirements (PIR). We conducted a series of usability and system design evaluations that explored (a) information that users consider when making trust judgments, (b) unobtrusive behavioral measures that integrate into cognitive models to predict when trust falls, and (c) trust calibrations when cognitive model predictions did not match user actions—providing the AI an opportunity to build trust by intervening at the right time in the right way. User behavior, impressions, and self-report responses were examined to understand what user behaviors emerge when users perceive a tool to be working collaboratively. Specific guidance on designing recommendation AI that can leverage behaviors and cognitive modeling for naturalistic interaction as well as system calibration techniques to improve a user's perception system competency are discussed.

ABOUT THE AUTHORS

Dr. Maria Chaparro Osman, Associate Scientist, Aptima, Inc., works in the Training, Learning, and Readiness, Division. Her research includes decision making under novel conditions, training and instructional design, performer/learner engagement during complex monitoring tasks in operational and training contexts, and usability of mobile applications. She received a B.S. in Technical Communication and New Media from the University of South Florida and an M.S. in Aviation Human Factors from Florida Tech, and a Ph.D. in Aviation Sciences with a focus in Human Factors at Florida Tech's College of Aeronautics.

Dr. Summer Rebensky, Scientist, Capability Lead, Aptima, Inc. works in the Training, Learning, and Readiness Division. She has a background focusing on human performance, cognition, and training in emerging systems. Dr. Rebensky has previous experience as a research fellow as a part of the Air Force Research Laboratory's Gaming Research Integration for Learning Laboratory (GRILL) conducting research on drone operations and human-agent teaming utilizing game-based technology... Dr. Rebensky received her BA in Psychology, MS in Aviation Human Factors, and PhD in aviation sciences with a focus in Human Factors from Florida Tech.

Dr. Audrey Reinert, Research Engineer, Aptima, Inc. works in the Performance Augmentation Systems Division. Her research focus is on the intersection of human-centered computing, data visualization, and human-machine interaction. Dr. Reinert holds a PhD in Industrial Engineering from Purdue University, a master's degree in Human

Computer Interaction from the Georgia Institute of Technology, and a bachelors' degree in Cognitive Neuropsychology from the University of California, San Diego.

Valarie A. Yerdon, Scientist, her experiences in industry, academia, and government to human-centered engineering initiatives to the modernization and digitization of adaptive and dynamic learning and training with a systems perspective, utilizing machine learning and artificial intelligence systems. She is currently completing her Instructional Systems Designer Graduate CRT at UCF and is a PhD candidate in Human Factors Psychology (2022) at Capitol Technology University, MD. US Citizen.

Mr. Christopher Jenkins, Senior Software Engineer, Aptima Inc. works in the Performance Augmentation Systems Division. His technical interests include .NET, JavaScript, Python, AI/machine learning, microservices, front and back-end web technologies, RESTful web services, scalable cloud infrastructure, and security. At Aptima, he leads and contributes to a variety of projects that focus on using next-generation web technology to create a cutting-edge user experience. Mr. Jenkins holds a BS in Information Technology from the University of Central Florida.

Ms. Jianna Logue, Software Engineer, Aptima, Inc. has experience with programming in C#, C, C++, Java, Angular, and React.js. Her work at Aptima has supported the development and integration of full-stack software applications, with a focus on microservices and their integration with outside systems. Ms. Logue has a BA in Russian language and literature from Georgetown University and an AS in Computer Programming and Analysis from Seminole State College. She is currently working toward a second BS in Computer Science at University of Central Florida.

Mr. Charles Jusko, Associate Software Engineer, Aptima, Inc. works within the Intelligent Performance Analytics Division. He uses his prior experience with Object-oriented Programming in conjunction with his knowledge of software architecture to support projects' software needs. He has knowledge of various programming languages including Python, C++, C#, and JavaScript, and a background not only in software development, but in data science as well. Mr. Jusko received his BS in Computer Science from the University of Akron in Akron, OH

Mr. Gabriel Ganberg, Principal Software Architect, Aptima, Inc. has more than 20 years of experience leading and architecting software projects in the R&D space. Serving as the Lead Architect for Aptima's Research & Engineering group, his focus is championing cutting edge internal R&D programs and maturing early-stage research prototypes into deployable applications. Mr. Ganberg received a BA in Computer Science and Economics from Vassar College in New York.

Wires Crossed in A Digital World: How to Prevent Misalignments in Human and AI Decision Making

Maria Chaparro Osman, Summer Rebensky, Audrey Reinert, Valarie Yerdon, Chris Jenkins, Jianna Logue, Charles Jusko, Gabe Ganberg
Aptima, Inc.
Woburn, MA
mosman@aptima.com, srebensky@aptima.com, areinert@aptima.com, vayerdon@gmail.com, cjenkins@aptima.com, jlogue@aptima.com, cjusko@aptima.com, ganberg@aptima.com

INTRODUCTION

The adoption and implementation of Artificial Intelligence (AI) has rapidly increased over the past 30 years due to advances in both the availability and power of computing hardware and software. The technical definition of AI is a computer system that can perform tasks that normally require human intelligence, such as collecting and parsing information, visual perception, speech recognition, decision-making, and translation between languages. Much of the computing technology we use is powered or enabled by some form of AI. As the military domain continues to shift from action-based tasks to supervisory decision-based tasks using massive amounts of data; recommendation technology is likely to be implemented in DoD systems. In particular, the Joint All Domain Decision Command & Control (JADC2) concept is an approach of interconnected systems and networks connecting all branches of the armed forces. Air, land, sea, cyber, and space will provide data and information from each of these sources in one interconnected system. Just one highly automated and information rich system, can be overwhelming for a user—let alone information from various kinds of systems, locations, and domains. As a result, it is anticipated that JADC2 will require a large amount of predictive analytics, machine learning, and AI to support collecting information, interpreting it, and supporting decisions and actions in the battlespace (DoD, 2022). Similarly, collaborative combat aircraft, or unmanned AI wingman, will require optimizing teammate's strengths, ensuring warfighters can trust and depend on autonomy, and ensuring teaming workloads are manageable for humans (Penney, 2022). The design of human machine teams will require C2 interfaces with high complexity and increasing automation responsibility. Tools to support the C2 operator must be observable, directable, predictable, incorporated from the beginning of training, support a human in adapting to a task, and help manage the state of the environment (Rebensky et al., 2021). Therefore, systems must be designed in an intuitive way that promotes users' trust while providing them with the support needed as decision-support AI is integrated into intelligence, mission planning, and C2.

