Trust Exercises and Automation Transparency: The Big Fish

Amanda J. H. Bond, Lauren Glenister
Soar Technology, Inc.
Orlando, FL
amanda.bond@soartech.com
lauren.glenister@soartech.com

Emily Anania, John Killilea, Beth F. Wheeler Atkinson
Naval Air Warfare Center Training Systems Division
Orlando, FL
john.killilea@navy.mil, emily.c.anania1@navy.mil,
beth.atkinson@navy.mil

Brian Stensrud
Soar Technology, Inc.
Orlando, FL
stensrud@soartech.com

Jacqueline McSorley
Embry-Riddle Aeronautical University
Daytona Beach, FL
mcsorlej@my.erau.edu

ABSTRACT

Team-based platforms, such as the U.S. Navy's P-8A aircraft, perform missions through interaction among multiple specialized individuals requiring clear and concise communication. Integrated autonomous agents offer a technical solution to allow students to perform aircrew tasks within an individual simulated training environment without affecting instructor workload associated with role-playing. While the goal of these autonomous agents is to track mission phases and provide verbal reports as human operators do, maximizing human-automation teaming dynamics is required. Automation transparency – the level to which autonomous agents provide human operators insight into decisions, reasoning, status, and outcomes through verbal communication – impacts trust in automation, which can impact workload and task performance. To maximize the benefit of autonomous agents, it is critical to ensure an appropriate level of automation transparency to support instructor situational awareness and accurate assessment without over-burdening the instructor. This paper outlines a study, conducted from January-March 2020, to examine the effects of automation transparency, specifically the amount and types of information needed to maintain situational awareness. For this study, the P-8A domain was represented by a primary task of fishing with drone teammates by attempting to find and catch specific fish with varied levels of information from the drones about the environment. During sessions, a total of 45 participants monitored drones that provided information about the area and direct drones to move and collect specific fish. Participants were restricted to verbal communication with the drones to represent aircrew coordination, using either pre-defined verbal commands or user-generated verbal commands to investigate the impact of verbal communication types with autonomous systems on task performance and workload. Performance, situational awareness, workload, and automation trust were assessed across three levels of automation transparency (low, medium, and high). Results of this study indicated that while fishing performance and situational awareness were not impacted by Transparency Level, workload was negatively impacted by Transparency Level with participants becoming overwhelmed as transparency increased.

ABOUT THE AUTHORS

Amanda Bond is a Lead Scientist at Soar Technology, Inc. and a PhD candidate at the University of Central Florida in Modeling and Simulation / Human Systems. Her background includes twenty years in Human Factors, specializing in human cognition and learning, including game-based training. She is also the deputy lead for the Serious Games Showcase & Challenge for I/ITSEC 2021.

Lauren Glenister is a Research Associate at Soar Technology Inc. She is currently pursuing a Doctorate in Modeling and Simulation at University of Central Florida. She has earned both a Master of Science in Software Engineering and Bachelor of Science in Human Factors Psychology from Embry-Riddle Aeronautical University, Daytona Beach.

Dr. Emily C. Anania is a Research Psychologist at Naval Air Warfare Center Training Systems Division (NAWCTSD) in the Basic and Applied Training and Technology for Learning and Evaluation (BATTLE) Laboratory. Her PhD is in Human Factors from Embry-Riddle Aeronautical University (ERAU). Her research interests include human-automation interaction, human factors analyses of training systems, and aviation human factors.

Dr. John P. Killilea is a Research Scientist at NAWCTSD in the BATTLE Laboratory. He holds a PhD in Modeling & Simulation (M&S), with a special focus on training, from the University of Central Florida. His research interests include training and interacting with synthetic agents, automated speech recognition, manned-unmanned teaming, and maritime training tactics and trend analysis.

Dr. Brian S. Stensrud joined SoarTech after completing his Ph.D. in Computer Engineering from University of Central Florida in 2005. During his tenure he has served as a technical contributor, researcher, customer liaison and principal investigator. He has acquired and led efforts funded out of each of the primary DoD services, including DARPA, ONR, NAWCTSD, and AFRL. Additionally, he has served in a technical lead role in the development of several artificial-intelligence based platforms and toolsets, and has personally contributed to the development of human behavior models and intelligent systems for use within simulations, serious games, intelligent user interfaces, and robotic platforms. Dr. Stensrud's work was recognized with the Army SBIR achievement awards for excellent performance in 2010 and 2012.

