CAPTURING INTENTIONAL TESTING OF AN AUTOMATED SYSTEM

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

By

ABRAHAM HASKINS

M.S., Wright State University, 2021
M.A., Capella University, 2015
B.S., University of Texas at Dallas, 2010
B.A., University of Texas at Dallas, 2010

2021

Wright State University

Wright State University

Graduate School

January 21, 2021

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Abraham Haskins ENTITLED Capturing Intentional Testing of an Automated System BE ACCEPTED IN PARTIAL FULFLILLMENTS OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Debra Steele-Johnson, Ph.D.
Dissertation Director

Scott Watamaniuk, Ph.D.
Graduate Program Director

Barry Milligan, Ph.D.
Vice Provost for Academic Affairs
Dean of the Graduate School

Committee on Final Examination

Ion Juvina, Ph.D.

Nathan Bowling, Ph.D.

Corey Miller, Ph.D.

Debra Steele-Johnson, Ph.D.

ABSTRACT

Haskins, Abraham. Ph.D. Department of Psychology, Wright State University, 2021. Capturing Intentional Testing of an Automated System.

Users change their behavior when interacting with automated systems based upon their trust levels. Users faced with an unknown system will adjust their trust levels as they learn more about that system. Past automation trust research has implicitly assumed that users are passive recipients of information when interacting with new systems. Feedback-seeking behavior, a pattern of behavior involving actively eliciting information about one's performance, is a well-researched concept within interpersonal research. Applying this interpersonal research to the domain of automation, I examined cases in which individuals sought feedback regarding the reliability of an unfamiliar automated system by asking for answers the user already possessed. I found evidence that feedback-seeking behavior exists within interactions with automation and called these behaviors *intentional tests* of the automated system. Users conducted more intentional tests on the system when faced with increased uncertainty (H1) and when encountering relatively early (H2) or easy (H3) trials. During these tests, users spent relatively little time assessing the system responses (H4). The effect of these intentional tests upon trust was significant yet relatively short-lasting (H5). This research shows another example of a case in which researchers may generalize the results of interpersonal research to the domain of automation. Engineers may also use these results to begin addressing a long-standing problem in automation trust: the inability for interventions to interact with long-term user behavior. These results demonstrated that intentional tests exist, can be a

useful tool, may be able to be identified automatically, and have at least some unintuitive properties that merit further study.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

Capturing Intentional Testing of an Automated System

Automation integration continues to increase in the business world (e.g., Frey & Osborne, 2013). Understanding how users interact with automated systems is critical to maximizing performance in jobs with increasingly high levels of automation. Some researchers have studied how people interact with automated systems by examining trust, defined as "risk-taking in a relationship" (Mayer, Davis & Schoorman, 1995). There is an optimal level of trust that maximizes productivity in interactions with an automated system (Bahner, Huper, & Manzey, 2008; Parasuraman & Miller, 2004). Prior research has focused on the effect of the stable trait, *dispositional trust* (i.e., propensity to trust), on behavior when interacting with an automated system (e.g., Hoff & Bashir, 2015). Comparatively less research has examined malleable trust, i.e., trust levels that change during an interaction, which researchers have labelled *learned trust* (Marsh & Dibben, 2003). Most past research has assumed implicitly that individuals are passive recipients of information from automated systems (French, Duenser, & Heathcote, 2018). However, feedback seeking research has shown that participants actively seek out information regarding their performance (Ashford & Cummings, 1983, Ashford, De Stobbeleir, & Nujella, 2016). Many properties of feedback seeking behavior hint that this behavior may be more common during interactions with automated systems than in interpersonal interactions (Morrison & Bies, 1991). If this is true, learned trust may be a dynamic bidirectional process in which users actively seek information to adjust their trust level to its optimal level. I will examine how individuals seek feedback regarding the reliability of an unfamiliar automated system, which I will call an *intentional test*. My purpose is to

examine the factors that influence the number and probability of intentional tests and distinguish intentional tests from genuine requests for help.

Automation

The automation boom has led to a number of alarming predictions and explanations in recent years. The rise of automation has been used to explain the unusual 2016 election, the alarming decoupling of GDP from income, and rising wealth inequality (Brynjolfsson & McAfee, 2013; Frey, Berger, & Chen, 2017; Hémous & Olsen, 2014). Researchers have debated the impact of the rise of automation on job availability in recent years. Whereas some estimates have placed the risk of job loss from the current automation boom in the United States as high as 47%, other analyses have placed the predicted job loss at 9% (Arntz, Gregory, & Zierahn, 2017; Frey & Osborne, 2013). These differences arise primarily from disagreements about how to account for jobs that may be partially automated and how many human positions will be eliminated as a result of this technological shift. However, most of these analyses agree that upwards of 90% of existing jobs today will involve heavy interaction with automated systems within the next two decades (e.g., Arntz, Gregory, & Zierahn, 2017; David, 2015; Frey & Osborne, 2013). To ignore such a dramatic shift in the nature of work is to court irrelevance for the field of industrial/organizational psychology.

Defining Automation. Defining automation in an absolute sense can be difficult. Finding a definition that includes modern factory machinery whereas excluding common tools such as crowbars has led some researchers to use alternative terms such as robotics (Goldberg, 2011). To solve this problem, Parasuraman and Riley (1997) opted instead for a transitional definition. Processes that were previously the domain of humans and are

now carried out by a non-human are defined as automation. Whereas this definition is the most robust in the face of unknown future technologies, for the purposes of this analysis it is not useful because it is intended to be a general domain definition rather than one that can be usefully operationalized. So instead, I will use a less comprehensive definition given by the society for Transactions on Automation Science and Engineering: "Automation… emphasizes efficiency, productivity, quality, and reliability, focusing on systems that operate autonomously, often in structured environments over extended periods, and on the explicit structuring of such environments" (Goldberg, 2011, p. 1).

Optimal Automation Usage. Automation usage does not always imply an increase in productivity (Parasuraman, Molloy, & Singh, 1993). Errors in usage can result in a decrease in productivity, and researchers have identified two common errors: automation complacency and automation bias (Parasuraman & Manzey, 2010). Automation complacency is defined as an overreliance on automated action. Automation bias is defined as placing too much weight on automated advice while decision-making. These or other types of automation error can lead to a decrease in productivity for a task involving automation (Parasuraman, Molloy, & Singh; 1993).

More recent researchers have suggested that too little automation bias and complacency are also errors (e.g., Bahner, Huper, & Manzey, 2008). Bahner, Huper, and Manzey argued that "complacency does not only involve a disadvantage. In contrast, it can clearly add to the performance gains provided by automated aids, at least insofar as these aids work correctly" (2008, p.696). This should be unsurprising because the benefit of automation comes in part from the freeing of cognitive resources to deal with other tasks. For example, consider users who opt to complete a task without using automated

assistance in any way. Their levels of automation complacency and bias would effectively be at zero. Instead of displaying automation complacency and bias by placing too much weight on the recommendations of the automated system, these hypothetical users are outright dismissing all recommendations as irrelevant. Comparing these suspicious users to those who take only the absolute safest and most certain recommendations, the suspicious users will be outperformed due to spending comparatively more cognitive effort on unnecessary tasks. Extending the logic, users who take only the absolute safest recommendations may be outperformed by other users who take twice as many, depending on the distribution of confidence levels within the automated system's recommendations. This line of reasoning would continue until at some point the disadvantages of taking risky advice outweighs the productivity benefits of selectively applying a user's cognitive effort. Thus, there is an optimal level of automation complacency and bias that maximizes productivity while taking into account the risks and benefits of the results of each error.

