Can Accountability be Instilled, in the Absence of an Authority Figure, in a Way Which Enhances a Human-Automation System?

by

Adam Wilkins

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved March 2019 by the
Graduate Supervisory Committee:

Erin Chiou, Chair
Robert Gray
Scotty Craig

ARIZONA STATE UNIVERSITY

May 2019

ABSTRACT

As automation becomes more prevalent in society, the frequency that systems involve interactive human-automation control increases. Previous studies have shown accountability to be a valuable way of eliciting human engagement and reducing various biases, but these studies have involved the presence of an authority figure during the research. The current research sought to explore the effect of accountability in the absence of an authority figure. To do this, 40 participants took part in this study by playing a microworld simulation. Half were told they would be interviewed after the simulation, and half were told data was not being collected. Eleven dependent variables were collected (accountability, number of resources shared, player score, agent score, combined score, and the six measures of the NASA- Task Load Index), of which statistical significance was found in number of resources shared, player score, and agent score. While not conclusive, the results suggest that accountability affects human-automation interactions even in the absence of an authority figure. It is suggested that future research seek to find a reliable way to measure accountability and examine how long accountability effects last.

TABLE OF CONTENTS

Page

LIST OF FIGURES

INTRODUCTION

While accountability is a commonly utilized word in many facets of society, especially politics and business, there is surprisingly little research on how it fits into today's technologically advanced world. Much of the accountability research over the years has looked at human-human interaction, but there is increasingly more human-automation interaction in our daily lives. This change in types of interaction creates a whole new arena in which accountability should be studied, so we can properly design human-automation systems. Past research on accountability has shown that it has the ability to reduce certain biases (Tetlock, 1983a; Tetlock, 1985), promote more complex thought (Tetlock 1983b; Tetlock 1987), and increase performance (Skitka, Mosier, & Burdick, 2000; Shah & Bliss, 2017); but it is unclear if this previous research directly pertains to today's human-automation systems.

Humans often fail to effectively interact with automation, leading to issues such as reduced system resilience and inefficient system operation. This issue can be seen in many situations; such as people sleeping behind the wheel of self-driving vehicles, despite being aware that they are supposed to remain alert and ready to take control. Accountability is a promising tool for overcoming these negative conditions, by manipulating the abilities noted above. However, if human-automation systems are to reap the potential benefits of accountability in their designs there is still much research that needs to be completed. It is hypothesized that, even in the absence of an authority figure, accountability can be instilled in a manner which increases human engagement with automation thereby enhancing the performance of a human-automation system.

A literature review will seek to establish what we currently know about accountability, automation, and the studies performed on accountability up to this point which are most relevant to the social sciences, particularly as related to human-automation systems. This paper will also present a discussion of some gaps in the current literature, and a research question based on these gaps is offered: will accountability have a significant effect on how participants collaborate with an interactive automation system, on how they rate their preparedness to account for their actions, and on how they perceive their workload, even in the absence of an authority figure moderating the study? The chosen study design and methodology for answering the question will be discussed. Results of the collected research data will be presented. Then, finally, a discussion of the implications of the findings, and potential next steps for understanding accountability, will be offered.

LITERATURE REVIEW

Accountability. The concept of accountability is utilized in many areas of society. In the broadest of terms accountability can be seen as giving or demanding of reasons for conduct (Roberts & Scapens, 1985). Mulgan (2000) notes that accountability has been a dynamic concept, but the general consensus is that it regards the process of accounting for actions to some authority. Accountability frequently takes on different, yet similar, meanings across fields, depending on the context it is being used in. For instance, in public administration it is often viewed as power relationships between principal and agent, primarily referring to duties which agents owe to those with leverage and control of their tenure in office (Uhr, 1993). On the other hand, business often views

accountability as justification of actions and beliefs, often in the context of rewards or punishments (Gitter & Masicampo, 2007).

It is important to note that, while accountability and responsibility are often used interchangeably, the concept of accountability is fundamentally different from responsibility. Bivins (2006) notes that responsibility denotes predetermined explicit obligations, whereas accountability refers to the readiness to explain actions, intentions, judgments, or omissions to relevant others. While there is often significant overlap between responsibility and accountability, accountability does not rely on the existence of explicit obligations. Simply put, a responsibility is an obligation, and accountability is a preparedness to account.

Much of the accountability research literature focuses on public administration or management. Some valuable insights can be garnered from this research. In public administration four systems of accountability (bureaucratic, legal, professional, and political) have been identified; each system consists of either an internal or external source of control, and either a high or low degree of control over actions (Romzek & Dubnick, 1987). This is important because it informs of the impact that the sources and levels of control have on accountability. In business five forms of accountability are often recognized, those being political, managerial, public, professional, and personal (Sinclair, 1995). These forms of accountability have varying effects on people, so understanding them is important when designing research, especially in the social sciences.

While the previously mentioned views and research on accountability are informative, they are not optimal for understanding accountability in the social sciences. In most fields, accountability is tied to understanding the outcomes of performance

expectations, and not necessarily the underlying motivators of accountability on people. Because of this difference in objectives, accountability in the social sciences may be viewed as "pressure to attend to more information and to also integrate this information in more complicated ways" (Skitka et al., 2000, p. 704). While the social sciences view may elicit accountability in similar manners to other fields, the outcomes are often viewed quite differently. Instead of simply seeking to understand why a performance objective was or was not met, the social sciences seek to understand how accountability affects many aspects of human thought and motivation.

Human-human accountability studies. A handful of seminal studies on accountability have been conducted in the area of human-human interaction. These studies found accountability affects thought processes. Accountability was shown to reduce the primacy effect, causing individuals to ascribe less weight to the first information obtained than those in the nonaccountable condition (Tetlock, 1983a). It was also found to reduce the fundamental attribution error; the tendency to place more emphasis on internal explanations without properly taking the situational powers into account (Tetlock, 1985). Yet another finding on accountability was that it can cause more complex judgment processing; in that it motivates more vigilant, thorough, and self-critical information processing (Tetlock & Kim, 1987).

A useful study that examines accountability's impact on the overall complexity of thought was completed by Tetlock (1983b). It found that individuals either shifted their views or thought about them in a more complex manner when they believed they would have to justify their stance to someone with a different position. This study showed that when and how accountability is induced can be important to the outcome. Accountability

is a multi-faceted complex topic which still requires much research to utilize in the most efficient and effective way possible.

A review of the social sciences accountability literature found that, in most situations, the benefits of accountability were most pronounced when accountability was induced pre-decision, and to an unknown audience (Lerner & Tetlock, 1999). While this finding is interesting, it is not optimal; in many real-world situations it is unlikely that the audience is unknown (e.g., managers, voters, etc.). It is also unclear if this research into human-human interactions directly correlates with human-automation interactions. In order to assess the research needs of the accountability topic, it is necessary to understand automation and how accountability has been studied in human-automation systems.