Beth F. Wheeler Atkinson is a Senior Research Psychologist at NAWCTSD, a NAVAIR Associate Fellow, and the Lab Lead of the BATTLE Laboratory. She has led several research and development efforts to investigate capability enhancements for training and operational environments, and has successfully transitioned a post-mission reporting and trend analysis tool that leverages automated performance measurement technology. Her research interests include instructional technologies (e.g., performance measurement, post-mission reporting/review), Human Computer Interaction (HCI)/user interface design and analysis, and aviation safety training and operations. She holds an MA in Psychology, Applied Experimental Concentration, from the University of West Florida (UWF).

Jacqueline McSorley is a doctoral candidate and graduate research assistant in the Human Factors department at Embry-Riddle Aeronautical University. She held an internship at NAWCTSD in the BATTLE laboratory during the summer of 2020. Her current research includes gamification of aviation weather training and assessment of pilots' preflight weather briefing habits.

Trust Exercises and Automation Transparency: The Big Fish

Amanda J. H. Bond, Lauren Glenister
Soar Technology, Inc.
Orlando, FL
amanda.bond@soartech.com
lauren.glenister@soartech.com

Emily Anania, John Killilea, Beth F. Wheeler Atkinson
Naval Air Warfare Center Training Systems Division
Orlando, FL
john.killilea@navy.mil, emily.c.anania1@navy.mil,
beth.atkinson@navy.mil

Brian Stensrud
Soar Technology, Inc.
Orlando, FL
stensrud@soartech.com

Jacqueline McSorley
Embry Riddle Aeronautical University
Daytona Beach, FL
mcsorlej@my.erau.edu

INTRODUCTION

The U.S. Navy's P-8A aircraft perform patrol and reconnaissance activities requiring the gathering and analysis of data from several sources. These missions are carried out with a crew comprising multiple people including acoustic warfare operators (AWO), electronic warfare operators (EWO), a tactical coordinator (TACCO), co-TACCOs (Co-TAC), and the flight crew. The P-8's missions require the crew to perform a concert of collaborative actions that must be communicated clearly and concisely in order to meet mission objectives. The TACCO leads the crew and receives and requests a multitude of verbal communications from both inside and outside the aircraft. The TACCO must simultaneously direct the flight crew on where they need to position the aircraft as well as make tactical decisions to execute their assigned mission based on the verbal communications he or she receives.

The complicated and workload-heavy role of the TACCO is therefore trained in two ways: via collective whole-task trainers, which include the entire P-8A mission team and several instructors, and via individual part-task trainers with one instructor and one or more TACCO trainees, each at individual stations. In the whole-task trainer, the entire mission crew is included, with multiple instructors directing the scenario, following and assessing trainee performance, role-playing additional entities, and moving the simulation forward to mimic the mission. This training is understandably challenging to orchestrate due to the manpower required; however, it is extremely valuable training wherein the crew can practice their coordination of mission objectives with reasonable task, workflow, communication, and workload fidelity. Unlike in the whole-task trainer, part-task trainers focus on individual training for aircrew like the TACCO. In this environment, the instructor role-plays to feed data that other positions or entities would provide while also running the simulation and assessing the student TACCO(s). A workload burden is placed on instructors to keep a mental model of scenario progress, role-play well enough to challenge and support the trainees, manage the nuances of the simulation environment, and evaluate trainee performance to provide feedback.

In order to increase training fidelity and opportunities, while also supporting the instructional paradigm in both the part-task trainer and in the whole task trainer, the Navy has investigated a demonstration system of autonomous agents that can perform the tasks of the AWO, EWO, flight, and Co-TAC within the simulated training environment. The agents can track the various mission phases and work in concert with the verbal commands of the TACCO, just as human operators do. In order to maximize the benefit of these autonomous agents for the instructor, it is critical to ensure effective human-automation teaming. This includes ensuring that the automation has the appropriate level of transparency, facilitates proper situational awareness, mitigates instructor workload, and enables accurate student performance assessment. Thus, the authors conducted a study that focused on the effects of automation transparency and verbal communication with automation in order to identify the amount and types of information that are sufficient to enable P-8A instructors to maintain situational awareness.