With high levels of data coming from multiple sources in the JADC2 interface, AI recommendation systems will be necessary to process the large amounts of information to provide users with suggested actions in meaningful and efficient ways. A relevant example of this technology in action in a civilian context is content recommendation systems and "Trending Now" features on streaming services, that tailor to user preferences. Content recommendation systems are one of the most frequently used types of AI. A content recommendation system requires a large volume of data to develop association models that link patterns of user behavior with specific outcomes from which new recommendations can be generate. In simple terms, these probabilistic association models generate recommendations by learning what content categories are linked, for example making the connection that those who search category A types of films tend to also search for category B types of films. Thus, when the next user follows a similar behavior pattern, they will be recommended other item previous users have also searched. A strong use case for these types of content recommendation systems is the intelligence domain, where large volumes of information must be sorted, filtered, and assessed related to an intelligence request. Often, intelligence analysts are tasked with answering information requests that require the application of subjective judgment and sorting a large volume of information from disparate sources, e.g., *how did COVID-19 impact space test programs?* However, an important aspect of AI adoption is the level of trust and perceived competency of the system by the human (Hancock et al., 2011). In intelligence domains, the sources of information must be trustworthy to correctly inform mission commanders of the likely state of the environment. Trust is commonly defined as the willingness of a party to be vulnerable to the actions of another party whether human or agent based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party (de Visser et al., 2019; Mayer

et al., 1995). When trust is low, often users will abuse, misuse, and disuse systems (Parasuraman, 1997). With large amounts of military investment in these systems, it is important that the warfighters build these tools into their workflow to improve efficiencies and lead to more effective decision-making. Active disuse of these systems will lead the warfighter to spend more time to circumvent these decisional aids which will only further hinder the mission in an information dense world. When trust is too high, users can become complacent and trust all recommendations made by a system which can lead to catastrophic outcomes when systems underperform. In the context of this paper, we target specifically competency-related trust— "do I trust the system to carry out the task it is assigned to do?" (de Visser, 2019). Often it is found that inaccurate mental models of what a system is intended to do can cause competency-related trust to break. For example, a user may blame their computer for being unable to print a document but does not understand that their internet connection may be the real reason they are unable to print wirelessly—leading to decreases in trust in their computer regardless of the cause Therefore, it is not necessarily the reliability of a system, but if the system is performing the specific functions in the way the human *perceives* it should. When expectations are conveyed and interactions are seamless, humans are provided with results that are germane to their needs, historical interests, and are more naturalistic in interaction; ultimately increasing their perception of the systems competence (Low et al., 2021).

This paper presents the design of a novel web-based system that was built on a cloud platform that ingests, stores, analyzes and produces semi-structured data such as a Portable Document Format (PDF) – called the Analytic for Federated Data Tool for Human Efficiency: Behavior and Updates Tracker for Learning Expectations and Relevance (ALFRED THE BULTER, hereby referred to as the system). The tests described in this paper intersect human factors, cognitive modeling, and recommendation AI to explore approaches for collaborative human-AI teaming. What AI lacks today is the ability to consider your own individual needs and respond to your behaviors, goals, and mental models—the personalized approach (Dafoe et al., 2021). We aimed to create a recommendation system leveraging cognitive science, user interactions, and intuitive and easy-to-use interface design to create an interactive and dynamic system that will provide different recommendations not only to each user, but also shift over time depending on the individual's needs.

METHODS

Under this effort, a web-based decision support AI provided recommendations for publicly available articles to answer an intelligence analysis Priority Intelligence Requirement (PIR). Of the military domains available in which recommendation systems would benefit, intelligence analysis provided the most agnostic and accessible environment to isolate and explore trust strategies. We conducted a cognitive walkthrough and series of usability and system design evaluations that explored (a) information that users consider when making trust judgments, (b) unobtrusive behavioral measures that integrate into cognitive models to predict when trust falls, and (c) trust calibrations when cognitive model predictions did not match user actions—providing the AI an opportunity to build trust by intervening at the right time in the right way.

Each of these tests were conducted utilizing a web-based tool system built on a cloud-based data science platform for ingesting, storing, analyzing, and producing semi-structured data. The system ingests a wide variety of data types, and then translates them into to a universal document-based format that can be interacted with by the user providing a graphic agnostic interface to assess trust in textual format. This system allowed us to perform a more controlled experimentation on the specific approaches and data while minimizing as many confounding variables as possible. The system also utilized pre-built analytic wrappers for open-source libraries used to support ingesting web-based articles and tracking user behavioral data. Behavioral data were tracked to validate both behaviors indicative of trust and how much these behaviors contribute to a user's trust to provide the system with optimal data to improve the model and impact user's system use. For this effort, the system used an algorithm that extracts keywords from a PIR that are used to search a corpus of documents for relevant news articles. Resulting articles are presented to the