Automation. More and more frequently, automation is being integrated into our daily lives. Automation can be defined as the execution of a function, previously carried out by a human or one which could conceivably be carried out by a human, by a machine (Parasuraman & Riley, 1997). There are multiple factors which influence how people interact with automation. One important factor is the level of automation. The literature on levels of automation range from comprehensive listings of the levels and how they are distinguished (Sheridan, 1992; Parasuraman, Sheridan, & Wickens, 2000), to abbreviated levels simply ranging the automation level between no automation and full automation (Parasuraman, 2000). Another important factor is the level of control a human has in the human-automation system, viewed as manual, advisory, interactive, supervisory, or automatic control (Van Wezel, Cegarra, & Hoc, 2010). As we know from research in the public administration domain, control can be a critical variable in accountability.

Generally speaking, levels of automation and control can affect not just how people use automation, but how they misuse it, whether it becomes disused, and even if its use becomes abused (Parasuraman & Riley, 1997). The levels of automation and control can also have a profound impact on the mental state of the people utilizing the automation. Some of the areas which can be affected by automation are mental workload, situational awareness, complacency, and skill degradation (Parasuraman, 2000). These are often the types of effects social scientists are interested in when it comes to human-automation systems, and it is these areas that are likely to be affected by accountability.

As automation becomes more and more prevalent in all areas of society the relationships between humans and automation become more complex, and eliciting the proper interaction becomes more important. In the past, much of the automation involved either advisory or supervisory control, but more automation is now utilizing an interactive type of control. This dynamic coordination to accomplish a task can be viewed as interactive team cognition (Cooke, Gorman, Myers, & Duran, 2012). To achieve interactive team cognition between humans and automation requires flexibility and proper engagement of system participants (Flemisch et al., 2012). Achieving a form of interactive team cognition in human-automation systems is one area where accountability may be beneficial, but more knowledge on accountability in human-automation systems is required.

Studies of accountability in human-automation systems. As of today, few studies on accountability in human-automation interaction have been published. The two studies which were found focused on supervisory control, and measured observed/unobserved errors as a critical dependent variable (Skitka et al., 2000; Shah &

Bliss, 2017). These studies found that accountable participants had significantly different scores in their observation tasks than nonaccountable participants. This proved that scoring metrics can be a valuable tool for assessing the behavioral difference between those being held accountable and those not being held accountable. While these studies are informative, and certainly add value to the existing accountability literature, the focus on supervisory control means they are concerned not with effective human-automation interactions, but rather on effective human monitoring of the automation. It is possible that the outcomes of accountability may be different depending on the level of automation and type of control.

In an article by Cummings (2006), it is pointed out that understanding accountability in human-automation systems can inform design of these systems in a manner that reduces biases and increases system functioning. The existing literature regarding human-human and supervisory human-automation interaction (involving accountability) may pertain to other types of human-automation interaction, but more research is needed to establish a correlation. Given the minimal amount of existing literature on this topic, it is easy to see that this is a relatively unexplored area. An important area which has remained unexplored in the accountability literature is how to measure accountability.

Measuring accountability. While researching the accountability literature no way of measuring accountability was found. To assess accountability levels the following question is being proposed: How prepared do you believe you are to justify your decisions and actions in the game to an interviewer? Participants will answer this question by selecting a point on a seven-point Likert-type scale ranging from “not at all

prepared" to "very prepared". This question relies on face validity. A key component of accountability is the "accounting" for actions or decisions in an interview or interrogation. By asking the participant how prepared they are to justify their decisions and actions, it will show how mindful they were of their decisions and actions during game play, and therefore how accountable they felt during the game. It is hypothesized that the accountable participants will rate themselves as more highly prepared to justify their decisions and actions than will the nonaccountable participants. If this is the outcome achieved, then it will help to show that accountability can indeed be instilled even without the presence of an authority figure. It is also helpful, however, to examine how accountability impacts perceptions of workload if we are to understand the role accountability may play in human-automation interaction design.

NASA-Task Load Index (NASA-TLX). The NASA-TLX is a well-established research method, utilized to understand how individuals perceive the workload of a given task. The NASA-TLX, developed by Hart and Staveland (1988), asks subjects to subjectively rate a task on six factors; those being mental demand, physical demand, temporal demand, performance, effort, and frustration level. By assessing these six metrics it is possible to gain information on the cognitive workload evoked during a task. This is valuable in the context of accountability, as an accountable individual whom is more actively engaged with automation should exert a higher cognitive effort, and will therefore likely experience a greater workload than a nonaccountable individual on many of these metrics (Salehi & Chiou, 2018).

Discussion of research gaps in accountability literature. Our current understanding of accountability shows promise for its ability to be effectively integrated

into human-automation system design. There is still much research to be done though, as there are currently significant gaps in the literature. The primary gap of interest in the current research is that it is unclear whether accountability will be effective in human-automation systems in the absence of an authority figure. In the past literature the researcher posed as an authority figure and was present during the experiment. This presents a couple of potential issues for real-world application. The first being that in most situations it is unlikely that an authority figure will be in close proximity at all times. This leads to the second issue of social influence pressure on participants. Authority figures exert social influence which increases compliance (Cialdini & Goldstein, 2004). A well-known example of this involves research participants supposedly electrically shocking an unseen person at the authority figure's prompting (Milgram, 1975). This is important because if individuals are willing to inflict harm on another person when an authority figure is present and instructing them to do so, authority is clearly going to affect accountability.

A second gap, which was not researched in this study but should be in future studies, is whether accountability can be elicited in a manner which guides humans towards exerting more effort on a specific goal. Previous research focused on whether accountable and nonaccountable participants differed in behaviors, but not if those behaviors could be directed. Increasingly complex human-automation systems can be too cognitively demanding for an individual to be engaged in every aspect of the system, so it may be important to guide their engagement through directed accountability. An example of this is an Unmanned Aerial System (UAS); it is unrealistic to expect an individual to be engaged with every aspect of an incredibly complex UAS, without significant

cognitive overload, so it may be helpful to manipulate accountability to focus the human's cognitive effort on desired metrics. While these desired metrics are likely to be different for each type of human-automation system, a UAS may focus on duties such as navigation, reconnaissance, or flight control. It is important to understand where the most human-automation collaboration is needed, in the given system, if accountability is to be directed properly.

A third gap of great interest, but beyond the scope of the current research, is that it is currently unknown how long the effects of accountability elicitation last. Most of the experiments appear to be short in duration. It is important to know if the effects of accountability are short-lived or long-lasting, as this can greatly influence system design. If accountability needs to be frequently reestablished it may not be beneficial to some system designs.

These are just a few of the gaps present in the accountability literature, especially as it relates to human-automation interaction. As stated before, there is much that still needs to be known about accountability to properly utilize it in many systems. It is a promising concept, but the research needs to be developed. It is with this in mind that the following research question has been investigated: does accountability have a significant effect on how participants collaborate with an interactive automation system, on how they rate their preparedness to account for their actions, and on how they perceive their workload, even in the absence of an authority figure moderating the study?