Types of Automation Error. Researchers have separated automation error and its constituent behaviors into a variety of categories under multiple frameworks. The two most common are the reliance/compliance and use/misuse/disuse/abuse frameworks (Dixon & Wickens, 2006; Parasuraman & Riley, 1997). Different fields within automation research have used different frameworks (e.g., Geels-Blair, Rice, & Schwark, 2013; Merrit & Ilgen, 2008), and both have distinct advantages. However, if researchers want to understand and integrate results from these different fields, then it would be beneficial to reconcile these two frameworks.

Reliance and Compliance. Drawing upon signal-detection theory, the reliance and compliance framework is concerned with categorizing action from a user in either the presence or absence of a signal from the automated system (Dixon & Wickens, 2006). Reliance refers to any case in which a user acts in the absence of a signal from an automated system. Compliance refers to any case in which a user acts in the presence of a signal from an automated system. Both categories can be further subdivided into appropriate or inappropriate depending on whether this activity from the user resulted in an error. For example, appropriate reliance refers to cases in which a user correctly assumed that the automated system was functioning without human intervention. Alternatively, inappropriate compliance refers to cases in which a user acts upon an erroneous prompt from an automated system.

Use, Misuse, Disuse, and Abuse. Parasuraman and Riley focused instead on the potential for a user to disagree with signals provided by a system (1997). Use refers to instances in which a user accepts and acts upon correct advice from an automated system. Misuse refers to cases in which a user accepts and acts upon incorrect advice from an automated system. Disuse refers to instances in which a user correctly disagrees with (and so fails to act upon) incorrect advice from an automated system. Also, Parasuraman and Riley defined disuse to cover situations in which a user actively avoids eliciting aid from the automated system, but in a later literature review researchers found the first definition to be most used (Hoff & Bashir, 2015). However, Parasuraman and Riley then break the pattern in their definition of automation abuse. For their definition of automation abuse, the behavior refers to actions of the engineer of the system rather than the user. Parasuraman and Riley defined automation abuse as cases in which the

engineer of an automated system fails to account for a user's potentially maladaptive behavior. Instead of using this final definition for automation abuse, some researchers have defined automation abuse by fitting the term into the missing case that would finish Parasuraman and Riley's pattern. When fit into this pattern, automation abuse is defined as instances in which a user incorrectly disagrees with correct advice from an automated system (Hoff & Bashir, 2013). I will use this newer definition (i.e., incorrectly agreeing with correct advice) for automation abuse, along with Parasuraman and Riley's (1997) original definitions for use, misuse, and disuse.

Reconciliation. Figure 1 shows the reconciliation of these two frameworks. Whereas primarily I will be using the terms for use, misuse, disuse, and abuse, I will refer to reliance and compliance when appropriate. Note that the example behaviors discussed earlier, automation complacency and automation bias, now fit clearly into this framework under misuse. Automation complacency refers to cases of inappropriate reliance whereas automation bias refers to cases of inappropriate compliance.

Figure 1

Reconciliation of Automation Error Frameworks

	"Emergency warning" given		No "emergency warning" given	
	Emergency present	No emergency	Emergency present	No emergency
User acts upon emergency	**Use** (Appropriate Compliance)	**Misuse** (Inappropriate Compliance & Automation Bias)	**Disuse**	**Abuse**
User does not act	**Abuse**	**Disuse**	**Misuse** (Inappropriate Reliance & Automation Complacency)	**Use** (Appropriate Reliance)

Trust

Researchers have identified automation trust as a variable of interest that may be used to predict automation errors (e.g., Hancock, Billings, Schaefer, Chen, De Visser, & Parasuraman, 2011; Madhavan & Wiegmann, 2007; Schaefer, Chen, Szalma, & Hancock, 2016). The most frequently used definition for automation trust comes from Lee and See (2004, p. 54): "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Lee and See derived their definition from the definition of interpersonal trust put forth by Mayer, Davis, and Schoorman (1995, p. 712): "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other part."

Higher levels of trust are associated with higher levels of use and misuse, i.e., agreeing with an automated system regardless as to whether it is correct or incorrect (Hancock et. al, 2011). In particular, trust predicts automation complacency and bias, both of which are referred to as misuse (Parasuraman & Manzey, 2010). Trust is used as the main predictor of automation errors, and there exists an optimal level of trust related to maximum productivity. If trust is too high, then automation bias and complacency rise above optimal levels, and users rely on automation in situations in which it is not correct to do so. If trust is too low, then users fail to capitalize on the cognitive benefits that could be gained from offloading some of the work onto the automation.

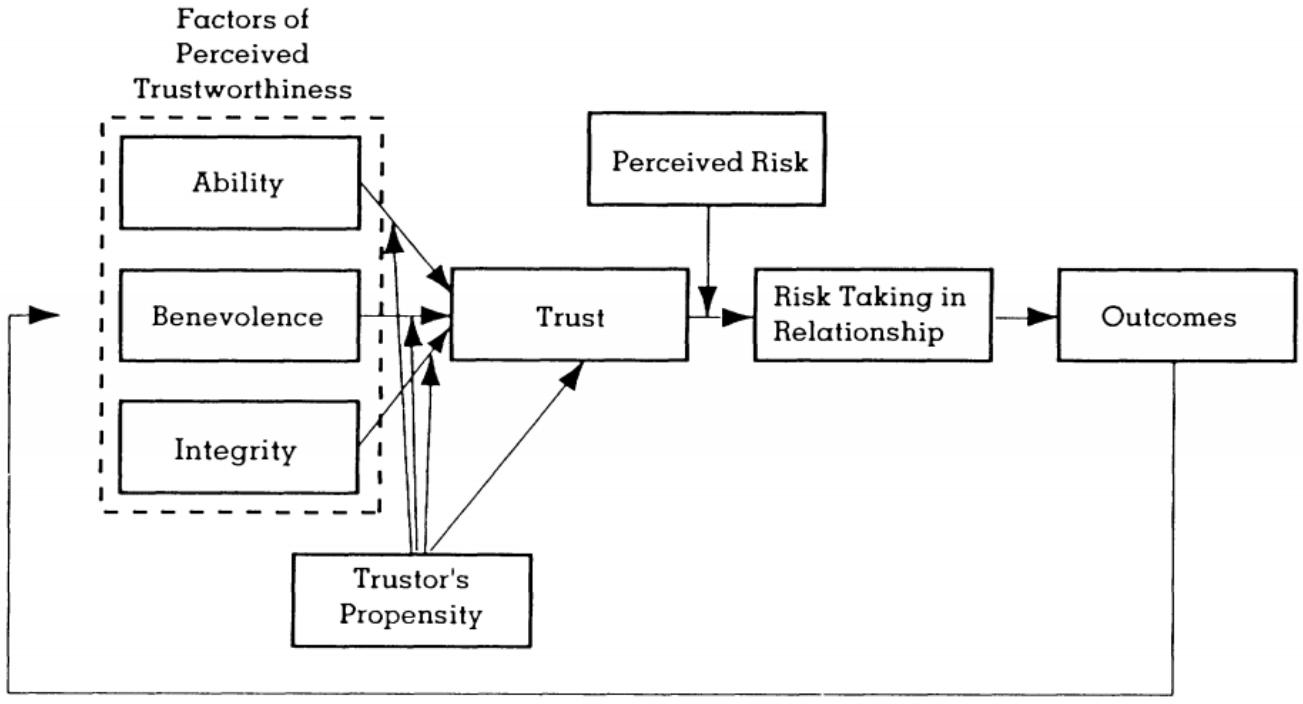
Types of Trust. Often, researchers have subdivided automation trust into a number of constituent factors (Hoff & Bashir, 2015). These subdivisions consist of dispositional trust, situational trust, and learned trust. Whereas the current study focuses on learned trust, the other categories bear mention for the sake of understanding the upcoming models.