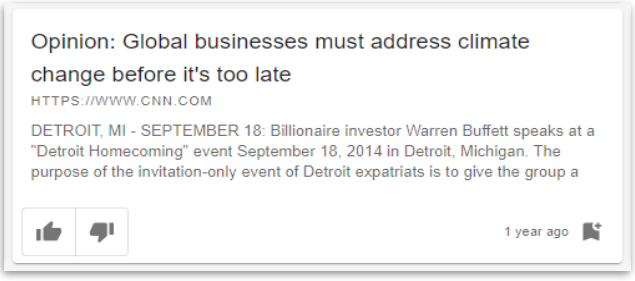

Figure 1. Example of Article Recommendation within the System

user for their analysis and feedback. On the main screen of a PIR, users are presented with 8 recommended articles by the system at any given time. Users are presented articles ranked by their relevancy to the PIR with snippets on information presented on the main screen (e.g., a portion of the title and content, source, and date). Users could select the article and it would provide the full text along with color-coded highlights to draw the reader's attention to relevant keywords and relevant organizations highlighted within the article. A user can additionally choose to (a) click on a thumbs-up icon to save the system's article recommendation, (b) select the thumbs-down icon to reject the recommendation or (c) ignore the recommendation altogether. Users could also select a banner icon to save the article outside of the PIR, similarly to the bookmark function in browsers (see Figure 1 for an example of a recommended article).

Cognitive Walkthrough – Understanding Trust Factors

The goal of the first effort was to perform a cognitive walkthrough to determine the factors that users take into consideration when choosing a recommendation of a relevant resource appearing in a search for more information, prior to decision-making (i.e., what makes you accept an article). The cognitive walkthrough illuminated whether the system (a) extracted relevant information (b) presented users with relevant information, and (c) further understand what factors to integrate into the Inquiry Based Learning (IBL) model. Two individuals (1 female, 1 male) participated in an in-depth cognitive walkthrough. Although the cognitive walkthrough utilized only two individuals, this was performed early in the design of the system and was focused on the quality of decision-making to ensure the full process of decision-making and mental models could be captured within the design of the system. As a result, we decided to perform an in-depth one-hour walkthrough with two individuals to ask detailed questions throughout. Both individuals had no prior experience with the system. The version of the system that was being tested was not yet implemented, therefore we utilized a Microsoft PowerPoint (PPT) mockup that introduced the system's user interface as well as mimicked each decision that could be made to a set of eight articles was used. The mockups that were provided were direct screenshots from the current version of the system with the provided new designs layered on top—exactly as they would be shown in the actual system. In other words, there was a PPT slide that included an image of the actual system presenting what the user would see if they had made that decision for themselves. There was a screenshot of every one of the system's potential paths for all the cards shown. Users were presented with the static system image of what they would see if they had just made a PIR and would point or state what they wanted to "click" next. The researcher would then navigate to the appropriate PPT slide image. For example, if a user said they would select to open an article, there was a slide that showed the user the inside of the selected article.

The cognitive walkthrough was completed in person with two researchers present. One researcher acted as a proctor and navigated to the correct slide depending on the selection the user suggested. The proctor sat next to the individual for the entirety of the session. The second researcher acted as a scribe, recording actions, feedback, and comments made by the individual. Users were presented with four high-quality recommendations and two poor recommendations to understand how decision-making strategies may change dependent on the relevance of the recommendation. Individuals were asked to share any thoughts aloud as they went along, what pieces of recommendations they were considering when making decisions, and any points of confusion. After the cognitive walkthrough portion, individuals were asked interview questions such as: "How would you describe your experience using ALFRED THE BUTLER to help your research?", "What are your thoughts on ALFRED THE BUTLER's recommendations?" and "Could you tell us more about your strategy in deciding whether to use ALFRED THE BUTLER's recommendations?"

The results of the cognitive walkthrough uncovered insights into users' decision-making process and factors that some users consider when reviewing recommendations. The following factors were deemed important when deciding whether to accept or reject an article recommendation in no particular importance order: (a) article date, (b) article title, (c) article source, (d) author, (e) relevance of article title to PIR, (d) presence of keywords in the title, and (e) presence of keywords within the article. The two users tended to open the "high quality" recommended articles to better understand the system or the article's relevance before making their decision. For the "low quality" or irrelevant articles users tended to make the decision to reject the articles before opening them. These findings revealed two main features necessary to design a system that could calibrate trust in a naturalistic way. First, the system must be able to track and identify which factors are important to users and in what order. Second, the system needed to provide insight into its functionality and reasoning for the recommendations it was providing. In other military contexts, such as JADC2, these fundamental findings still apply. Mission commanders in a JADC2 environment will likely also have factors that are relevant or specific to them. Additionally, some insight into the rationale for the recommended actions provided by AI in a JADC2 interface would be necessary.

Design Workshop and Review – Exploring Trust Intervention Techniques

The results of the cognitive walkthrough led to a two-day workshop iterating on designs for feedback features that would address the comments from the users. Research on trust repair strategies has suggested a myriad of approaches such as apologies, explanations, setting up expectations, providing opportunities to provide feedback, explaining failures or limitations within the system, requesting assistance, or providing warnings (de Visser, 2020). A design workshop was held considering these approaches and leveraged current interface feedback designs such as those in Netflix and Facebook and examined how they could be integrated into the current system. Further, the development of feedback features was evaluated via user testing. The goal of the user testing was to evaluate the implemented system updates, specifically after accepting or rejecting an article (Figure 2). In addition, a "Recommend New Articles" button was added. This feature provided the user with a new set of eight articles to allow us to identify when users actively decided to ignore a recommended article and proceed forward to new recommendations (Figure 3). The updates included the addition of two human and AI feedback features. One of the features, the "Tell Us Why" feature allowed users to provide the system feedback relative to which factors identified in the cognitive walkthrough contributed to accepted and rejected recommendations (Figure 4). The listed reasons were: source trustworthiness, date, title relevancy, content relevancy, source familiarity, readability, as well as a qualitative "other" field. A "Why Was This Recommended?", feature was added that explained to the user that the article was presented based on the inclusion of a set of keywords (Figure 5). In conjunction with the evaluation, we wanted to ensure that these feedback mechanisms were intuitive to users. Two individuals volunteered to participate in user testing in which cognitive walkthroughs were conducted. Both individuals had limited experience with the system, the recommendation, and the decision aid system prior to the test. Individuals were asked to provide feedback related to a list of icons for the article recommendation refresh button to identify the icon that was the most intuitive. Users were asked to perform a similar task to the first usability test, wherein they were asked to identify relevant articles for the same PIR from the same set of articles. However, the cards now included all the aforementioned feedback features included in the goals section. Individuals performed semi-structured simulated user testing scenarios including look and feel questions. All interactions and comments by individuals were documented and synthesized in the report of these findings.