The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

Effective Human-Automation Teaming

Effective human-automation teaming relies on aspects of the automation, aspects of the human(s) involved, and aspects of the environment in which they are interacting. Parasuraman and Riley (1997) succinctly discuss the potential pitfalls of improper human-automation teaming – misuse, disuse, and abuse. If automation does not function the way a user expects (e.g., confusing design, too many false alarms) it is likely that the operator will disuse, or underutilize, the system. If operators do not understand the limitations of the automation, it is possible that they will misuse and have an overreliance in the system. Parasuraman and Riley (1997) also note the ability for abuse: sometimes developers or managers will automate functions without understanding the effects on human-automation performance. This can result in poorer overall performance than an operator acting without automation. Therefore, it is important to understand the proper ways to design and develop automation to support human use and to maximize performance. To this end, much research has investigated ways that automated systems can be designed for appropriate trust and reliance (Lee & See, 2004) and ideal communication behaviors for human-automation teaming (Demir et al., 2016). Research has also begun to investigate human-agent teaming in order to better understand how to integrate synthetic agents into teams (e.g., McNeese et al., 2018). Though many factors affect human-automation teaming, automation transparency has been shown to have an impact, given its importance in communication between the human and automation and its ability to impact trust (e.g., Yang et al., 2017).

Automation Transparency

Prior research suggests that a system's clarity—its transparency in action and intent—impacts trust in automation (Christoffersen & Woods, 2002; Sarter et al., 1997). Automation transparency—the level to which an autonomous agent or process provides insight into the agent's decisions, reasoning, status, and outcomes—can mitigate trust issues (Chen et al., 2014). According to a recent driving simulation study, automation transparency may enhance users' trust in automation and situation awareness (Kunze et al., 2018). Similarly, according to Helldin (2014), automation transparency positively impacted performance and trust; however, it may have negative effects on workload and decision-making time.

Under investigation in this study was the effect of automation transparency on instructor workload. Although it has been shown that conveying an intelligent agent's intent is beneficial, the amount of information provided and its effect on user performance and self-report measures was examined to try to identify potential effects on dependent variables – situation awareness and workload.

Situation Awareness

Situation awareness (SA) is described as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1995). Endsley proposed three levels of SA by which to assess users' perception of elements, comprehension of their meaning, and projection of future status. In addition, these three levels have been reworked into a SA-based Agent Transparency model that includes transparency with situation awareness by providing insight on an agent's task parameters, logic, and predicted outcomes. (Chen et al., 2014). P-8A instructors must have sufficient proficiency in applying these three major components of SA (perception, comprehension, and projection) within the context of these intelligent agents in order to establish and maintain an accurate picture of the training environment. Adding autonomous agents to the training environment should enable the instructor to focus on the instruction itself rather than on providing scenario cues to students. However, given that the instructor is not as involved in the control and role-playing of some entities, there is the potential for decreased instructor situation awareness. In addition, though the agents will hopefully take on tasking previously allocated to instructors, the instructors will also need to monitor the agents. This has the potential to increase instructor workload, depending on the usability of the automation, and ways that the instructor has to interact with, or monitor, the agent.

Workload

The perceived workload of a task – the physical and/or mental resource "cost" of performing a task (Hart & Staveland, 1988) can dramatically impact task performance. With the investigation of the transparency levels and impact on situational awareness is the concern that the level and nature of automation can be adding to instructors' workload, potentially negatively impacting the ability of the instructor to attend to the trainee. The goal of automation for P-8A

instructors is to maintain or decrease workload while maintaining or increasing situational awareness and training performance measurement efficiency, rather than force the instructor to monitor the automation as well as the student(s). Prior research also indicates that the level of automation provided within a system, particularly cognitive systems, can impact both workload and situational awareness (Endsley & Kaber, 1999). Level of automation can be categorized based on whether the human or the automation performs the task. Related to level of automation, is how much information – if any – the automation provides to the human in the loop (transparency).

CURRENT STUDY

This study examines how automation transparency levels and communication types impact workload, situation awareness, and performance within a game-based approximation of the P-8 task.

Generalized Domain – Drone Fishing

For this study, some of the major decision making activities within the P-8 domain were replaced within a fictional drone-based fishing game in order to investigate general automation principles from a larger population of participants that would have external validity to the P-8 community. This gamified task was created by developing and employing a task breakdown of the P-8A Anti-Submarine Warfare (ASW) task as a model for game design. The key players, types of information being reported, the verbal communication method, and the types of cognitive decisions were assessed to approximate the TACCO job. Subject matter experts provided consultation on the game design to ensure similarity to the P-8 ASW task. In this game, the participants managed five drones that provide information about their fishing zones and were assigned a timed mission to catch specific quantities and types of fish. The participant verbally communicated with two drones, named "Fish" and "Assault" drones, and commanded them to use their resources either bait (Fish drone) or flashbangs and nets (Assault drone) to complete the mission of catching fish. The other three drones – "Radar", "Sonar", and "Imaging" - were automated and provided different reports to communicate the fish characteristics (size, pattern, and movement). The automated drones did not take commands from the participant. To add additional workload, a sea monster called "The Kraken" was inserted randomly across the game area. The Kraken was described as having eaten a jamming device, making the automated drones unable to see in that area. Additionally, the Fish and Assault drones could not employ payloads in that area, mimicking another aircraft or surface / subsurface platform in the area.