Interpersonal Trust. Before I can discuss how these categories relate to the current study, I must first briefly discuss interpersonal trust. Many models of automation trust are derived from interpersonal trust models (e.g., Lee & See, 2004). Prior to 1995, there was a proliferation of trust conceptualizations with 157 definitions and dozens of models proposed in a single 30-year span (Moyano, Fernandez-Gago, & Lopez, 2012). Eventually, researchers came to something of a consensus of support regarding Mayer and Davis's (1995) Integrative Model of Trust. They defined trust as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to

monitor or control that other part." Mayer and Davis's (1995) model can be seen in Figure 2. Their trust model describes factors of perceived trustworthiness interacting with a propensity to trust to influence trust, which influences perceived risk and in turn outcomes, which then loop back into affecting the factors of perceived trustworthiness. Though Mayer and Davis did not differentiate between malleable and stable trust, they did specifically mention that propensity to trust was a relatively stable trait.

Figure 2

Mayer and Davis's (1995) Interpersonal Trust Model



Whereas many of the features of Mayer and Davis's (1995) model map onto automation trust, there are enough examples of divergence that researchers have developed trust models specific to automation trust (e.g., Ghazizzadeh, Lee, & Boyle,

2012; Hancock, Billings, Schaefer, Chen, De Visser, & Parasuraman, 2011; Hoff & Bashir, 2015; Lee & See, 2004). The simplest example of a divergence relates to factors of perceived trustworthiness, which in Mayer's interpersonal model include ability, benevolence, and integrity. Concepts such as benevolence and integrity on the part of the trustee do not have obvious analogues within the context of automation, and researchers would have to address the discrepancy were they to use Mayer and Davis' (1995) interpersonal model to describe interaction with automated systems. In addition, many properties unique to automation trust are not included within Mayer's model, such as situational trust, i.e., the tendency for users to rely on extraneous indicators of automation trustworthiness such as familiarity with the domain in which that automated system is deployed. Despite these discrepancies, I highlight Mayer's interpersonal trust model here to show the critical feedback loop between outcomes and factors of perceived trustworthiness. It is this aspect of Mayer's model that researchers later developed into the concept of learned trust (e.g., Lee & See, 2004). In addition, Mayer's propensity to trust construct was adapted into the concept of dispositional trust (Marsh & Dibben, 2003).

Dispositional Trust. In the automation trust literature, dispositional trust refers to the overall willingness to trust any unknown automated system (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Hoff & Bashir, 2015). Reinterpreted, this definition refers to the trust that a user has in automation in general. Dispositional trust in automation is defined similarly to and has many of the same properties as interpersonal propensity to trust within Mayer's 1995 trust model. Most notably, both refer to the relatively stable trait component of trust within their respective domains. In addition,

cognitive ability correlates positively with both interpersonal propensity to trust and dispositional trust in automation (see Juvina et al., 2019 for a review of literature and a modeling argument). Also, researchers have uncovered many personality factors that correlate with dispositional trust, including extraversion and openness (Merritt & Ilgen, 2008). Other correlates include age, gender, and culture (Hoff & Bashir, 2015).

Situational Trust. Situational trust refers to the trust that a user has in automation in a particular situation (Hoff & Bashir, 2015). This does not refer to familiarity with a specific automated system the user will operate but rather extraneous properties expected to be involved in that system's use. Basically, situational trust is domain-specific trust. For example, users familiar with farm equipment may trust automated systems present in a tractor more than they trust automated banking equipment (Madhavan, Wiegmann, & Lacson, 2006). Several situational factors (i.e., factors other than properties of the automation) inform a user's situational trust level. Hoff and Bashir (2015) separated situational trust into external and internal factors, with internal factors describing user states such as expertise and external factors describing situational states such as perceived risk. The internal factors are self-confidence, subject matter expertise, mood, and attentional capacity. The external factors are system type, system complexity, task difficulty, workload, perceived risks, perceived benefits, organizational setting, and task framing.

Learned Trust. Learned trust refers to malleable aspects of automation trust that change over the course of interacting with a single, specific automated system (Hoff & Bashir, 2015). In one of the few studies on learned trust, Desai, Kaniarasu, Medvedev, Steinfeld, and Yanco (2013) observed results similar to those found in research on

interpersonal trust. Desai et al. (2013) found that timing of both 1) breaking interpersonal trust and 2) automation failures affect trust assessment. As in interpersonal trust, early breaks in trust and automation failures lead to larger decreases in learned trust (see also Juvina et al., 2019). The current study will focus on this relatively unexamined area of the research.

Of the three types of trust addressed in the automation trust literature, learned trust is the least studied (French, Duenser, & Heathcote, 2018). There are several measures of dispositional trust, a few measures of situational trust, and *no* measures of learned trust. Trust is a combination of a user's dispositional, situational, and learned trust (Hoff & Bashir, 2015). Because dispositional trust is a relatively stable construct over time and situational trust should remain stable within a single situation, I expect any changes to overall trust levels within a single experiment to reflect the effects of learned trust.

Users calibrate their trust level to match the reliability of an automated system through the mechanism of learned trust (Ghazizadeh, Lee, & Boyle, 2012). However, no automated system is perfectly reliable and automation failures are at least partially random. At least some of the time, a set of untimely errors could cause learned trust to be miscalibrated, with a user expecting higher or lower levels of automation reliability than is correct for a given system. In addition, there are cases in which the naturally occurring stable level of trust is not optimal because it does not take into account risk levels (Mosier, Skitka, Heers, & Burdick, 1998). For example, a pilot may be dealing with an automated system that has a failure rate of less than .0001% and never see a failure.

However, due to the potentially catastrophic risk inherent in any individual failure that pilot must learn to check the system as if it has a much higher chance of failing.

However, the mechanism through which learned trust occurs is largely unstudied, and attempts to alter learned trust have been often unsuccessful and followed by a subsequent regression toward the natural trust level (e.g., Bisantz & Seong, 2001; Skitka, Mosier, & Burdick, 1999). Some tested interventions include training regarding automation properties and introducing intentional failures. In all cases, the resulting decrease to learned trust lasted only in the short term. Subsequent repeated successful interactions with an automated system raised the user's trust levels to their original positions.

Automation Trust Models. Researchers have developed numerous automation trust models (see French, Duenser, & Heathcote, 2018 for a review).

Correlate Categorization Models. Of the models that describe automation trust, most of them focus primarily on grouping known correlates of automation trust and categorizing them (e.g., Hancock, Billings, Schaefer, Chen, De Visser, & Parasuraman, 2011; Hoff & Bashir, 2015; Lee & See, 2004). This is primarily useful in studying dispositional trust. Because the focus of this study is on the actual process of trust formation and the nature of learned trust, these models are not relevant here. The most widely cited of these is Lee and See's 2004 automation model. Whereas Lee and See are generally cited for their definition of automation trust, their model is not useful to me due to being unclear, untestable, and largely irrelevant to the current study.

Muir's (1994) Trust Model. Muir's (1994) automation trust model describes a set of referents influencing behavior, which influences a mental model of the automation,

which then influences trust (which may lead to feedback into the referents if trust is not calibrated), which then influences behavior. Muir published her model one year prior to Mayer and Davis's (1995) interpersonal trust model, yet the models seem very similar in many regards. They both show trustor perceptions being influenced by propensity to trust, which influences behavior and then feeds back into perceptions. In addition, Muir's model shows a number of modifications that appear in more recent automation trust models (Ghazizadeh, Lee, & Boyle, 2012; Lee & See, 2004). The main differences between Muir's (1994) trust model and Mayer and Davis's (1995) interpersonal trust model are where the feedback mechanism is started, the inclusion of a behavior variable between the initial factors and the trust component, and an inclusion of a mental model of the user. Because Muir's background is in computer science, her model differs from those found in human factors research in both its design and vocabulary. Muir's model is appropriate for my study, but I will avoid it because of researchers' lack of familiarity with it and a resulting lack of validation studies examining it.