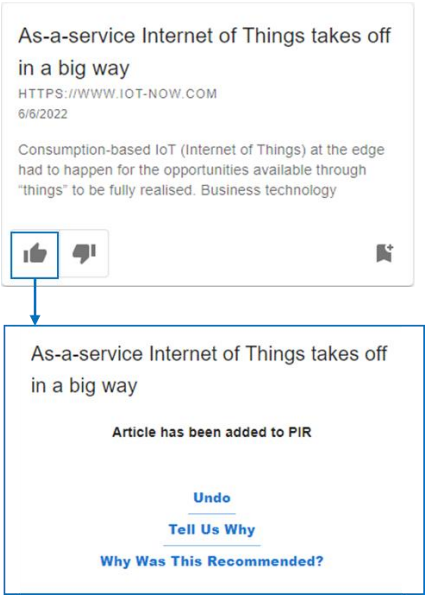

Figure 2. Accepting an Article

Figure 3. Recommend New Article Flow

The path individuals took while using the system's interface and any comments made during their interaction were documented to identify whether they used the feedback features as intended or outside of expectations, whether in levels of use, misuse, or disuse. Additionally, the following questions were asked: "How would you describe your experience using the system to help your research?", "What are your thoughts on the system's recommendations?", "Could you tell us more about your strategy in deciding whether to use article recommendations?" "Do you prefer the inclusion of "I statements" when you are being asked about why you liked or disliked the articles? Why?", and "Are there any feedback options you felt were missing?". The results of the test uncovered that some individuals may have factors outside of the predetermined factors and therefore an "Other" option was added to the "Tell Us Why" feature. For the "Why Was This Recommended?" feature, users had varying opinions of what they believed this functionality would do and how it would alter their search, therefore it was clear more explainability would be needed related to what the system uses to search and how users can change the search.

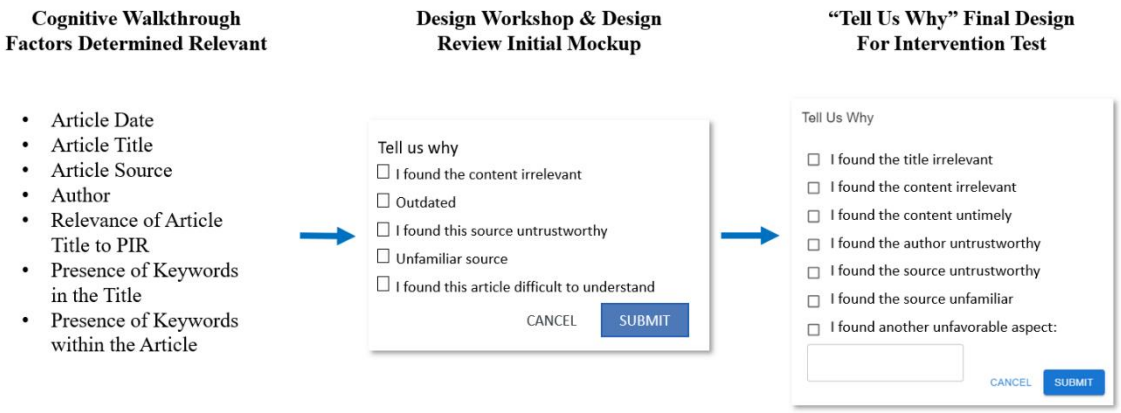

Figure 4. "Tell Us Why" Design Iterations

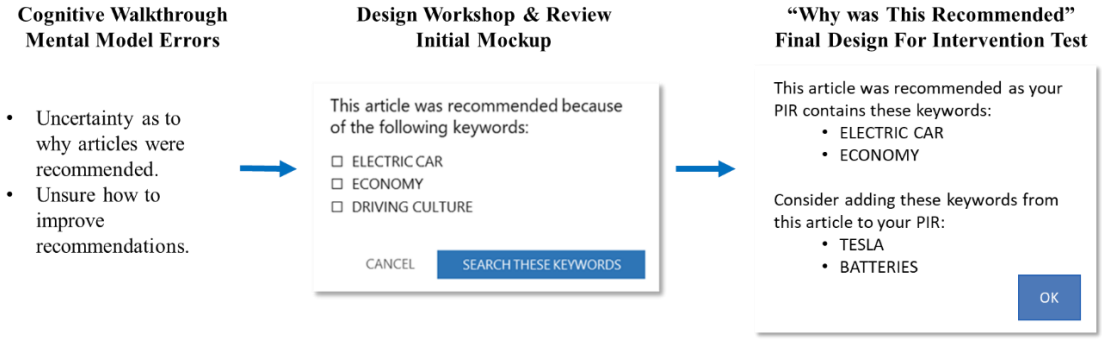

Figure 5. "Why was This Recommended" Design Iterations

Test 1 – Finding Behaviors Indicative of Trust

Concurrently with the design review, a second user evaluation was conducted to identify and examine the types of actions taken by users when interacting with a competent and an incompetent version of the system. The goal of this effort was to confirm the findings of the cognitive walkthrough in an unobtrusive format and to begin to understand when user interventions were needed by examining naturalistic interactions with the model. The competent version of the system would present the user with articles that contained a high density of keywords related to Internet of Things (IoT) devices while the incompetent version presented articles that contained relatively few IoT keywords. In practical terms, the competent system appears to actively assist the user by providing what it determines to be relevant documents while the incompetent system appears to be actively providing irrelevant articles for the effort. We hypothesized that user behaviors would change depending on their perception of the system's competence.