Accountability has been shown to be effective in minimizing certain cognitive biases (Tetlock, 1983a; 1983b; 1985; Tetlock & Kim, 1987), reducing errors of omission and commission (Skitka et al., 2000; Shah & Bliss, 2017), as well as increasing human

performance in supervisory controlled human-automation systems (Skitka et al., 2000; Shah & Bliss, 2017); and interactive controlled human-automation systems are similar in that they frequently involve comparable human-agent interactions. It was, therefore, hypothesized that, even in the absence of an authority figure, accountability can be instilled in a manner which increases human engagement with automation thereby enhancing the performance of a human-automation system.

The hypothesis was greatly informed by the existing literature. Previous research showed that accountability increases the complexity of thought (Tetlock, 1983b; Tetlock & Kim, 1987) and reduces automation bias (Skitka et al., 2000), thereby increasing human engagement in the researched systems. The existing literature also found that accountability increased the performance of supervisory controlled human-automation systems due to the increased human engagement; it is therefore believed that accountability will also increase the performance of an interactively controlled human-automation system. This previous research was, however, carried out by researchers posing as authority figures, so the hypothesis researched in this study sought to remove the social influence presented by the close proximity of an authority figure. As stated, the hypothesis was informed by previous research, while attempting to delve into an unexplored area of accountability.

Based on the hypothesis, it was predicted that when accountability is induced: accountable individuals would exhibit more preparedness for the need to account for their actions, an interactively controlled human-automation system would perform better on performance metrics, and accountable individuals would perceive their workload differently than nonaccountable individuals. More specifically, it was predicted that

accountable participants would rate themselves as more prepared to account for their actions and motivations than the nonaccountable participants, as rated by the proposed accountability question. For the performance metrics, it was predicted that accountable participants would share more resources due to a greater concern for the effectiveness of the holistic system, and they would obtain higher player scores, agent scores, and combined scores than the nonaccountable participants, due to their higher engagement levels. For the NASA-TLX workload measures, it was predicted that accountable participants would rate themselves higher on all six metrics: mental demand, physical demand, temporal demand, performance, effort, and frustration. Mental demand would be rated higher by accountable participants, due to the increased complexity of thought instilled by accountability. Physical demand would be higher in accountable participants, because they would be more engaged with the system, increasing physical activity. Temporal demand would be higher in accountable participants, as they would perceive a greater urgency regarding time. Accountable participants would rate themselves higher in performance due to their increased engagement with the system, therefore increasing confidence in their performance. Effort would be higher in accountable participants because of their increased engagement with the simulation. Finally, accountable participants would rate themselves higher on the frustration metric, due to the increased engagement, mental demand, physical demand, and temporal demand.

METHODS

Design

The research used the basic randomized design comparing one treatment to a control. The independent variable of this research was social accountability (henceforth

referred to as “accountability” in this paper), operationally defined as pressure that motivates increased attention to greater amounts of information and a more sophisticated integration of this information. Testing of accountability consisted of one level: generalized accountability. Accountability, was implemented through the inclusion of a cue in the training, as discussed previously. A control group was utilized for comparing the accountable group to a nonaccountable group. The control group was told that data was not being collected due to a hard-drive malfunction; this story was facilitated by an external computer hard-drive being left unplugged in plain sight next to the study computer. By manipulating accountability, and by using a control group, the study was able to collect data to test the hypotheses. The design allowed for comparison between the accountability and non-accountability groups.

Data was collected on 11 dependent variables. Four performance variables were collected in the microworld simulation: (a) participant score, (b) agent score, (c) combined participant and agent score, and (d) number of resources shared. One dependent variable was collected by the proposed question to assess an individual’s level of accountability: (e) accountability. Finally, six dependent variables were collected through the NASA-TLX questionnaire, to assess perceived workload: (f) mental demand, (g) physical demand, (h) temporal demand, (i) performance, (j) effort, and (k) frustration level. All data was analyzed by performing t-tests for each dependent variable, comparing the accountable condition to the control. If the hypothesis is to be confirmed, the participants in the accountable condition will have significantly higher scores in the performance measures, they will rate themselves as more prepared, and will rate themselves higher on the NASA-TLX measures of mental demand, performance, effort,

and frustration level than the control group. The hypotheses could, however, be disconfirmed by a lack of significance found.

Participants

The study utilized 41 participants (28 male, 12 female, 1 no response; M_{age} = 21, age range = 18 – 30). This particular study was too novel and exploratory to utilize existing studies for a power analysis so as to accurately estimate the needed sample size, but it was believed that this sample size would be adequate, based on assumptions made from previous studies. The participants were recruited from the Human Systems Engineering subject pool at the Arizona State University Polytechnic campus. The only exclusion criteria being that participants which are blind or have uncorrectable vision which renders them unable to accurately view a 15.6" computer monitor, from a distance of approximately 18 inches, would be excluded. Participants recruited from the subject pool received one class credit as partial fulfillment of course requirements, in accordance with the subject pool policies.

Materials

The experiment was conducted in a lab in the simulator building at the Arizona State University Polytechnic campus. The lab is sparsely decorated, has two desks, three chairs, and two laptop computers. This sterile environment minimizes distractions and potential confounds stemming from perceptions garnered from extraneous visual stimuli.

Microworld Simulation. The primary material used in this experiment is a microworld environment that simulates a hospital scheduling task (see Figure 1). In the simulation participants play the game jointly with a computer agent. This microworld environment was created in Java. The interface has four main sections (e.g., your panel,

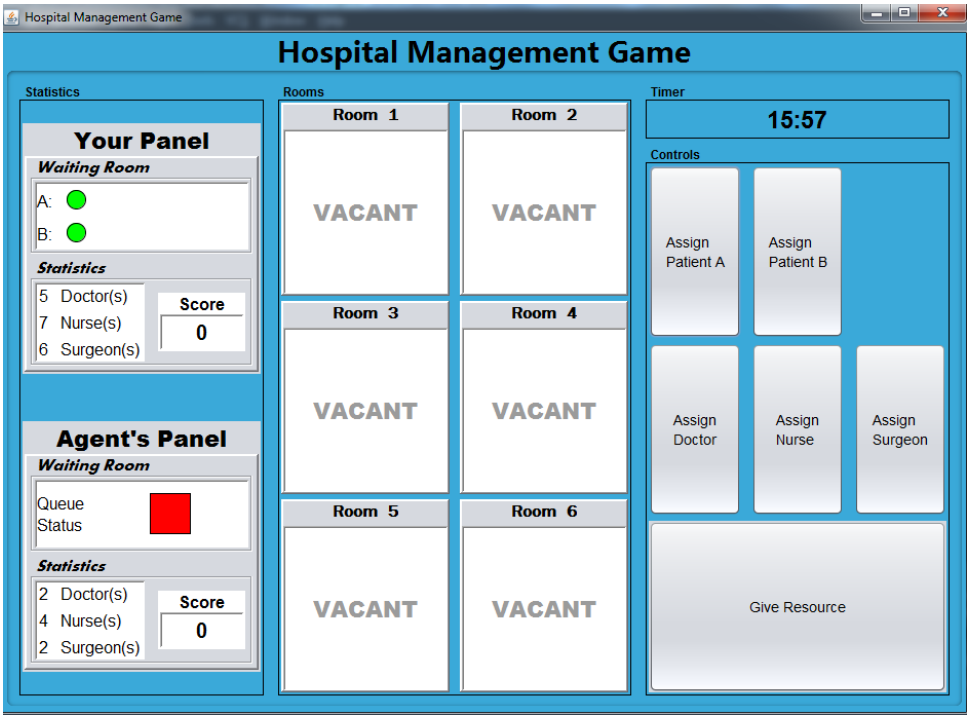

Figure 1: A screenshot of the microworld interface at the start of a trial. At the start, participants' waiting rooms were less crowded relative to the agent's waiting room.