Levels of Transparency

Table 1. Transparency Levels

Transparency	Description
Level 1 (Low)	This level provides minimal information about the actions of the drones. The user is provided a transcript, but does not know which drone spoke, the location of the drone at the time of a transmission, or the time the transmission was sent. The user also needs to deduce (with a small amount of information) why a drone has reported an error or cannot complete a command. **Example:** If the fishing drone cannot drop a fishing line, it reports "Error." The user then has to figure out what type of error ocurred : deployment, movement, or speech related.
Level 2 (Med)	This level provides more information than level one by indicating the type of error the drone has reported and the location of a transmission. The user has more transparency about actions the drones have taken but has to infer why a drone cannot complete a command. **Example:** If the fishing drone cannot drop a fishing line, it reports "Error, Deployment." This tells the user that the error from the drone is related to attempting to deploy the line. However, it is unclear if the deploy error is related to the user being out of a specific bait or hook or if the drone currently has a fish on the line.
Level 3 (High)	This level provides the most information. The user has the most transparency about actions the drones have taken and do not need to guess about any of their actions (past, present, or future).

	Example: If the fishing drone cannot drop a fishing line, it reports "Error, Deployment, out of spotted bait" This informs the user that the drone could not drop the fishing line because they are out of the bait type the user wanted to apply.

This study delineates three levels of automation transparency: low, medium, and high. Low transparency is defined as having minimal information regarding the actions of the drones, requiring the user to determine status information that is sent via information displayed on the screen regarding the operational picture. Medium transparency provides status information along with additional information in the form of brief descriptive statements. It also provides additional current information, but still requires the user to investigate and understand reasoning why a drone is in a specific state. High transparency provides the most information, including the information from Low and Medium transparency levels, as well as its state and cause of it. The High transparency level is intended to eliminated the need to infer information based on the operational picture (see **Table 1**).

Grammar Sets

To examine the impact of language—specifically, verbal command formats—on the participants as they performed the task, two different grammar sets were used for the automation. The first was a set of pre-defined utterances to be provided and rehearsed at the beginning of the lesson. The second grammar set was a set of utterances customized by participants (retaining critical words) prior to starting the study. Participants were able to request verbal help from the researcher if they forgot the command. See Table 2.

Table 2. Drone Pre-Set Grammar

Fishing Drone Pre-set Grammar	
Move to a Cell	"Fish move to cell [Alpha/Bravo/Charlie/Delta] [1/2/3/4]"
Catch Fish	"Fish drop line with [spotted/striped/solid] bait on a [large/small] hook"
Keep Fish	"Fish Store"
Release Fish	"Fish Release"
Assault Drone Pre-set Grammar	
Move to a Cell	"Assault move to cell [Alpha/Bravo/ Charlie/Delta] [1/2/3/4]"
Drop Net	"Assault drop net"
Drop Flash-Bang	"Assault drop flash-bang"

Research Hypotheses

The following are the eight hypotheses for the study. See Figure 1 for the hypothesis diagram.

H1: An interface that provides more transparency than another interface will significantly increase participants' fishing performance.
H2: An interface that provides more transparency than another interface will induce significantly lower workload.
H3: An interface that provides more transparency than another interface will induce significantly higher participant situation awareness (SA).
H4: An interface that provides more transparency than another interface will induce significantly higher participant trust in the automation.
H5: Participants who are allowed to generate their own grammar sets for communicating with drones will produce significantly better fishing performance than those who use a pre-determined grammar set.
H6: Participants who are allowed to generate their own grammar sets for communicating with drones will commit significantly fewer speech errors than participants who use a pre-determined grammar set.