A total of 6 users participated in the competence test where they were exposed to both the incompetent and competent systems. The order was randomized, with three individuals receiving the competent system first and the other three receiving the incompetent system first. Individuals first reviewed a PowerPoint presentation that told them they would be playing the role of an intelligence analyst with the goal of understanding IoT devices as they are becoming more popular and the security threats, they pose by saving articles they felt would sufficiently answer their problem statement. They were told that their problem statement was:

> *What is the projected growth rate of personal IoT devices in the U.S.? Are there large-scale techniques to protect IoT devices within the U.S. from global adversaries?*

They were then introduced to the system and presented a screenshot showing the screen they would see when they accessed the system. This was followed by callouts with screenshots to show users the different icons they would see and their functions. Each individual was told that they would first open an Excel spreadsheet that included four tabs that coincided with each of the surveys they would be filling out. They were to begin by completing the pre-survey and then access the first system link provided to them. Within the system, individuals were told that they could: (a) open articles, (b) accept a recommendation, (c) reject a recommendation, (d) take no action on the recommendation. Individuals were instructed to review 24 articles (by reviewing and refreshing recommendations three times) and then return to the Excel spreadsheet and answer the check-in survey and repeat the process two more times (i.e., 24 interactions and then check in survey) until they reviewed 72 articles. Next individuals were directed to click on the second system link and repeat the process they had gone through in the first link. After competing all interactions and check-in survey, individuals completed a post-evaluation survey and scheduled a debrief time with one of the researchers (see Figure 6). The total test took on average 1.5 hours for the task and 0.5-1 hour for the debrief.

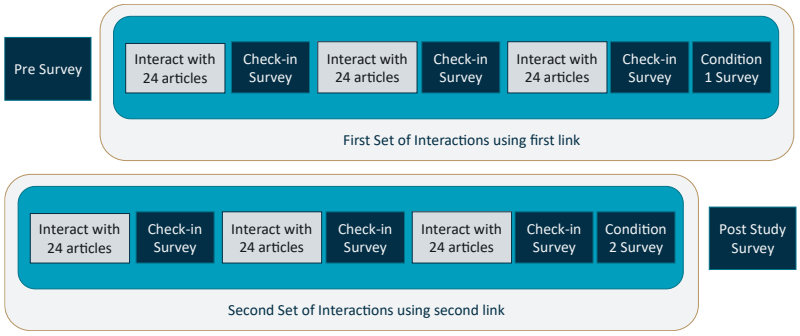

Figure 6. Task and Survey Timeline

Objective Behavioral Measures

To determine behaviors that are indicative of trust, we sought to record every click-based interaction possible within the system for analysis (a full list of behaviors that could be tracked within the system as a web-based platform is presented in Table 1). The system automatically collected a total of 33 user behaviors. However, only 9 total behaviors were predicted to be relevant or indicative of trust based upon the earlier system evaluations described above as these behaviors appeared consistent across users and were representative of decisions or investigative behavior. The following behavior types were hypothesized to be the most insightful for trust: (a) hovering, (b) selections, and (c) inactions. The system only collected one hovering behavior, whether the user hovered over a recommendation. The system collected 6 user selection behaviors: (a) an accepted system recommendation, (b) saving a recommendation to a bookmark-style folder, (c) rejecting a system recommendation, (d) clicking on a recommendation to open it for more information, (e) turning highlighting on or off inside of a recommendation, and (f) refreshing recommendations. The remaining behaviors presented in Table 1 were collected for exploratory purposes.

Table 1. Trackable Behaviors by Function

Trackable Behaviors	
• Selecting, Creating, and Deleting, a Project (and Title)	• Selecting, Saving, and Removing Article
• Selecting, Creating, Deleting, and Editing a Collection (and Title)	• Going outside of and Returning to the system
• Selecting, Creating, and Deleting a PIR (and Title)	• Viewing, Hovering Over, and Refreshing a Recommendation
• Selecting, Editing, and Removing a Highlight	• Adding, Editing, and Deleting a Comment
• Turning Highlight On and Off	• Selecting a URL
• Editing a Summary	• Conducting a Search

Survey measures

Users were asked to complete a series of surveys during the experiment. The pre-experiment survey included demographics, an assessment of Propensity to Trust (Jessup et al., 2019), the users' familiarity with search engines (Körber et al., 2018), and their familiarity with the system and similar platforms. After making 24 decisions, the users would complete a single trust measure Huang et al. (2020); and Körber (2018) including questions like "To what extent do you trust that ALFRED THE BUTLER will recommend useful articles?". Survey responses were given

every 24 articles as trust in a digital system generally stabilizes after approximately 24 interactions (Yu et al., 2019). After completing each condition, the user would assess the system based upon the five elements of analytical rigor e.g. "How confident are you in ALFRED THE BUTLER's ability to provide a diverse set of resources that cover multiple perspectives?", a trust scale, and was asked to provide open-ended statements on reasonings for trust ratings. After the test was completed, the user would complete a System Usability Scale (SUS) and provide additional comments during a brief interview.