The Exploration-Exploitation Model. Some research has addressed mechanisms through which people learn trust or increase levels of learned trust–one of which is research related to the multi-armed bandit problem (Berry & Fristedt, 1985). Researchers have identified a particular pattern of behavior in their analysis of the multi-armed bandit problem within which intentional tests fit (Berry & Fristedt, 1985). In the multi-armed bandit problem, a hypothetical participant is attempting to maximize the payout from an unfamiliar set of slot machines (i.e., multiple "one-armed bandits"). Participants tend to proceed through the two stages of exploration and exploitation (Audibert, Munos, & Szepesvári, 2009). In the exploration phase, participants sacrifice immediate payout in

order to learn more about the slot machines. In the exploitation phase, those participants use what they have learned to exploit the system for their own gain. Intentional tests are one of the methods that users might utilize to learn more about certain types of systems during the exploration phase.

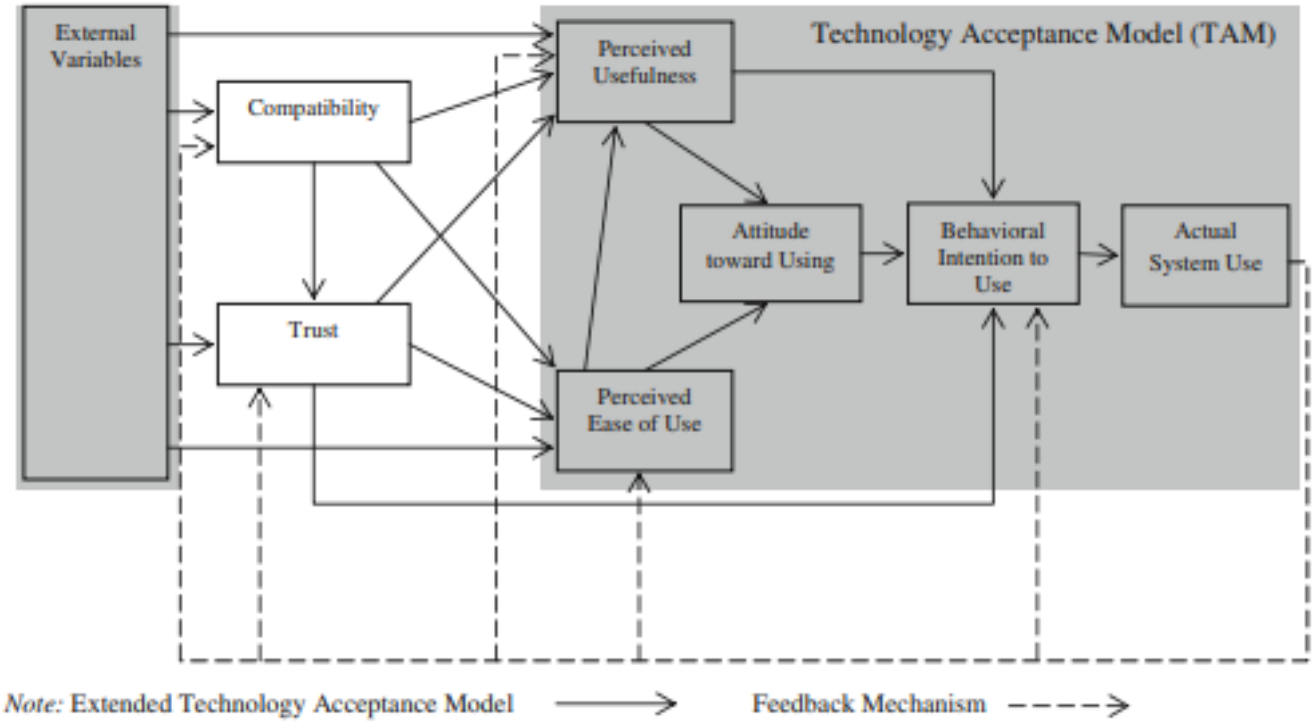
However, the exploration-exploitation model was designed for use in probability theory and machine learning applications (Farias, Vivek, & Ritesh Madan, 2011). Researchers used this model to find the optimal ratio of time spent exploring an unfamiliar system to time spent exploiting it. The model does not address the mechanism by which a user may explore or exploit the system nor does it address how these mechanisms may influence the trust level of a user. Because the exploration-exploitation model focuses primarily on the method by which a user may maximize earnings rather than the mechanism of exploration, I will not be using it as the basis of my research.

The Automation Acceptance Model. The Automation Acceptance Model has implications for learned trust (Ghazizzadeh, Lee, & Boyle, 2012). Before I can address the Automation Acceptance Model, I must address its predecessor: the Technology Acceptance Model (Davis et al., 1989). The Technology Acceptance Model describes the mechanism by which an individual adopts and uses an unfamiliar technology (Davis et al., 1989). This model points to external variables that lead to perceived usefulness/ease of use, which influences a user's attitude towards utilizing the technology, which leads to a behavioral intention to use, which leads to actual system use. Researchers have adapted the Technology Acceptance Model (Davis et al., 1989) for use in a number of disparate fields. Researchers have adapted the model for use in marketing, management, computer science, and human factors psychology (e.g., King & He, 2006). Ghazizadeh, Lee, and

Boyle (2012) extended the Technology Acceptance Model (Davis, Bagozzi, & Warchas, 1989) to include a trust, compatibility, and feedback loop to create the Automation Acceptance Model (Figure 3). In Figure 3, note that the gray area refers to the Technology Acceptance Model, and the areas outside the gray area were added by Ghazizadeh, Lee, and Boyle (2012) to create the Automation Acceptance Model.

Figure 3

Ghazizadeh, Lee, and Boyle's (2012) Automation Acceptance Model



One main disadvantage of the Automation Acceptance Model model is that it addresses dispositional and situational trust as a single construct. Whereas the adaptation of this model for the use in the automation trust literature limits its use in any study examining dispositional versus situational trust, the detail included within its feedback

loop is relevant to my research. Specifically, the Automation Acceptance Model "shows acceptance as a dynamic bidirectional process [between the user and the automated system] rather than a static single-directional process" (Ghazizzadeh, Lee, & Boyle, 2012, p.45). A static single-directional process in this case would refer to the processes in which individual's level of trust (dispositional and situational) affects system use but not the reverse. A dynamic bidirectional process in this case would refer to the processes of trust (dispositional, situational, and learned) influencing system use and system use influencing trust. Learned trust reflects a bidirectional process in which users utilize dynamic bidirectional processes to calibrate trust levels. The dynamic bidirectional process of learned trust provides feedback to the user about the system.

Feedback Seeking Behavior

There is substantial research on feedback that can inform the current study regarding the bidirectional process of learned trust. The positive effect of feedback on performance has long been identified as one of the most well-studied and dependable effects in psychology (e.g., Chapanis, 1964). However, an implicit assumption of much feedback research (and in particular earlier feedback research) was that individuals were passive recipients of feedback regarding their performance and the environment (see Ilgen, Fisher, & Taylor, 1979 for a review). More recent research on feedback (e.g., Ashford & Cummings, 1983) has posited that individuals actively seek feedback. For example, Ashford and Cummings (1983) hypothesized and found evidence that individuals were active seekers of feedback and engaged in behaviors that optimized the amount of feedback they received from their environment.