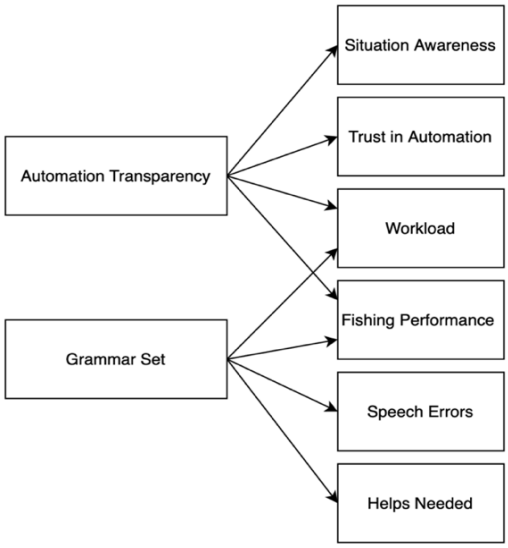

Figure 1. Hypothesis Diagram

H_7: Participants who are allowed to generate their own grammar set for communicating with drones will have a significantly lower workload.
H_8: There will be statistically significant lower number of "helps" needed (i.e., reminders of command grammars) for user-generated commands vs. canned commands.

METHOD

Participants

Fifty-one individuals participated in the research study, recruited via social media postings, emails, and word of mouth. Of those, data from six participants in Group B were eliminated due to technical issues; as a result, the total number of participants included in the research was 45. The breakdown per group is displayed in Table 3. Participants ranged in age from 18 to 63 (M = 30.40), with 29 males and 16 females. Participants were asked for their frequency of use for voice devices, with 42% indicating daily use, 16% indicating weekly use, 18% indicating occasional use, 16% indicating rare use, and 9% indicating they never use voice activated devices. Regarding video game play, 22% of participants reported daily gaming, 20% reported weekly gaming, 22% reported occasional gaming, 31% indicated rarely gaming, and 4% indicated never playing video games.

Table 3. Participants Per Group

***N* by Experimental Condition**				
Grammar Groups	**Transparency Groups**			**Total**
	1	**2**	**3**	
A	*8*	*9*	*9*	**26**
B	*6*	*7*	*6*	**19**
Total	**14**	**16**	**15**	**45**

Participants were screened for age, risk factors, hearing and vision, and English language proficiency (as the speech agents are programmed to give commands and understand commands in English). Forty-one participants reported English being their first language, with one participant reporting Chinese as a first language, two participants indicating Spanish as a first language, and one participant declining to respond. Of the language responses, 56% reported speaking only one language, 40% reported speaking two languages, and 4% (2 people) reporting speaking three or more languages.

All participants, save one, reported using GPS for local navigation; of those respondents, 16% always followed GPS instructions verbatim, 68% often followed GPS instructions verbatim, and 16% sometimes followed GPS instructions verbatim. Participants also self-reported the level to which they understood how intelligence-based GPS systems work: 20% indicated a very high understanding, 38% indicated moderate understanding, 29% indicated somewhat understanding, 9% indicated low understanding, and 4% indicated no understanding at all.

Apparatus

The authors created a game-based testbed using the Unity game engine called "DroneFisher". The goal of DroneFisher was to mimic the P-8A task in a generic domain with as much similar tasking to enable a larger subject pool to participate in the experiment. A game was developed wherein drones—some autonomous and some controlled by the player—could use various sensors to find fish in a playing field while using assets such as hooks, bait, nets, and flashbangs to herd and catch fish based on a fish list. This ensured a constant stream of communication between the drones and the player querying the sensors and directing actions. In addition, a sea monster—the Kraken—would roam the playing board and jam the sensors, forcing the players to further rely on the autonomous drones. During previous demonstrations, P-8A instructors provided feedback that the game was a reasonable approximation of the P-8 task. See Figure 2 for a screenshot of DroneFisher.

Procedure

All participants followed the workflow displayed in Figure 1 after completing the informed consent process. Next, participants received a brief with game instructions, completed the demographics questionnaire, and began the experimental gameplay (including situational awareness questions). Participants completed the DroneFisher task seven times. The first two rounds served as training rounds with smaller playing areas and fewer fish to catch. Participants were randomly assigned to low (1), medium (2), or high (3) transparency groups. In addition, the participants were randomly sub-divided into groups with pre-set grammar (A) or custom grammar (B). The only distinction occurred between the A and B groups: Group A received instructions on pre-set grammars for commanding the drones, while Group B received training on how to edit the grammar to create their own custom commands and were assisted in doing so via the software provided. Following the last round of gameplay, participants took the National Aeronautics and Space Administration - Task Load Index (NASA-TLX), an abridged version of the Human-Computer Trust (HCT) Scale (Mardsen & Gregor, 2000), and provided general feedback about the game. Participants were compensated $10 per half-hour of participation, paid via gift card.