agent's panel, rooms, and controls). The "Your Panel" display provides the number of patients in the participants waiting room (up to six of each patient type), displayed as dots with the dot's color correlating with the number of that patient type (e.g., one-two = green, three-four = yellow, and five-six = red). It also shows the number of available resources (e.g., doctors, nurses, and surgeons) and the participant's score (i.e., number of patients treated). The "Agent's Panel" provides the same information as the "Your Panel" except the information is regarding the agent's waiting room, available resources, and score. The agent's panel, however, does not show the exact number of patients in their waiting room, it displays a colored square that follows the same color pattern listed above. The "Rooms" section displays the six rooms available for treating patients and whether they are vacant or occupied by a patient and the resources currently assigned to

the patient. The "Controls" section has buttons for assigning either patient type A, patient type B, a doctor, a nurse, or a surgeon to a room. This is accomplished by clicking on the desired resource and then clicking on a vacant room. In the "Controls" section there is also a "Give Resource" button, which when clicked allows the participant to share either a nurse, a doctor, or a surgeon with the agent. The interface also has a countdown timer displaying the time left in the trial. In this simulation the agent's sharing behavior can be described as a "tit-for-tat" approach, in that it will share resources with the participant in a similar fashion as the participant shares with the agent. The pace of the game is designed so that when the participant experiences a high workload the agent experiences a low workload, and vice versa; the pace changes every four minutes. As mentioned before there are two patient types, patient type A requires a doctor, a nurse, and 45 seconds to treat, and patient type B requires a surgeon, a nurse, and 60 seconds to treat. The player receives one point for each patient treated, of either type, in their hospital, while the agent receives one point for each patient treated in the agent's hospital. It is important to note that the simulation is not meant to depict an actual hospital scheduling task, rather it was designed so that non-experts could complete the game, while creating an easy to understand narrative and goal (i.e., treat as many patients as possible).

To understand how accountability impacted participant's behaviors, during the simulation, four performance metrics were recorded through the microworld simulation. Those metrics being: the total number of resources shared, the player's total score, the agent's total score, and the combined player/agent total score. The number of resources shared metric showed the total number of nurses, doctors, and/or surgeons each participant provided to the agent. This is an important metric, because the number of

resources available in the game are finite, so how they are utilized informs where a participant's priorities lay. The total player/agent scores showed how many patients were treated in each respective hospital, with each treated patient resulting in a score increase of one point. The combined score was the sum of both the total player score and the total agent score. These scores are important because they allow an assessment of participant engagement with their hospital, the agent's hospital, and the system as a whole.

Training. Training on how to utilize the microworld environment was presented through a self-studied PowerPoint presentation. The presentation explained each section of the interface and how to interact with the environment. A slide in the accountability groups presentation mentioned an interview after the game. The accountability group saw the following line in the training: "You will be interviewed by a subject-matter expert at the end of the experiment to justify your strategies". The control group (nonaccountable group) did not have a slide mentioning an interview.

Accountability question. To assess the level of accountability displayed by participants this study asked the previously discussed question: "How prepared do you believe you are to justify your decisions and actions in the game to an interviewer?". Responses to this question were collected using a seven-point Lykert-type scale, with participants being able to select a level of preparedness ranging from "not at all prepared" to "very prepared". This data was coded numerically for quantitative analysis.

NASA-TLX questionnaire. This study is interested in how accountability affects aspects of perceived workload. Factors of the perceived workload were measured using the NASA-TLX (Task Load Index) questionnaire (see Appendix A for NASA-TLX questions). The NASA-TLX was chosen because it breaks workload into multiple factors,

allowing analysis of the individual components of task load so as to gain better information on how accountability impacts each factor. For this study the six factors of the NASA-TLX were measured using a 0-7 Lykert-type scale. Zero correlated with "low" and 7 correlated with "high" for five of the six factors (mental demand, physical demand, temporal demand, effort, and frustration level); the sixth factor (performance) used the same Lykert-type scale, but with zero correlating with "poor" and seven correlating with "good".

Demographic questionnaire. Some demographic information was collected in order to better understand certain aspects of the participants utilized in the study. The demographic information collected (see Appendix B for demographic questions) consisted of age, gender, education levels, and a range of questions assessing the participants use of technology. The questions in the demographic questionnaire were developed by the researcher.

Procedure

Upon arrival at the lab, the participants were welcomed and asked to take a seat at the desk in front of the utilized computer. The participant was then given the informed consent to read; and upon completion of the reading, verbal consent was obtained by the researcher. The participant was then presented the training material on how to utilize the microworld simulation and instructed to ask any questions they may have during training, as the researcher would be unable to answer questions during the trial. Upon completion of the training the participant was given two practice trials, of two-minutes each, to ensure they were comfortable with the functionality of the interface. If the participant had no questions after the practice trials they were asked to begin the 16-minute main trial.

Once the main trial was finished, they were asked to complete three brief questionnaires: the accountability question, the NASA-TLX, and the demographic questionnaire. Upon completion participants were informed they had completed the study and thanked for their participation.

RESULTS

Data was collected from 41 participants. Descriptive statistics were then obtained through analysis using the "Psych" package in R. As a result of the descriptive analysis one participant was removed from the accountable group due to a technical issue resulting in a failure to record the "agent score" through the microworld simulation. This resulted in two equal groups of 20 for each condition.

The "stats" package was then utilized in R to run a series of Welch's t-tests comparing the accountable group to the nonaccountable group for all 11 dependent variables (resources shared, participant score, agent score, combined score, accountability, mental demand, physical demand, temporal demand, performance, effort, and frustration). It was hypothesized that, even in the absence of an authority figure, accountability can be instilled in a manner which increases human engagement with automation thereby enhancing the performance of a human-automation system. The analysis produced mixed results, with some variables supporting this hypothesis, while others did not support it (see Appendix C for table of observed statistics).