Quantitative Response

One of the first variables of interest when assessing if users behaved differently when presented with a competent or an incompetent system was the time it took to accept or reject an article. Our initial assumption was that users would be quicker to reject articles in the incompetent condition than in the competent condition. Upon examination of the response time data for accepted and rejected articles, we obverse no significant difference in average response time (between 17 and 19 seconds regardless of condition). This aligns with the subjective responses during the debrief. Individuals noted that they would take some time reading the cards to parse out why the system felt they were relevant. Most users accurately perceived the system's competency (4 of the 5 total users) during their interactions with the system. Further, the average trust in the system increased or decreased corresponding to the system's competency in both conditions (see Figure 7). It should be noted that trust decreases are greater in magnitude than trust increases. When presented with the competent first condition, user trust was stable after 25 interactions. This aligns with previous research on the topic of user trust in a system. When presented with the incompetent condition first, trust did not have a natural shape and fluctuated up and down. This may indicate that starting off a system that is perceived as incompetent and can have a much more difficult time accurately building trust.

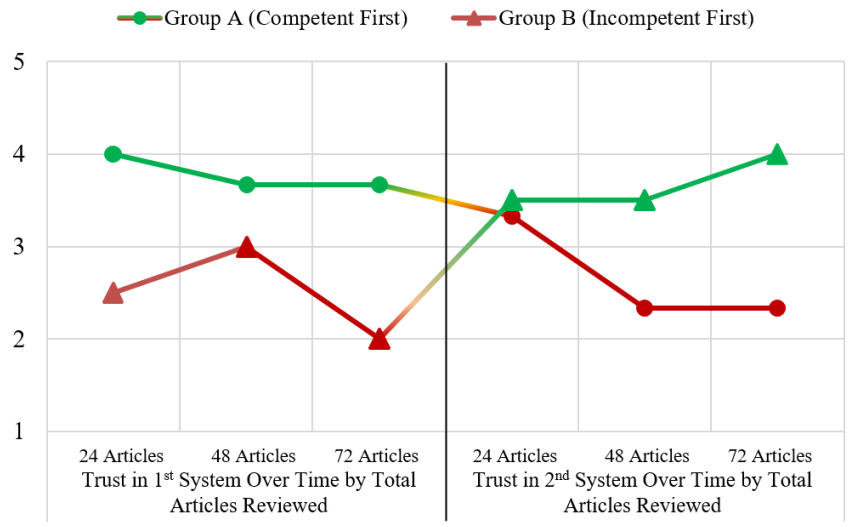

Figure 7. Average User Trust Over Time

Qualitative Reponses

As for the qualitative analysis of the debrief interviews, it was found some behaviors were unique to specific conditions noted by at least 2 responses. In the incompetent condition, users were more likely to reject more articles in hopes to improve the recommendations, skip over more articles, and remove any suggested highlighting by the system. In the competent condition, users were more likely to place faith in the system by accepting articles based on title, reading articles more thoroughly, as well as opening and rejecting less articles in general (see Table 2). Users rated their

Table 2. Qualitative Comments on Behaviors

Action	Incompetent *n*	Competent *n*
Less Time Reviewing Article	4	0
More Rejections	3	0
Removal of Highlighting	2	0
Refreshing Articles	2	0
Decision Based on Title	0	2
Less Rejections	0	2
Less Article Selections	0	2
Took More Time Within Article	0	1
Quick to Make a Decision	2	1
Less System Feedback	1	0
Verifying Article Authors Externally	1	0

Note. For the "Quick to Make a Decision" action, in the incompetent the decision was specifically a rejecting article while in the competent it was accepting.

perceptions on the system's ability to meet five key aspects of analytical rigor (Bruni et al., 2007). It was found that the competent version of the system was rated higher on all five dimensions. Particularly for the dimension of the system's ability to provide relevant recommendations (see Figure 8). Although we purposefully constructed versions of the system that provided relevant recommendations, and those that did not, these findings support that behaviors and perceptions do appear to differ if the human *perceives* the system to be lower competence. In JADC2 environments, computer-based interactions would also be necessary, with changes in behavior if the system is perceived to be making competent recommendations or not. Although the interactions leveraged in JADC2 would be different than those used in our tests, a similar structure and analysis could be used to understand when trust is low or over-trust may be occurring.

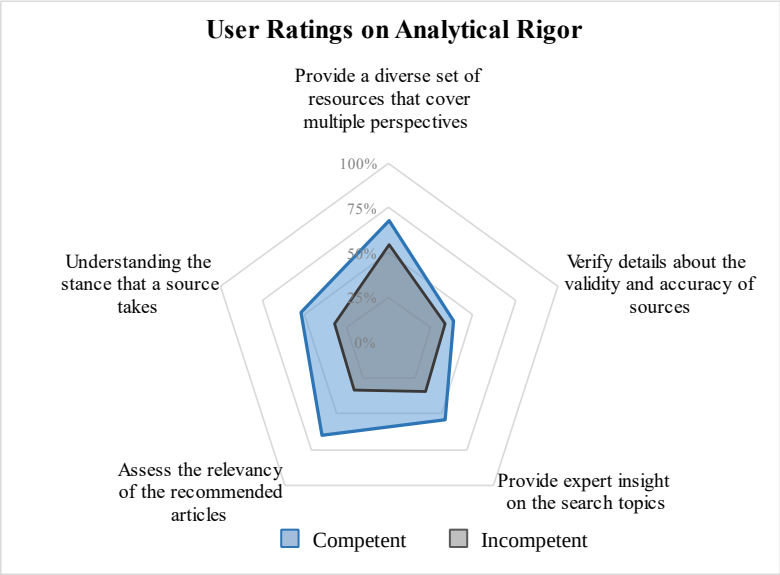

Figure 8. User Ratings of Analytical Rigor by Dimensions

Test 4 – Implementing the Trust Calibration Strategies

The goal of the fourth and final usability evaluation was to evaluate the system's ability to calibrate trust. This test utilized the findings from the previous tests as well as an Instance-Based Learning (IBL) model, to provide trust calibrations at the right time, in the right way. In other words, feedback may not always be needed if the system is already providing recommendations relevant to the user's goals. However, when the system is not providing relevant articles, user feedback will be prompted. Similarly, humans can detect when a conversation has gone off the rails, and that the parties are talking about two different topics. An intervention provides an opportunity to return the conversation to the intended purpose. To achieve a naturalistic intervention method, predictive modeling based on the attributes identified in the cognitive walkthrough was leveraged.