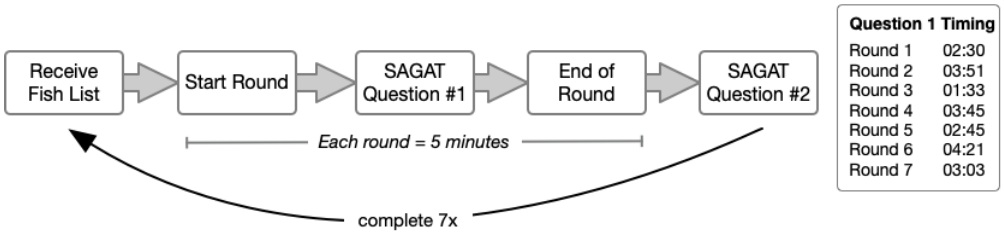

Figure 1. Experimental Task Order

RESULTS

Overview

A significant difference was found for mental workload between high and low transparency as well as for temporal demand (a component of workload) for high and low transparency. When all six subscales of workload are aggregated into a global workload score, a significant difference exists again between high and low transparency. No other significant results related to transparency were found.

Figure 2. Participant View of the DroneFisher Playing Area

Preliminary Analysis

Data were extracted from DroneFisher as .csv files and subjected to statistical analysis via Microsoft Excel and SPSS. Prior to running analyses, we examined the data and omitted participants who encountered technical issues that impacted gameplay and the participants to struggle with some of the basic commands. Those individuals were excluded from the data set prior to further analysis.

Prior to running analyses related to the hypotheses, we attempted to determine if there was a significant difference in performance between Rounds 1 and 2—the "training" rounds with a smaller playing field—and Rounds 3 through 7. If this analysis did not show a significant difference, we intended to include all rounds in the analysis. If performance did show a significant difference, we intended to include only Rounds 3–7 in the analysis.

Performance Analysis

In order to test for difference between the practice rounds and the testing rounds, we examined two measures: the accuracy of the fish caught and the situational awareness question accuracy. For both measures, we ran a paired two-sample *t*-test between Rounds 1 and 2 together and Rounds 3 through 7 together for each transparency level.

Transparency level two (T2) and transparency level three (T3) had significant performance differences between the training rounds and the testing rounds, but only for fishing performance (T2 Rounds 1 and 2 (M = 0.327, SD = 0.309) and Rounds 3 through 7 (M = 0.423, SD = 0.191): $t(15) = -1.816$, $p = 0.045$; T3 Rounds 1 and 2 (M = 0.347, SD = 0.271) and Rounds 3 through 7 (M = 0.463, SD = 0.219): $t(14) = -2.324$, $p = 0.018$).

We ran several additional two-sample t-test to investigate if transparency level impacted any performance dimensions other than accuracy. We found that T_2 participants (M = 0.250, SD = 0.022) used significantly more Assault Drone resources than T_3 (M = 0.196, SD = 0.050); $t(4) = 2.225$, $p = 0.028$., T_1 (M = 0.319, SD = 0.032), and they used significantly more fishing resources than T_2 (M = 0.256, SD = 0.011) or T_3 (M = 0.257, SD = 0.021); $t(5) = 4.145$, $p = 0.004$ and $t(8) = 3.619$, $p = 0.003$. Additionally, T_3 did not catch any incorrect fish.

Workload Analysis

We also conducted statistical analyses on workload effects based on transparency levels in order to understand potential impacts of transparency on the instructors. Looking across all workload dimensions into an Using the aggregated global workload score, an ANOVA indicated significant difference in aggregated workload $F(2, 42) = 4.49$, $p = .017$; partial $\eta^2 = .176$ specifically between the high (Level 3) and low (Level 1) levels of transparency ($p = .014$).

A MANOVA was used to assess automation transparency's effect on all six sub-scales of workload, and it did not indicate significance ($F(12, 74) = 1.47$, $p = .155$; Wilk's $\Lambda = .652$, partial $\eta^2 = .192$. However, between-subjects analysis showed that transparency had a statistically significant effect on Mental Demand ($F(2, 42) = 4.02$, $p = .025$; partial $\eta^2 = .161$) and Temporal Demand ($F(2, 42) = 3.25$, $p = .049$; partial $\eta^2 = .134$). This significant difference for Mental demand ($p = .024$) and Temporal demand ($p = .047$) was only between the low (Level 1) and high (Level 3) transparency levels.