Contrary to expectations, the accountable participants did not rate themselves as more prepared to account for their game play than did the nonaccountable group on the accountability measure, according to the observed means (Figure 2). In fact, surprisingly, the accountable group self-reported lower levels of preparedness to account (M= 5.55, SD

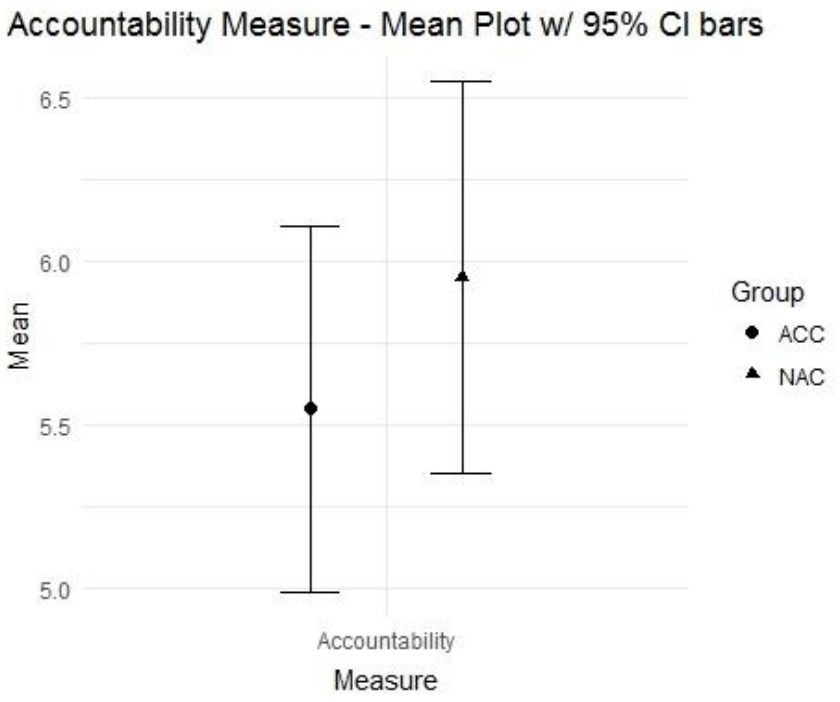

Figure 2: The observed means of participant preparedness to account, shown by group: Accountable (ACC) and Nonaccountable (NAC), with 95% Confidence Interval (CI) bars.

= 1.19) than did the nonaccountable group (*M*= 5.95, *SD*= 1.28); however, there was no significance in the accountability measure data (*t*(37.8)= -1.03, *p*= .312, *d*= .324).

It was hypothesized that accountable participants would score higher than the nonaccountable participants on the measures collected from the microworld simulation (number of resources shared, player score, agent score, and combined score). The mean scores of these measures (Figure 3) show that results of this hypothesis were mixed. For the number of resources shared measure, analysis showed the accountable group shared significantly more resources (*M*= 13.9, *SD* = 3.46) than the nonaccountable group did (*M*= 9.25, *SD* = 5.33; *t*(32.6)= 3.27, *p*< .01, *d= 1.035*). Analysis also found that the accountable group achieved significantly higher agent scores (*M*= 57.95, *SD* = 4.08) than

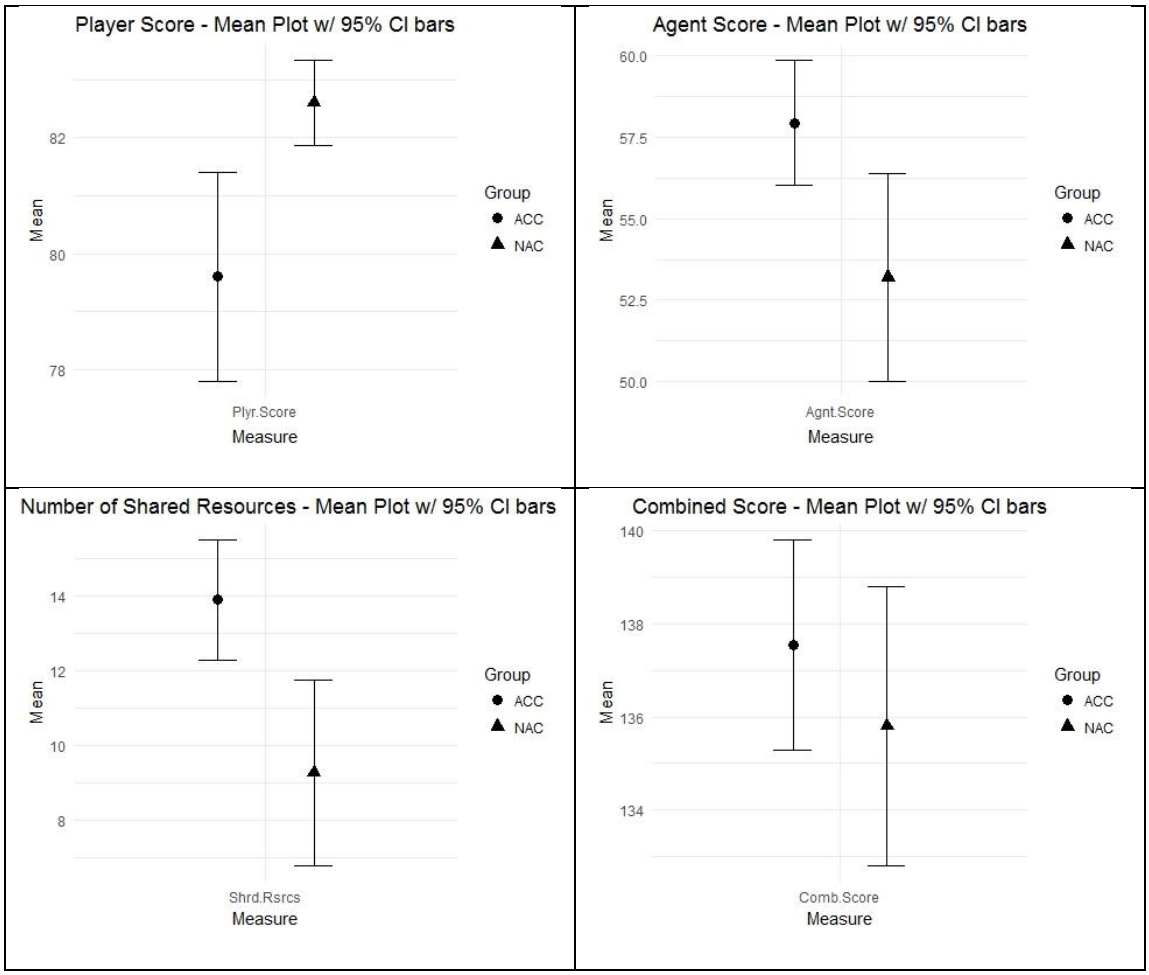

Figure 3: The observed means of the microworld simulation measures, shown by group: Accountable (ACC) and Nonaccountable (NAC), with 95% CI bars.

did the nonaccountable group (*M*= 53.2, *SD* = 6.81; *t*(31.1)= 2.68, *p*= .012, *d*= .846). Contrary to expectation the accountable group achieved significantly lower player scores (*M*= 79.6, *SD* = 3.84) than did the nonaccountable group (*M*= 82.6, *SD* = 1.57; *t*(25.2)= -3.23, *p*< .01, *d*= 1.022). No significance was found in analysis of the combined score measure when comparing the accountable group (*M*= 137.55, *SD* = 4.84) to the nonaccountable group (*M*= 135.8, *SD* = 6.43; *t*(35.3)= 0.97, *p*= .337, *d*= .308).