Researchers have examined active seeking of information from an unfamiliar system or activity within the unrelated domain of goal-oriented behavior research. Specifically, Ashford and Cummings suggested that individuals can use monitoring or inquiry strategies to actively seek feedback. Monitoring poses fewer interpersonal risks given that monitoring can provide useful feedback without other individuals being aware of the monitoring behavior. In contrast, individuals face greater interpersonal risks if they use the inquiry strategy because inquiry involves direct (overt) requests to obtain feedback from others (e.g., asking your supervisor to evaluate your performance). This research (see Ashford, De Stobbeleir, & Nujella, 2016, for a review) has implications for feedback seeking behavior when dealing with automation. An implicit assumption within the majority of automation research is that individuals are passive recipients of feedback about the system (French, Duenser, & Heathcote, 2018).

Categories of Motivation. Ashford, Blatt, and VandeWalle (2003) identified three motives that underlie feedback seeking. All three motives are relevant in the domain of automation interaction. The three motives are: instrumental, impression management, and ego-protection.

The instrumental motive refers to a motive to actively seek feedback in order to facilitate goal achievement, which would then enhance performance. Relevant to my research, an implication is that one could increase feedback seeking behavior by increasing the importance of goal achievement, for example, by linking rewards to performance. In my research, feedback seeking behavior would involve intentional testing of an automated system.

The impression management motive refers to a motive to actively seek feedback in order to manage how one is perceived by others. The impression management motive is consistent with the focus on impression management in the feedback seeking model proposed by Morrison and Bies (1991). That is, individuals seek to enhance or protect the impressions others hold of them. Relevant to my research, an implication of this motive is that one could increase feedback seeking behavior (i.e., intentional testing) by increasing an individual's awareness that interactions with an automated system are private and not observed by others.

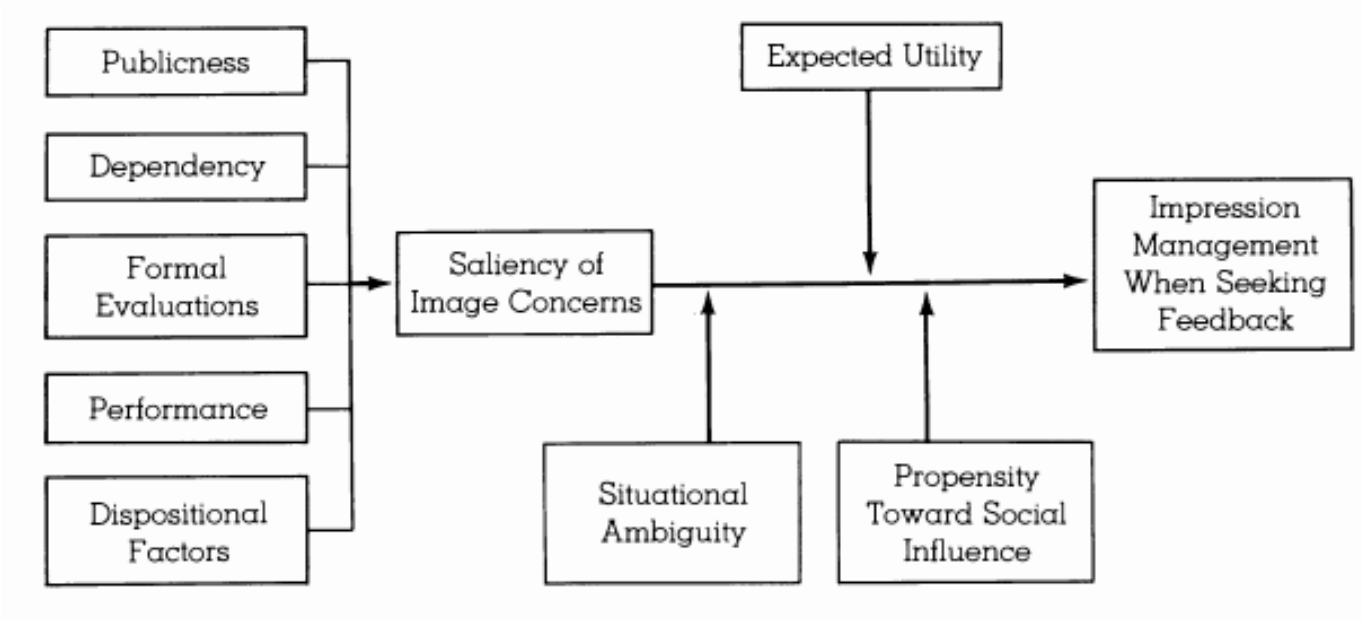
The ego-protection motive refers to the tendency for individuals to interpret feedback in a way that protects their self-image. The ego protection motive is less useful in informing the design of research studying intentional testing of automated systems. However, it does imply that when individuals interpret feedback from an automated system, they are likely to do so in a way that enhances or preserves their self-image. This mirrors the egocentric discounting of others' opinions found in the advice-taking literature (Yaniv & Kleinberger, 2000). Within the domain of intentional testing of automated systems, this implies that the responsibility for the failure of an automated system will be shifted onto the automation rather than upon the user of the automated system. Thus, failures of the automated system in intentional testing scenarios should lead to a decrease in trust of the capabilities of that automated system rather than a belief by participants that they have failed somehow in their use of that system.

Automation Applications. Morrison and Bies (1991) noted that impression management is a key predictor of feedback seeking behavior. Four of the five predictors related to image concerns that might limit feedback seeking behavior in other settings

should increase feedback seeking behavior when dealing with an automated system: publicness, dependency, performance, and dispositional factors (formal evaluations, the fifth predictor, have no analog in automation trust and will thus be ignored; Morrison & Bies, 1991). Figure 4 shows Morrison and Bies's model.

Figure 4

Morrison and Bies's (1991) Model of Factors Leading to Impression Management Behavior in the Feedback Inquiry Process



In the following, I define each of the four predictors relating to impression management that are relevant to human-automation teams. Publicness refers to the tendency for feedback seekers to regulate their feedback seeking behavior depending on the number of observers (Morrison & Bies, 1991). There are often no observers in human-automation teams. Dependency refers to the tendency for users to limit their

feedback seeking behavior if they are dependent on the source of that feedback's continued good impressions of the feedback seeker (Morrison & Bies, 1991). Automation has no such impressions. Performance refers to the tendency for feedback seekers to limit feedback seeking behavior if they feel they are doing poorly and expect this to change the source of the feedback's opinions on the seeker (Morrison & Bies, 1991). Even when performing poorly, feedback seekers should not limit their feedback seeking behavior when dealing with automation. Dispositional factors refer to a tendency for seekers to limit feedback seeking behavior particularly for seekers with high self-monitoring tendencies and higher levels of self-consciousness (Morrison & Bies, 1991). In a human-automation team, self-consciousness should be much less of a concern.

Early Testing. Also, feedback seeking literature points to one of the predictors for individual events of intentional testing within interactions with automated systems. Notably, an individual in a situation of ambiguity should display higher levels of feedback seeking behavior. Ashford and Cummings (1983) suggested that feedback seeking behavior is likely to be lower when individuals are performing routine tasks, including using technology routinely. However, the opposite of this assertion implies that feedback seeking behavior may be greater when performing unfamiliar tasks, including using unfamiliar technology. This implies that intentional testing of automated systems, i.e., feedback seeking behavior, is likely to occur at higher levels in initial stages of working with an unfamiliar automated system.

Previous Examination of Intentional Testing

An unpublished thesis was used to explore the possibility of intentional tests using an in-person sample (Haskins, 2018). It was the exploratory analysis done in this thesis

that led to the research questions in the current study. In addition, the research method used in the current study builds upon that which was done in Haskins's (2018) thesis. Both the current study and the unpublished thesis make use of variations of an X-Ray screening task, and the relative difficulties of the stimuli used in the current study were established by the 2018 thesis results (Haskins, 2018). In Haskins's thesis, participants carried out a version of an X-Ray screening task in which researchers instructed them to identify the presence of weapons. Participants had the option of asking for assistance from an automated system. Haskins (2018) constructed the stimuli used in the image from constituent pieces of the images used in Merrit and Ilgen's (2008) X-Ray screening task though they were assembled into a different set of composite images.