Grammar

Participants who could create their own grammar (Condition B; M = 0.57, SD = 0.16) had significantly higher mean fishing performance than participants using the pre-set grammar (Condition A; M = 0.39, SD = 0.19): $t(43) = -3.45$, $p = 0.001$, $d = 1.05$ based on fishing accuracy and situational awareness as measured by the Situation Awareness Global Assessment Technique (SAGAT) questions. Participants who created their own grammar (Condition B; M = 0.194, SD = 0.12) also committed significantly fewer speech errors than participants who used the pre-set grammar condition (Condition A; M = 0.295, SD = 0.17): $t(43) = 2.159$, $p = 0.018$, $d = 0.68$, meaning they had fewer incidents of using incorrect phraseology to the automated drones. Finally, participants who could create their own grammar (Condition B; M = 1.11, SD = 1.60) requested significantly fewer helps – reminders of what the commands were – than the pre-set grammar condition (Condition A; M = 3.115, SD = 4.90): $t(32) = 1.954$, $p = 0.030$, $d = 0.55$.

Finally, the custom grammar condition (Condition B) resulted in significantly lower workload scores on the Frustration sub-scale (A (M = 68.08, SD = 22.45), B (M = 45.00, SD = 27.13): $t(43) = 3.118$, $p = 0.002$, $d = 1.06$).

DISCUSSION

There were no significant results related to the effect of transparency on fishing performance and situational awareness. However, workload was negatively impacted by transparency level indicated by higher reported workload with the highest level of transparency than with the lowest level of transparency. The increase in workload was likely due to a constant stream of long utterances coming from the autonomous drones, causing the participant to continually monitor what was happening and to wait for an opening to issue a command. In summary, the participants in the high transparency level were possibly suffering from information overload as a result of receiving more information to process. Conversely, the low level of transparency did not negatively impact performance

From these additional analyses, we might further hypothesize that the lower automation transparency levels offered less detailed information, therefore hindering the participant from taking more precise actions. Alternately, the lower automation transparency levels may have afforded more time for the participants to examine their available resources since their perceived workload was reported as lower, providing increased mental capacity to store resource

availability in working memory. Overall, however, the task may have simply been too difficult for a typical participant to master within the given time.

Grammar

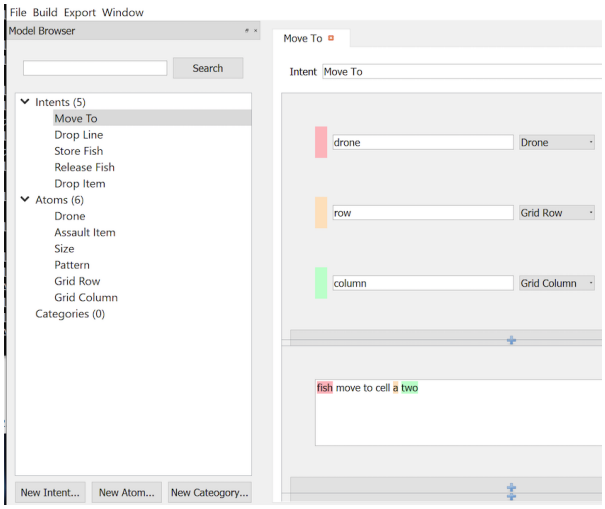

Figure 3. Interface for Creating Custom Grammars

Providing individuals using a voice-based system with the opportunity to write their own grammars (within the limits of a pre-defined actor-action set) resulted in user benefits within the context of the fishing game. These benefits manifested in four ways: a reduction of speech errors, a reduced need for help remembering the correct command vocabulary, improved task performance as measured by fishing accuracy, and lower reported workload as compared to the pre-defined grammar condition. This supports a recommendation for how to design speech-based systems, which are increasing in prevalence, particularly in the commercial world.

The demonstrated benefit of user-defined custom grammars may be that the act of creating the grammar itself provides a metacognitive learning opportunity to the users. Metacognition, the strategy of thinking about and analyzing the task to be performed, has been a proven strategy for learning improvement since the 1970s (see Lin, 2001 for a review). When the participants created their own grammar, they were presented with *intents* – actions that the system can take (see Figure 3). The intents have *atoms*, which are nouns that can describe the various actors (e.g., drones, grid rows). The act of parsing the statements and understanding the implicit order of the message may help the participant understand how the agents function; this makes it easier for participants to use the agents effectively. Additional research to investigate the potential of metacognitive effects would include a similar interface for those using the pre-set commands.