Analysis of the NASA-TLX data, obtained through performing individual t-tests on each measure (mental demand, physical demand, temporal demand, performance,

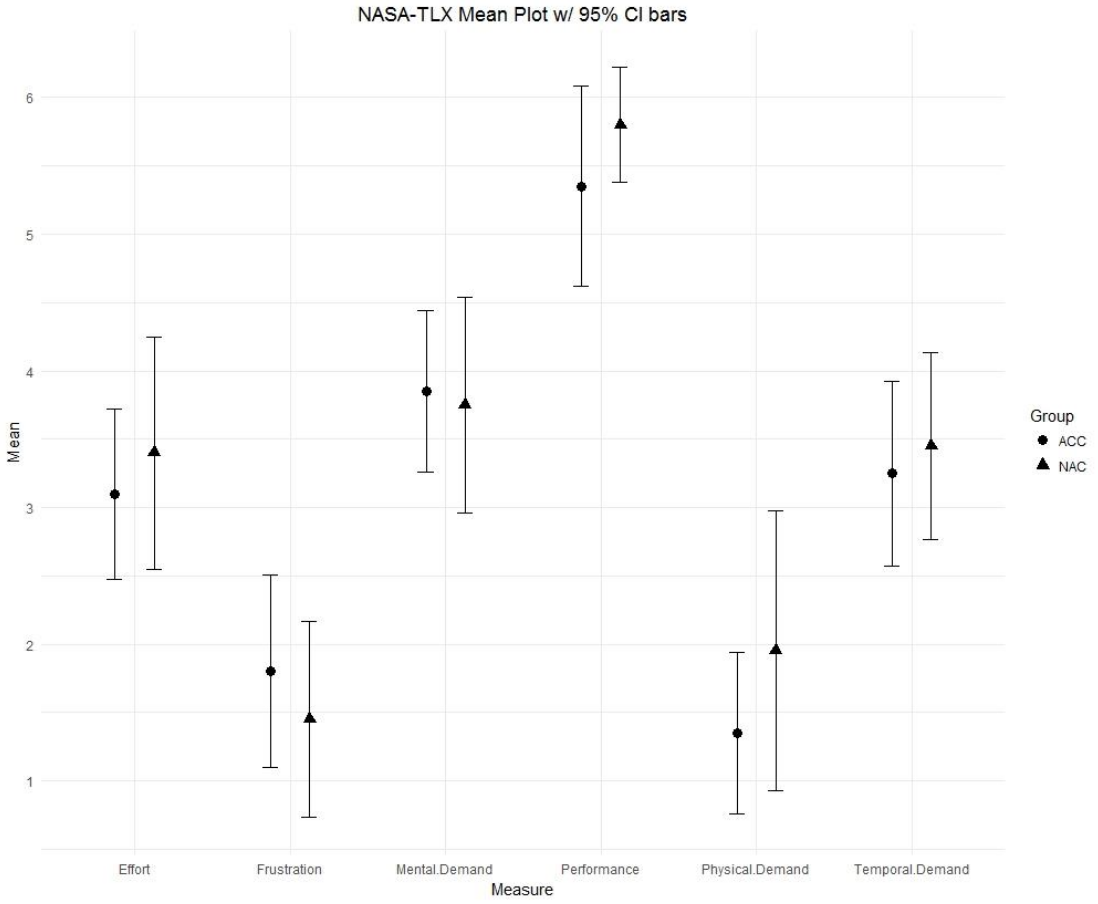

Figure 4: The observed means of the NASA-TLX measures, shown by group: Accountable (ACC) and Nonaccountable (NAC), with 95% CI bars.

effort, and frustration level), did not return any significant findings. Figure 4 shows the observed means of each of these measures. The accountable group self-reported slightly higher mental demand (M= 3.85, SD = 1.27) than the nonaccountable group (M= 3.75, SD = 1.68); however, the findings were not significant (t(35.3)= 0.21, p= .833, d= .067). The accountable group self-reported slightly lower physical demand (M= 1.35, SD = 1.27) than the nonaccountable group (M= 1.95, SD = 2.19); again, the findings were not significant (t(30.1)= -1.06, p= .297, d= .336). The accountable group also self-reported slightly lower temporal demand (M= 3.25, SD = 1.45) than the nonaccountable group

(*M*= 3.45, *SD* = 1.47); the findings were not significant though (*t*(38)= -0.43, *p*= .667, *d*= .137). The accountable group self-reported lower performance scores (*M*= 5.35, *SD* = 1.57) than the nonaccountable group (*M*= 5.8, *SD* = 0.89); however, the findings were not significant (*t*(30.2)= -1.12, *p*= .27, *d*= .353). The accountable group also self-reported lower effort levels (*M*= 3.10, *SD* = 1.33) than the nonaccountable group (*M*= 3.4, *SD* = 1.82); but, again, the findings were not significant (*t*(34.9)= -0.6, *p*= .556, *d*= *.188*). Finally, the accountable group self-reported higher frustration levels (*M*= 1.8, *SD* = 1.51) than the nonaccountable group (*M*= 1.45, *SD* = 1.54); these findings were, however, also not significant (*t*(37.9)= 0.73, *p*= .472, *d*= .23).

DISCUSSION

The implications of this study are wide-ranging and warrant an in-depth discussion. It is important, however, to begin this discussion by noting that the conducted research was exploratory and not meant to provide conclusions pertaining to the role of accountability in human-automation systems. There is still much to research on accountability before any conclusions can be made. This discussion simply seeks to provide an assessment of the results obtained, potential directions for future research, and an acknowledgment of the limitations of the current research.

The results of the accountability measure, created for this experiment, did not support the hypothesis. While a lack of significance in the results is not terribly surprising, the nonaccountable participants self-reporting higher levels of preparedness to account than did the accountable participants, on average, was unexpected. More research would be required to accurately determine the reasoning this result was obtained or if the results are repeatable, however, it is believed this result was due to one of a couple

possible factors, or a combination of them. First, the nonaccountable participants were told that due to a technical issue data was not being collected and the researcher posed as uninformed research aid as opposed to an authority figure, this potentially led to increased confidence in the nonaccountable participants' ability to account, as they believed there was nothing to dispute their account. Second, it is possible that accountability put pressure on the accountable participants which led to a reduced confidence in their preparedness to account. More research is needed to understand this outcome, however, performing this research could be beneficial for establishing a, much needed, measure of accountability which currently does not exist.