Human-Like Calibration Methods

The IBL model was developed by Carnegie Mellon University's (CMU) dynamic decision-making lab. Under the hood, articles are enriched by the IBL model to include a probability and confidence score of receiving an accept, reject, or ignore on the recommendation by the user and ranked by relevance to the PIR. For example, the model could say that an article has a score of 0.5 for a thumbs up, 0.2 for a thumbs down and a 0.3 for an ignore. Meaning, the model predicts a 50% chance the user will accept the recommendation, a 20% chance the user will reject the recommendation, and a 30% chance they will ignore the recommendation. The system then saves this information to use for future recommendations that are generated. Once the user provides feedback, the events are relayed back to the IBL model to enhance future recommendation predictions. The IBL model includes the factors identified in the cognitive walkthrough and therefore can adapt over time to the factors that users are prioritizing. The system communicated with the model via REST API functionality. When a user first selects a PIR, a list of recommended articles will be generated. Each of these recommended articles will be sent to the model at this time. The model then generates and returns to the system a confidence score for the 3 main actions a user may ultimately perform on each article.

Once the user acts on an article, that is they accept, reject, or ignore an article, the action is sent to the model. This allows the model to learn if it was correct or incorrect in its prediction(s) and adjust accordingly. When predictions from the IBL model did not align with the user's actions, the "Tell Us Why" feature was automatically prompted. This ensured avoiding a "boy-cried-wolf" scenario. The system would only prompt the user for feedback when it was clear that the recommendations it was providing did not align with user's mental models. Those factor ratings within the

"Tell Us Why" were sent back to the model to improve the weighting of the factors to the user's specific factor importance breakdown. The system sends both the action feedback and the "Tell Us Why" feedback to the model in real-time, as soon as the feedback is generated, to help the model refine itself and thus give more accurate predictions for the next article recommendations that the system generates. Lastly the "Why Was This Recommend" prompt was expanded to present the user to keywords present within the PIR as well as considerations for adding new keywords to improve the relevance of the article for the user. These prompts would dynamically change the keywords that were utilized in the user's search within the PIR. These two features together would lead each user to completely different recommendations as well as interventions at different times depending on how important the factor was to them. Users could still investigate "Tell Us Why" and "Why Was This Recommended" features at will after reviewing each recommendation, however, the prompts were automatically prompted when IBL predictions were off.

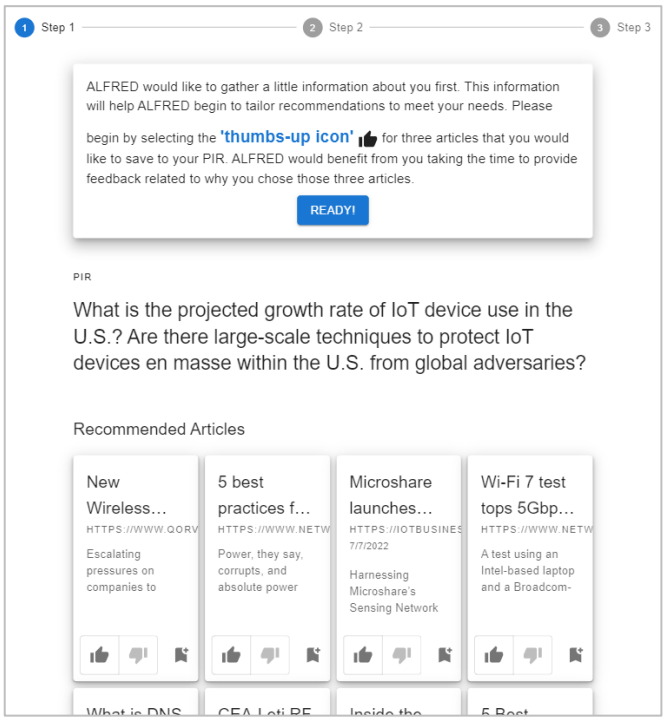

Figure 9. Test Set Interface

To provide the system with initial data about the user, we examine the use of a test set to obtain users' initial trust levels for the IBL model. A dialog is displayed immediately after the user creates a new project (i.e., PIR) when the project is a clean slate—no previous feedback from the user has been received yet. The user was presented with a set of eight articles. The user only had the option of the specified feedback type being solicited. For example, when prompted to accept three articles, the reject recommendation buttons are disabled. This allows the user to focus on the task at hand (i.e., providing feedback of the solicited type) and ensures both kinds of feedback have been received. The number of feedback inputs provided is tracked; when they reach the target number required for that step, the test set dialog automatically moves on to the next step. This initial set, designated as a "test set" was meant to calibrate the system's recommendations. Once users review the rest set and soliciting feedback has been completed, a 'Save' button is enabled that allows the user to close the dialog and return the main project (i.e., PIR) view.

Users were recruited on a volunteer basis and interacted with the system's integrated calibration strategies. Users interacted with a total of 13 sets of articles (or 104 articles). The same types of measures were given as Test 1 with the difference that a few trust questions were given after every set to capture their current trust level in the system's ability to provide useful recommendations, if their search behavior changed at all, and whether they had left for any extended period of time. The same post-surveys were given as Test 1. Data collection for test 2 is ongoing with promising results to guide naturalistic human-machine teams to ensure an accurate mental model between the team members.