Limitations

The most notable limitation of the study, though unavoidable, was the small sample size. A future iteration should include a larger sample in order to increase statistical power. An additional limitation is the lack of a control group to enable comparison to a condition with no experimental treatment.

Future Directions

In summary, more research is needed to understand automation transparency levels; however, an increase in information did not provide any benefit to performance or situational awareness within the confines of this study. Rather, it negatively affected workload, specifically SA and performance over time. Ideally, the DroneFisher study would be replicated with several changes. A follow-on experiment should replicate the study as-is but use members of the P-8A community. Using participants from an end-user community would mitigate the concern that the task is too difficult for a non-expert participant to successfully complete; however it might introduce new biases with respect to trust in automation derived from previous experiences. The concern of the task being difficult to successfully complete by non-experts emerged from the need to multitask and rare completion to catch all fish each round by any participant. The second new experiment would both revisit the use of non-expert participants and would simplify the task considerably. By altering factors such as the number of fish variants, or reducing the number of autonomous agents, the task difficulty could be reduced.

CONCLUSION

This study demonstrated that workload was negatively impacted by Transparency Level such that the more data that was provided, the more overwhelmed the participants became. The increase in workload is likely due to a constant stream of long utterances coming from the autonomous drones, as briefly mentioned in the Project Approach section.

This caused the participant to constantly monitor what was happening and wait for an opening to issue a command. In summary, the participants in the high Transparency Level were possibly suffering from information overload. This study examined the implications of automation transparency and control over voice commands on workload, situation awareness, and performance.

As new ways to input information (e.g., voice commands, hand gestures) are developed, it is imperative to ensure that usability of the system is not compromised. As the P-8A community investigates the use of autonomous agents for training support, more research is recommended before the instructor graphical user interface (GUI) is finalized. As seen in this research, as well as other recent literature, providing too much transparency (which in this study equated to amount of information) causes information overload, which negatively affected perceived workload without demonstrating improved SA or performance. If the P-8A community were to implement synthetic teammates without conducting additional research, it would be recommended to provide automation transparency at a low level or allow the user to choose the level of transparency desired.

ACKNOWLEDGEMENTS

The research reported in this paper was funded by the Naval Air Warfare Center Training Systems Division (NAWCTSD), Contract Number N68335-18-C-0198. The authors wish to thank Charles Newton, Eric Tucker, Robert Picking, Blake Johnson, Dajia Ortiz, and Kristen Mills, without whom the research would not have been possible. The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives.

REFERENCES

Chen, J. Y. C, and Barnes, M.J. (2014). Human–agent teaming for multirobot control: a review of human factors issues. *IEEE Trans Hum-Mach Sys*., *44*(1):13-29.

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). Situation awareness-based agent transparency. *Army Research Lab Aberdeen Proving Ground Human Research and Engineering Directorate*.

Christoffersen, K. and Woods, D.D. (2002). How to make automated systems team players. In E. Salas (ed.), *Advances in human performance and cognitive engineering research*, volume 2, 1–12. JAI Press, Kidlington, UK

Demir, M., McNeese, N. J., & Cooke, N. J. (2016, March). Team communication behaviors of the human-automation teaming. In 2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA) (pp. 28-34). *IEEE*.

Durso, F. T., Hackworth, C. A., Truitt, T., Crutchfield, J., Nikolic, D. and Manning, C., (1998), Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Association Institute, Inc*.

Endsley, M. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors 37*(1), 32-64.

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics, 42*(3), 462-492.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

Helldin, T. (2014). *Transparency for Future Semi-Automated Systems: Effects of transparency on operator performance, workload and trust* (Doctoral dissertation, Örebro Universitet).

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, *62*(3), 345-360.

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

Lin, X. (2001). Designing metacognitive activities. *Educational Technology Research and Development*, 49(2), 23–40.

Madsen, M., & Gregor, S. (2000, December). Measuring human-computer trust. In *11th Australasian conference on information systems* (Vol. 53, pp. 6-8).

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human factors, 60*(2), 262-273.

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

Sarter, N.B., Woods, D.D., and Billings, C. (1997). Automation surprises. In G. Salvendy (ed.), *Handbook of human factors & ergonomics*, 19–35. Wiley, New York, NY, 2nd edition.

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017, March). Evaluating effects of user experience and system transparency on trust in automation. In 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI (pp. 408-416). *IEEE*.