In Haskins's (2018) study, participants were seated alone in a small room with a computer. After a guided training set of three stimuli slides, participants were left to complete a set of 150 stimuli without assistance that had been separated into three blocks with feedback and the option to take a break between blocks. An example of a single stimulus is shown in Figure 5. In order to examine a set of hypotheses related to anthropomorphism, half of the participants had access to an anthropomorphized automated assistant and half had access to a non-anthropomorphized automated assistant.

Figure 5

Haskins (2018) Stimulus Example

Participants' unassisted accuracy was an average of 68% across all stimuli, and the specific accuracy was recorded on a per-stimulus basis to determine which stimuli would be defined as "easy" in the current study. For reference, the example shown in Figure 5 is of a relatively easy stimulus. Researchers did not record response times and confidence levels in this thesis study. However, I conducted an examination of intentional testing using Haskins's (2018) thesis data that defined intentional tests using a simplified definition of "a relatively easy image found in the first block of 50 images in which assistance was requested and a correct answer was given." A single intentional test using this simplified definition in which a correct answer was given by the automated assistant resulted in an average of a .4% increase in subsequent behavioral trust displayed by a participant. A single intentional test using the simplified definition in which an *incorrect*

answer was given by the automated assistant resulted in an average of a 1.6% decrease in subsequent behavioral trust displayed by a participant.

Intentional Testing

To summarize, learned trust in automated systems reflects user reactions to uncertain systems (Marsh & Dibben, 2003). The exploration-exploitation model indicates that users will intentionally explore unknown automated systems (Berry & Fristedt, 1985). The bidirectional process of learned trust described by the Automation Acceptance Model indicates that users take an active part in seeking information about unknown automated systems (Ghazizzadeh, Lee, & Boyle, 2012). Feedback seeking literature describes this process in an interpersonal setting (Ashford & Cummings, 1983). Feedback seeking literature describes a number of predictors of feedback seeking behavior that indicate that this behavior should be common in humans dealing with automated assistants (Morrison & Bies, 1991).

I will define the active seeking of feedback from an automated system with the goal of determining the properties of that system as an *intentional test*. An intentional test involves asking an automated system to solve a problem that the user believes they have already solved. A passed intentional test refers to cases in which the automation responds to a test with what the user believes is correct advice. A failed intentional test refers to cases in which the automation responds with what the user believes is incorrect advice. All requests for system advice that are not intentional tests will be referred to as *genuine requests.*

Situational Factors relating to Intentional Testing. Ashford and Cummings (1983) expected that situations that included increased uncertainty would be associated

with an increased number of intentional tests (Ashford & Cummings, 1983). Also, situations that included relatively low stimulus difficulty were expected to be associated with an increased number of intentional tests. The factors of uncertainty and stimulus difficulty informed my first three hypotheses.

Situational Factors: Uncertainty. A user might be less certain either because the task itself involves some amount of inherent uncertainty or because the user is unfamiliar with a task. In either case, I would expect to see higher levels of feedback seeking, i.e., the use of intentional tests. In the former case, uncertainty refers to either uncertainty regarding an inherently opaque system or uncertainty regarding an unfamiliar automated system or both. An example of an inherently opaque system is a slot machine, as opposed to a relatively clear ATM that a user may simply be unfamiliar with. In an opaque system, uncertainty could be controlled by manipulating user knowledge about the system. In an unfamiliar system, increased uncertainty would be present in earlier trials. There is less need for feedback seeking (i.e., intentional tests) when users are familiar with a task and understand the nature of the stimuli.

Hypothesis 1: Individuals in an uncertain condition will conduct a higher number of intentional tests than individuals in a certain condition.

Hypothesis 2: There will be a greater number of intentional tests in earlier trials.

Situational Factors: Stimulus Difficulty. Stimulus difficulty may be related to the rate of intentional testing. If the task to be accomplished is simple, there would be less need for automated decision support aid. If the task is difficult, due to stimuli difficulty or pace, there is a greater need for an automated decision support aid. In tasks with one or both of these features, one would expect to see the need for automation to

increase, and I would expect to see intentional tests as part of the overall usage behavior. As part of this behavior, users should conduct intentional tests on specific stimuli that they are confident in so they can assess the advice given by the system. Thus, individuals are much more likely to conduct an intentional test for a relatively easier stimulus.

Hypothesis 3: There will be a greater number of intentional tests conducted on easier stimuli.

Personality and Cognitive Factors Related to Intentional Testing. In the preceding sections, I addressed two situational factors. However, in addition to situational factors, researchers studying personality and cognition also have addressed potential antecedents of intentional tests. Many of the personality and cognitive antecedents of feedback seeking behavior may be useful in identifying individuals predisposed to intentional testing behavior (Ashford, De Stobbeleir, & Nujella, 2016). I focused on one of these factors (interpersonal trust) at length above. The feedback literature has found support for many of the predictors of feedback seeking behavior. Two dispositional variables that predict feedback seeking behavior are learning goal orientation and a high level of openness to experience (Krasman, 2010; Van der Rijt, Van de Wiel, Van de Bossche, Segers, & Gijselaers, 2012). Also, Krasman (2010) found that extraversion, conscientiousness, and neuroticism predicted feedback seeking behavior.

Other factors that might play a role in intentional testing are overall cognitive ability, an innovative cognitive style, and a high level of feedback orientation (De Stobbeleir, Ashford, & Buyens, 2011, Linderbaum & Levy, 2010). It is possible that higher cognitive ability may decrease the period in which individuals with higher cognitive ability feel unfamiliar with a new automated system, and this effect could be

somewhat explained by the overall tendency of such individuals to seek feedback at a higher rate in general. Though supporting such an assertion is outside the scope of the current study, it is possible that these effects are linked. That is, increased levels of feedback seeking behavior (and thus intentional testing) on the part of individuals with high cognitive ability might be related to increased speed in individuals becoming familiar with new automated systems. Also, the serial position effect (i.e., increased recall of items at the beginning and end of lists) would suggest that people remember more from earlier interactions, and the recency and primacy effects might play a role, but these are beyond the scope of this study. There are other factors that may play a role as well that also are beyond the scope of this study. This study is not attempting to identify the dispositional factors that predict the likelihood of conducting an intentional test. Instead, this study is focused on the situational factors that might be manipulated by a design team to artificially increase or decrease the rate of intentional tests and what cognitive measures might be used to assess intentional tests.

Mental Chronometry. Mental chronometry from cognitive psychology might increase researchers' understanding of when and why intentional testing behavior on new automated systems might occur, and mental chronometry is a focus in my study. In general terms, mental chronometry refers to measurement of response times for the purpose of predicting the amount of time that a given cognitive task takes (Wong, Haith, & Krakauer, 2015). Measurement of response times allows the study of intentional testing to move from a) predicting the effects of tests after they occur and the likelihood of a test having been previously administered to b) predicting the likelihood that any particular current interaction is actually an instance of an intentional test. This is a key

issue if engineers are to include within their designs a model of intentional tests that predicts potential interactions. With predictive power, engineers could flag specific behaviors as examples of a user intentionally testing the system and alter system behavior accordingly.

A potential intentional test of an automated system acting as a decision support aid usually will follow a predictable pattern. First, a stimulus is presented. Next, a user takes time to consider their answer and whether they will request help from the automation. Then, the user requests help. After, the user will have some time during which they are considering the advice from the system. Last, the user will submit their finalized response to the stimuli.

There are two potential times that may be measured (assuming the assistance of the automated aid is encapsulated within a discrete request). When dealing with an automated decision support aid there are two response times that may be measured. The first is the block of time after a stimulus is presented but before a request for assistance is requested (i.e., “pre-request time”). The second is the block of time after the system has given its advice but before the user has submitted their final answer (i.e., “post-request time”).