The results of the performance metrics gathered from the microworld simulation partially supported the hypothesis. The accountable group sharing significantly more resources, than the nonaccountable group, shows a higher level of engagement in the well-being of the overall system. This greater level of resource sharing resulted in the accountable group obtaining significantly higher "agent score" scores than the nonaccountable group. The greater sharing of resources allowed the agent to treat more patients, and therefore obtain a higher score. This finding reinforces the stance that accountable participants were more engaged with the overall system than were the nonaccountable participants. While the finding that the accountable group had a significantly lower "player score" than the nonaccountable group contradicts the hypothesis that performance would increase in all performance metrics, in hindsight this finding makes sense. The higher level of resource sharing hindered the accountable participants' ability to heal as many patients in their hospital, resulting in lower player scores. As stated, this result does not support the initial hypothesis, but may ultimately

show that accountable participants were willing to put their own needs behind those of the automated agent at times. The last game metric, “group score”, also contradicts the hypothesis. For this metric the accountable group scored slightly higher than the nonaccountable group, but the findings were not significant. The initial hypothesis was formed with the assumption that nonaccountable participants would be less engaged with the microworld simulation across all metrics, but it appears this belief did not hold true. It is possible that because the simulation was framed as a “game”, and participants could see scores during the simulation, this potentially induced motivation to engage more with the game than the participants may have otherwise. Further research which reduces these extraneous motivations may better show the effects of accountability on this particular human-automation system. Ultimately, however, the data collected from the game does more to support the hypothesis than it does to contradict it.

The results of the NASA-TLX metrics did not support the hypothesis. Although none of the metrics were statistically significant, it is worth looking at the results a bit more closely. Interestingly, the accountable group self-reported lower scores than the nonaccountable group on four of the six metrics. The four metrics which the accountable group self-reported lower on were physical demand, temporal demand, performance, and effort. It is possible that because accountable participants shared more resources, they reduced their workload; therefore, requiring less physical and temporal demand during game play, and reducing the amount of effort needed, as they had fewer resources to manage in their own hospital. The lower self-reporting on the performance metric for the accountable group is potentially the result of reduced confidence as a result of the accountability, as previously discussed. The two metrics self-reported higher by the

accountable group, mental demand and frustration, may imply a higher cognitive load due to accountability in this research. The accountable group potentially perceived a higher mental demand due to their desire to take in a greater amount of information to ensure they prepared to account for their game play. It is possible that this greater cognitive effort also led to greater levels of frustration, however, the frustration is also potentially the result of having given away resources which they may have not received back when they would have liked. To better understand why exactly participants reported as they did further research would be required.

While not all measures produced significant results, and some even point in the opposite direction of what was originally expected, it appears clear that inducing accountability had an effect on participants. Most notably, the significance found in the objective performance metrics is important, as these results show accountability had a significant impact on the human-automation system. As mentioned throughout this discussion, though, much research is still needed to understand the effects of accountability. Manipulation of the accountability question and/or the microworld simulation may help to further answer the many questions that still need answered when it comes to accountability.

It may be tempting to credit specific observed effects to competing theories, such as priming effects, but that would be erroneous. It could be argued that telling participants they will be interviewed elicits priming, and while it is true that this primed them to be accountable, the researcher was careful to avoid giving any indication that accountable or nonaccountable participants should act in any certain manner. This means that while they were primed to be accountable, their actions and behaviors were solely

the result of the accountability, not any specific priming. If priming is a concern, future accountability research would benefit from a longitudinal study design. While there is no consensus on how long priming effects last, the research of Squire, Shimamura, and Graft (1987) shows that "healthy" participants exhibited less influence from priming four days after the initial priming was elicited. By performing a well-designed longitudinal study, on accountability, researchers will be able to determine the duration of accountability effects while ruling out priming effects as a concern.

Limitations

The sample size constitutes a limitation. The sample size did not correlate with many of the effect sizes, therefore reducing the ability to find statistical significance. The sample size is also too small, and lacks appropriate diversity, to make conclusions regarding the larger population. This study will, however, provide data which will potentially help to inform the design of future studies.

An important limitation is that this study did not analyze speed as a factor. It is possible that speed impacted some of the outcomes, especially the performance metrics. For instance, accountable participants may have obtained lower player scores due to slower reaction times stemming from increased cognitive effort. It would be beneficial for future accountability studies to measure speed, as it could provide valuable information on the impact of accountability on factors such as cognitive effort and decision-making speeds.

Another limitation of this study involves the duration of the study's length. This study lasted approximately 45 minutes, meaning it is unclear if the effects of accountability will hold up for a longer duration. This is important given that many

human-automation systems need to work collaboratively for long periods of time. No research has investigated the length of time accountability elicitation lasts; future research would benefit from investigating this issue.

CONCLUSION

This research has attempted to expand on the body of knowledge regarding accountability in human-automation interaction. Building on previous accountability research (Skitka et al., 2000; Shah & Bliss, 2017), the current research investigated the effect of accountability on a human-automation system in the absence of an authority figure. While results of the eleven dependent variables were mixed, it does appear that accountability had an impact on the behaviors of the participants in the accountable condition when compared to the nonaccountable condition. Most notably, accountability had a significant impact on metrics of the human-automation system's performance. The accountable participants shared more resources with the automated agent, indicating an increased concern for the effectiveness of the automation. Accountable participants also obtained significantly higher agent scores, showing a greater engagement with the automation and a willingness to put the needs of the automation ahead of self at times. The argument that accountable participants were more engaged with the automation, than were the nonaccountable participants, is further strengthened by the finding that accountable participants had lower player scores than the nonaccountable participants, indicating that accountable participants behaved more selflessly. More research is needed if accountability is going to be utilized in the design of human-automation systems, but this research certainly constitutes a valuable contribution to our understanding of accountability and its potential use in system design.

REFERENCES

Bivins, T. H. (2006). Responsibility and Accountability. In K. Fitzpatrick & C. Bronstein (Eds.), *Ethics in Public Relations: Responsible Advocacy* (pp. 19-38). Thousand Oaks, CA: SAGE Publications. doi: 10.4135/9781452204208.n2

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: compliance and conformity. *Annual Review of Psychology, 55*, 591. doi: 10.1146/annurev.psych.55.090902.142015

Cooke, N. J., Gorman, J. C., Myers, C. W. and Duran, J. L. (2013), Interactive Team Cognition. *Cognitive Science*, 37(2), 255–285. doi:10.1111/cogs.12009

Cummings, M. L. (2006). Automation and accountability in decision support system interface design. *Journal of Technology Studies, 32*(1), 23-31. doi: 10.21061/jots.v32i1.a.4

Flemisch, F., Heesen, M., Hesse, T., Kelsch, J., Schieben, A., & Beller, J. (2012). Towards a dynamic balance between humans and automation: Authority, ability, responsibility and control in shared and cooperative control situations. *Cognition, Technology & Work, 14*(1), 3-18. doi: 10.1007/s10111-011-0191-6

Gitter, S. & Masicampo, E. J. (2007). Accountability. In R. F. Baumeister & K. D. Vohs (Eds.), *Encyclopedia of Social Psychology.* doi: 10.4135/9781412956253.n2

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, *52*(C), 139–183. Retrieved from http://www.stavelandhfe.com/images/NASA-TLX_paper.pdf

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255. Retrieved from https://search-proquest-com.ezproxy1.lib.asu.edu/docview/1791724876?accountid=4485&rfr_id=info%3Axri%2Fsid%3Aprimo

Milgram, S. (1975). *Obedience to authority: An experimental view* (Harper torchbooks; TB 1983). New York: Harper & Row.