CONCLUSION

The following conclusions can be drawn from the studies listed above thus far. First, individual end users will adopt different selection criteria depending on their personal preferences, biases and subject matter expertise when choosing to accept or reject a recommendation. These individual preferences can be quantified as personas (see Reinert et al., 2023). Users will explore and adopt a series of strategies to improve their experience with an AI-enabled system. Like buying a car, some people may prioritize features and gas mileage over color and year. A tailored recommender should be able to consider the needs and goals of the user in front of them. Although the system aimed to explore the use case of intelligence analysts through publicly available articles, similar approaches could easily be used in mission planning and JADC2 (Rebensky et al., 2023). In other military domains, these factors may instead be the likelihood of mission success, the number of backup plans available, or the resources used. Each mission commander may have different priorities of those factors but will still require a decision-aid system that can intervene and shift the recommendations

with as minimal detractions from the mission as possible. Under this effort, we explore some potential methods for designing individualized decision-aid systems. However, the effectiveness of these strategies is dependent on how transparent the system is and the user's mental models and its ability to capture those behaviors.

The current results conform with prior research that suggests that user trust in a system stabilizes after 25 interactions. After this point, user behavior tends to conform to an expected pattern. Sometimes, the behavior patterns for high-trust and low-trust individuals will appear to be functionally similar. For example, both groups of users may make decisions without opening the card. When a system is perceived to be competent, this is because they place faith in the system's recommendations. When a system is perceived to be incompetent, this is because they no longer wish to invest any resources in recommendations that have not proven fruitful in the past. Some users adopted new strategies as they went along in attempts to improve the recommendations they were given (before the dynamic system in Test 4). Therefore, the validity of a user's mental models may be questionable, thus must be given transparent and explainable rationales for what each function does in improving recommendations. Otherwise, actions users take on their own to try and meet their needs in the system may lead to further decreases in trust.

In future research, we aim to model an ideal intelligence analyst based on the 5 elements of analytical rigor (Bruni et al., 2007). Although the system can adapt to the user's needs—the user may not always be right. We hope to extend this model to intervene when the user's search has fallen short of analytical rigor and prompt the user to diversify their search. Users may also expect the system to accomplish more than it is capable of so exploring trust-dampening techniques will also be key. Both the AI and the human prompting their teammate to improve their efforts will be the most team-like approach for collaborative human-AI decision-making. Additionally, we hope to explore including more difficult to capture metrics unobtrusively such as how long someone spent reviewing an article.

To align user expectations with the system's capabilities, there will be times when an intervention is necessary. These interventions should be as naturalistic as possible and mimic interpersonal conversations where humans ask for clarification. These requests wouldn't need to happen all the time, as that would be disruptive, just when indicators show something is amiss. Users may not be well versed in how recommendation systems fully work, so instead of offering transparency, translucency when there appears to be doubt in the system can give users understandable insight into what the system is attempting to do. In Test 2 we allowed users to still explore the features that provided explainability at will so users can take a proactive approach. Like the desire to improve individualized training through AI recommendations and adaptive training systems, operational settings could also benefit from adaptive decision-support systems that evolve with the state of the environment and the decision-maker at hand. Future work may explore extending the approaches and findings here to other military and civilian relevant decision-making contexts.

ACKNOWLEDGEMENT

This material is based upon work supported by the United States Air Force under Contract No. FA8650-22-C-6421. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force. We would like to thank our engineering team members who created the functionality within ALFRED THE BUTLER to support this work Gabe Ganberg, Chris Jenkins, Jianna Logue, and Henry O'Conner as well as our partners at Carnegie Mellon University including Cleotilde Gonzalez, Don Morrison, Baptiste Prebot, and Erin Bugbee.

REFERENCES

Bruni, S., Marquez, J. J., Brzezinski, A., Nehme, C., Boussemart, Y., (2007). Introducing a human-autonomation collaboration taxonomy (HACT) in command and control decision-support systems. International Command and Control Research and Technology Symposium. 1-13.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., Graepal, T., (2021). Cooperative AI: machines must learn to find common ground. *Nature, 593*. 33-36.

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.

Department of Defense. (2022). Summary of the joint all domain command & control (JADC2) strategy [Report].

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. Human Factors: The Journal of the Human Factors and Ergonomics Society, 53(5), 517–527. https://doi.org/10.1177/0018720811417254

Huang, L., Cummings, M. L., & Ono, M. (2020). A Mixed Analysis of Influencing Factors for Trust in a Risk-Aware Autonomy. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64(1), 102–106. https://doi.org/10.1177/1071181320641027

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In *Virtual, Augmented and Mixed Reality. Applications and Case Studies: 11th International Conference, VAMR 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21* (pp. 476-489). Springer International Publishing.

Körber, M. (2018). Theoretical considerations and development of a questionnaire to measure trust in automation. *Computer Security & Reliability*. Triennial Congress of teh IEAF, Florence. https://doi.org/10.31234/osf.io/nfc45

Low, M. P., Cham, T. H., Chang, Y. S., & Lim, X. J. (2021). Advancing on weighted PLS-SEM in examining the trust-based recommendation system in pioneering product promotion effectiveness. Quality & Quantity, 1-30.

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Penney, H. R. (2022). Five imperatives for developing collaborative combat aircraft for teaming operations. *Michell Institute for Aerospace Studies.*

Rebensky, S., Chaparro-Osman, M., Yerdon, V., Reinert, A., & Nguyen, D. (2023). On the same cognitive wavelength: Reaching calibrated trust through cognitive modeling and adaptive human-agent systems. *Human Systems Conference 2023,* Arlington, VA., USA.

Reinert, A., Prebot, B., Rebensky, S., Morrison, D., Yerdon, V, Chaparro Osman, M., Nguyen, D., Gonzalez, C. (July 2023). Using Cognitive Models to Develop Synthetic Digital Twins of Known User Personas. AHFE 2023, San Francisco, CA, USA

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