Both the decision to test the automated system (part of the pre-request time) and the actual mental process required for the task itself (post-request time) are expected to add to response times differently depending on whether the user is conducting an intentional test. A user conducting an intentional test should have a longer pre-request time due to the additional mental processing required to decide whether to conduct an intentional test. A user conducting an intentional test should have a shorter post-request

response time due to already being confident that they have sufficiently addressed the stimulus prior to requesting help.

Problems with Response Times. Issues inherent to response time use must be addressed. Using response times to predict individual events is usually a gamble, due to the considerable variability within any individual's response times (Whelan, 2008). However, most of that variability comes from momentary attentional lapses, and attentional lapses always *increase* the overall time taken to complete a specific task. Effects that significantly decrease the time taken to complete a task, such as "flashes of insight," are both rare and insufficient to lower the added response time to zero (Bowden, Jung-Beeman, Fleck, & Kounios, 2005). If a participant takes longer than expected, then this could be the result of attentional lapses. However, if a participant takes significantly less time than expected, then this is likely the result of that participant having carried out fewer mental processes. The only other explanations for a shorter response time would be insufficient effort responding and extremely high levels of trust. In this study, I will address insufficient effort responding through a variable payout structure. However, a variable payout structure will not eliminate insufficient effort responding completely, and thus measuring pre-request response time comes with increased error.

When measuring pre-request response times it would be impossible to differentiate between cases in which a user took a potentially negligible amount of time to decide to conduct an intentional test and cases in which the user had an attentional lapse. As a result, I will not be measuring pre-request response times.

When measuring post-request response times, there are few explanations for why a user may have taken very little time to submit their final answer. If the user is engaging

in insufficient effort responding, then they may be simply accepting the system advice immediately. To minimize the effects of insufficient effort responding, the current study will use a variable payout structure. The other explanation for a user having a very short post-request response time is that they are conducting an intentional test. Because the user is already confident in their answer prior to requesting help, an intentional test of an automated system should be related to a relatively small post-request response time.

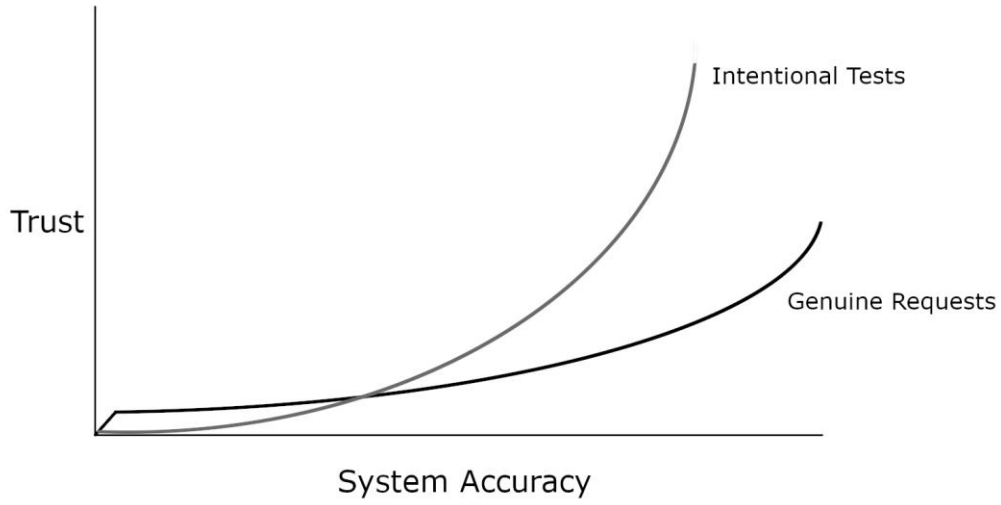
Hypothesis 4: Post-request response times will be shorter for intentional tests than for genuine requests for help.

Intentional Test Effects. One would expect that intentional tests play a different role than genuine requests in changes to learned trust. In genuine requests for help, the user is unsure of the answer and is implicitly trusting the quality of the advice provided by the automation. In contrast, in the case of an intentional test, the user is sure of the answer and is asking for advice from the automation as a means of assessing the quality of the advice provided. Thus, the purpose of an intentional test is to obtain feedback regarding the reliability of that system. The user specifically sought out this feedback to assess the reliability of that system. Furthermore, the user's confidence in the accuracy of this feedback should be higher due to the user's confidence in their own assessment of the correct answer to a stimulus prior to conducting an intentional test.

Hypothesis 5: Trust will be more strongly related to system accuracy during intentional tests than system accuracy during all other interactions (i.e., genuine requests; see Figure 6).

Figure 6

Predicted Relationship



Method

Participants and Design

Participants were randomly assigned to either the certain or uncertain condition. The certainty/uncertainty induction was administered using an online task tutorial (described in the task description section below). All participants completed 150 trials of an online task in 30 blocks of 5 trials each. Following each block, participants were provided with feedback on their accuracy. I recruited participants using Amazon's Mechanical Turk. I expected recruiting from Mechanical Turk to increase the level of intentional testing behavior by minimizing the influence of impression management, thus

maximizing the number of instances of intentional testing for study (Morrison & Bies, 1991). My Mechanical Turk sample size was 300, which was a sufficient number of participants to obtain a power level of .8 to detect an interaction effect. Given the size of this sample, I expected similar distributions on demographic characteristics between the two conditions. However, I verified that the participants across the two conditions were roughly equivalent in terms of demographics and traits measured by the pre-task survey. The only required limitation upon the sample was that all participants must speak English. The average age of the Mechanical Turk sample was expected to be 35 years, with a range of 18-77 years (Burnam & Piedmont, 2018). The Mechanical Turk sample was expected to consist of 55% women and is expected to be made up of 8% Black participants, 6% Asian, 77% Caucasian, 5% Hispanic, and 4% from other ethnicities.

Procedure and Task Description

Procedure

Individuals volunteering to participate in this online study were pre-screened for eligibility, and those who were unable to speak English were excluded from participating. MTurk includes a setting that screens out non-English speakers. Eligible participants completed a consent process (see Appendix A) and a pre-task survey. The pre-task survey included measures of openness, learning goal orientation, feedback orientation, automation trust, and cognitive ability. These measures are described below. Also, I informed participants that their data would be removed and they would not be compensated if they did not work to answer diligently and honestly. I embedded two measures intended to combat insufficient effort responding in the pre-task survey. The first was a set of three forced-response items. An example item is “for this question,

indicate that you neither agree nor disagree with this prompt." Participants who answered any of these incorrectly had their data removed. The second insufficient effort response check was a time measure. Participants who took an average of fewer than 2 seconds per item on any particular page of the survey had their results removed from the dataset.

Following the pre-task survey, participants were instructed to download the X-Ray Screening Task program to complete on their personal computer. Then, participants completed an online tutorial, during which they received the certainty/uncertainty induction, and then completed the X-Ray Screening Task. Participants who quit and restarted the program (for any reason) or who took less than 1 second per stimulus for 10 consecutive stimuli had their data rejected. I chose this restriction because one second per image is not enough time to properly assess the stimuli, and answering so quickly for ten consecutive images would require participants to ignore and then have a strategy for rapidly navigating the progress report every 5 trials. Finally, participants emailed the results from their X-Ray Screening Task to an address provided and then were debriefed (see Appendix B).