Mulgan, R. (2000). 'Accountability': An ever-expanding concept? *Public Administration, 78*(3), 555-573. doi: 10.1111/1467-9299.00218

Parasuraman, R. (2000). Designing automation for human use: Empirical studies and quantitative models. *Ergonomics,43*(7), 931-951. doi: 10.1080/001401300409125

Parasuraman, R., Riley, V. A. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 39*(2), 230-254. doi: 10.1518/001872097778543886

Parasuraman, R., Sheridan, T., & Wickens, C. (2000). A model for types and levels of human interaction with automation. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 30*(3), 286-297. doi: 10.1109/3468.844354

Roberts, J. & Scapens, R. (1985). Accounting systems and systems of accountability — understanding accounting practices in their organizational contexts. *Accounting, Organizations and Society, 10*(4), 443-456. doi: 10.1016/0361-3682(85)90005-4

Romzek, B., & Dubnick, M. (1987). Accountability in the Public Sector: Lessons from the Challenger Tragedy. *Public Administration Review, 47*(3), 227-238. doi:10.2307/975901

Salehi, P. & Chiou, E. K. (2018). *Human-agent interactions: Social accountability, social loafing, cooperation and performance in interactive control.* Manuscript in preparation, Department of Human Systems Engineering, Arizona State University, Mesa, AZ.

Shah, S. J., Bliss, J. P. (2017). Does accountability and an automation decision aid's reliability affect human performance in a visual search task? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,61*(1), 183-188. doi: 10.1177/1541931213601530

Sheridan, T. (1992). Telerobotics, automation, and human supervisory control. Cambridge, Mass.: MIT Press.

Sinclair, A. (1995). The chameleon of accountability: Forms and discourses. *Accounting, Organizations and Society, 20*(2), 219-237. doi: 10.1016/0361-3682(93)E0003-Y

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human - Computer Studies, 52*(4), 701-717. doi: 10.1006/ijhc.1999.0349

Squire, L. R., Shimamura, A. P., & Graft, P. (1987). Strength and duration of priming effects in normal subjects and amnesic patients. *Neuropsychologia, 25*(1), 195-210. doi: 10.1016/0028-3932(87)90131-X

Tetlock, P. E. (1983a). Accountability and the Perseverance of First Impressions. *Social Psychology Quarterly, 46*(4), 285-292. doi: 10.2307/3033716

Tetlock, P. E. (1983b). Accountability and complexity of thought. *Journal of Personality and Social Psychology, 45*(1), 74-83. doi: 10.1037/0022-3514.45.1.74

Tetlock, P. E. (1985). Accountability: A Social Check on the Fundamental Attribution Error. *Social Psychology Quarterly, 48*(3), 227-236. doi: 10.2307/3033683

Tetlock, P. E., & Kim, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality and Social Psychology, 52*(4), 700-709. doi: 10.1037/0022-3514.52.4.700

Uhr, J. (1993). Redesigning Accountability: From Muddles to Maps. *The Australian Quarterly, 65*(2), 1-16. doi:10.2307/20635716

Van Wezel W., Cegarra J., Hoc JM. (2010) Allocating functions to human and algorithm in scheduling. In J. Fransoo, T. Waefler, J. Wilson (Eds.), *Behavioral Operations in Planning and Scheduling* (pp. 339-370). doi: 10.1007/978-3-642-13382-4_14

APPENDIX A

NASA-TLX QUESTIONNAIRE: FROM HART AND STAVELAND (1989)

Instructions: Place a mark on each scale that represents the magnitude of each factor in the task you just performed.

1	Mental Demand	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?	Low_____High
2	Physical Demand	How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating. etc.)? Was the task easy or demanding. slow or brisk. slack or strenuous, restful or laborious?	Low_____High
3	Temporal Demand	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?	Low_____High
4	Performance	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?	Poor_____Good
5	Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?	Low_____High
6	Frustration Level	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?	Low_____High

APPENDIX B

DEMOGRAPHIC QUESTIONNAIRE

Age: _____

Gender (circle): Male / Female / Non-Binary / Prefer not to answer

Highest level of education:	**You**	**Your Father**	**Your Mother**	**Your Siblings**
Some high school or less				
High school diploma				
2-year college degree / trade school				
4-year college degree				
Master's degree				
Professional degree				
Doctorate degree				

During the past three months, I did _____ hours voluntary jobs for society.

If a current student, please list the following:

College _____
Major _____
Degree in pursuit of_____
Number of years pursuing this degree _____

If employed, please write in your occupation:

_____

I use a computer (Check only one): ____ Daily ____ Every couple of days
____ Once a week ___ Every couple of weeks ____ Less than once a month
____ Never

I use the computer for (Check all that apply):
____ Looking up information
____ Email
____ Word processing
____ Spreadsheets
____ Computer games
____ Other (Please specify)

I play the following categories of video games (check all that apply):
___ Sports ___ Real-time strategy ___ First person shooter ___ Racing
___ Role-playing (RPG) ___ Puzzles ___ Arcade
Other _____

In the past six months, I have played video games on the following system type (check all that apply):
___ Console ___ Hand-held ___ PC

I use the following technologies (check all that apply):
____ Mobile phone ____ PDA/Smart phone ____ iPod/MP3 player ____GPS Navigation

How much experience do you have playing video/computer games?
____ None ____ less than 1 year ____ 1 – 2 years ____ > 2 years

I play video/computer games (Check only one): ____ Daily ____ Every few days
____ Once a week ____ Every few weeks ____ Less than once a month ____ Never

APPENDIX C

TABLE OF OBSERVED STATISTICS

Measure	Condition	n	*M*	*SD*	df	t-value	p-value	Cohen's d
Accountab-ility	ACC	20	5.55	1.19	37.8	-1.03	0.312	-0.324
	NAC	20	5.95	1.28				
Resources Shared	ACC	20	13.9	3.46	32.6	3.27	0.002	1.035
	NAC	20	9.25	5.33				
Player Score	ACC	20	79.6	3.84	25.2	-3.23	0.003	-1.022
	NAC	20	82.6	1.57				
Agent Score	ACC	20	57.95	4.08	31.1	2.68	0.012	0.846
	NAC	20	53.2	6.81				
Combined Score	ACC	20	137.55	4.84	35.3	0.97	0.337	0.308
	NAC	20	135.8	6.43				
Mental Demand	ACC	20	3.85	1.27	35.3	0.21	0.833	0.067
	NAC	20	3.75	1.68				
Physical Demand	ACC	20	1.35	1.27	30.1	-1.06	0.297	-0.336
	NAC	20	1.95	2.19				
Temporal Demand	ACC	20	3.25	1.45	38	-0.43	0.667	-0.137
	NAC	20	3.45	1.47				
Performan-ce	ACC	20	5.35	1.57	30.2	-1.12	0.27	-0.353
	NAC	20	5.8	0.89				
Effort	ACC	20	3.1	1.33	34.9	-0.6	0.556	-0.188
	NAC	20	3.4	1.82				
Frustration	ACC	20	1.8	1.51	37.9	0.73	0.472	0.23
	NAC	20	1.45	1.54				

n= number of participants, M= mean values, SD= standard deviation, df= degrees of freedom