X-Ray Screening Task, Tutorial, and Certainty/Uncertainty Induction

X-Ray Screening Task. The X-Ray Screening Task is similar to tasks that might be performed by TSA agents, i.e., a simplified version of a luggage screening task. Participants viewed a set of images. Participants reported for each image whether a weapon (specifically, a knife or a gun) was present or absent or indicated they wished to ask for help from the automated assistant. The automated assistant had an accuracy of 80% for all participants. This set of images was the same set of 150 images used in Haskins (2018). The images varied in difficulty. Each image had an accompanying image

difficulty, defined as the percent of participants who accurately determined the presence versus absence of a weapon in that image in the data collected in Haskins (2018). The order of presentation was randomized for each participant, as determined by a random number generator. Weapons were present in 75 of the 150 stimulus images. Figures 7 and 8 show an example of a relatively easy and a relatively difficult stimulus, respectively.

Figure 7

Relatively Easy X-Ray Stimulus (Contains a Weapon)

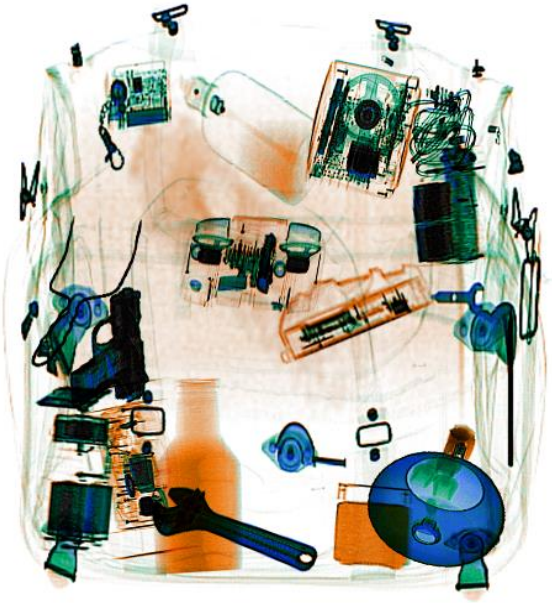
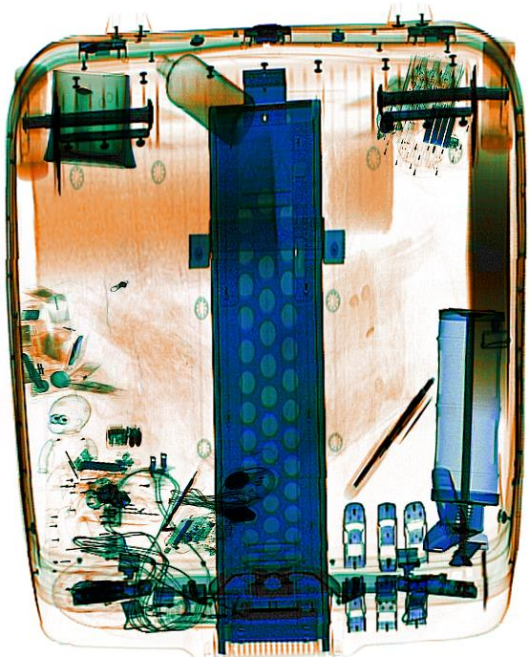


Figure 8

Relatively Difficult X-Ray Stimulus (Contains a Weapon)



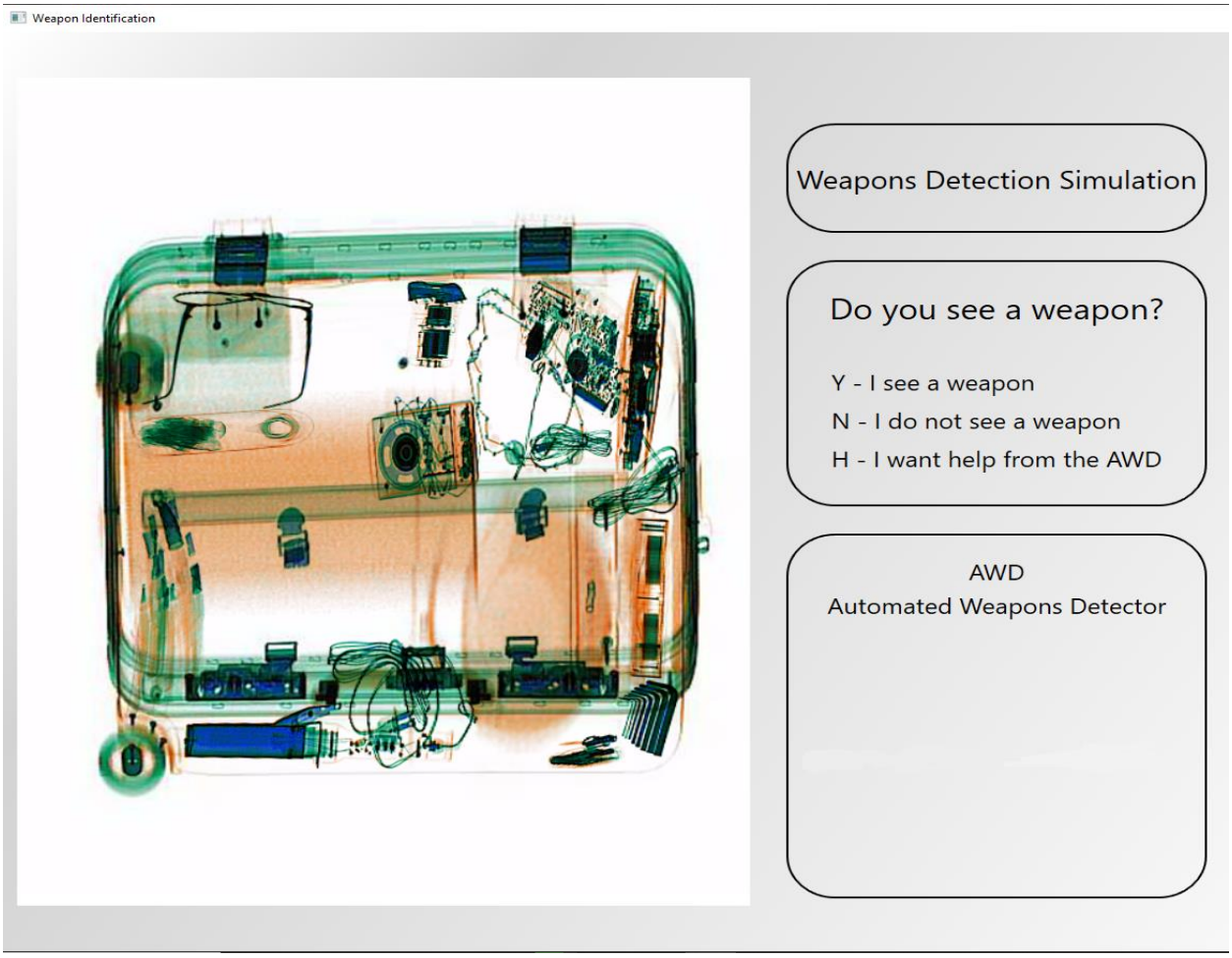
Tutorial and certainty/uncertainty induction. Prior to beginning the screening task, participants completed a brief online tutorial (see Appendix C). Participants in the uncertain condition were told that the automated assistant is not always accurate yet has recommendations that are notably better than chance. Participants in the certain condition were told the system's true accuracy of 80%. Participants in both conditions were told that the automated system's accuracy will remain constant throughout the experiment.

Following the tutorial, participants began the task. Participants were given unlimited time to examine each image to determine whether a weapon is present and pressed "Y" to indicate a weapon was present or "N" to indicate a weapon was not

present. Participants pressed “H” to request help from the automated decision support aid called the AWD (“automated weapons detector”). If a participant pressed “Y” or “N”, s/he proceeded immediately to the next trial. If a participant pressed “H” to request help from the AWD, the participant was taken through a set of steps described below. The three response options were presented adjacent to the luggage image in every trial. Figure 9 shows an example trial of the task.

Figure 9

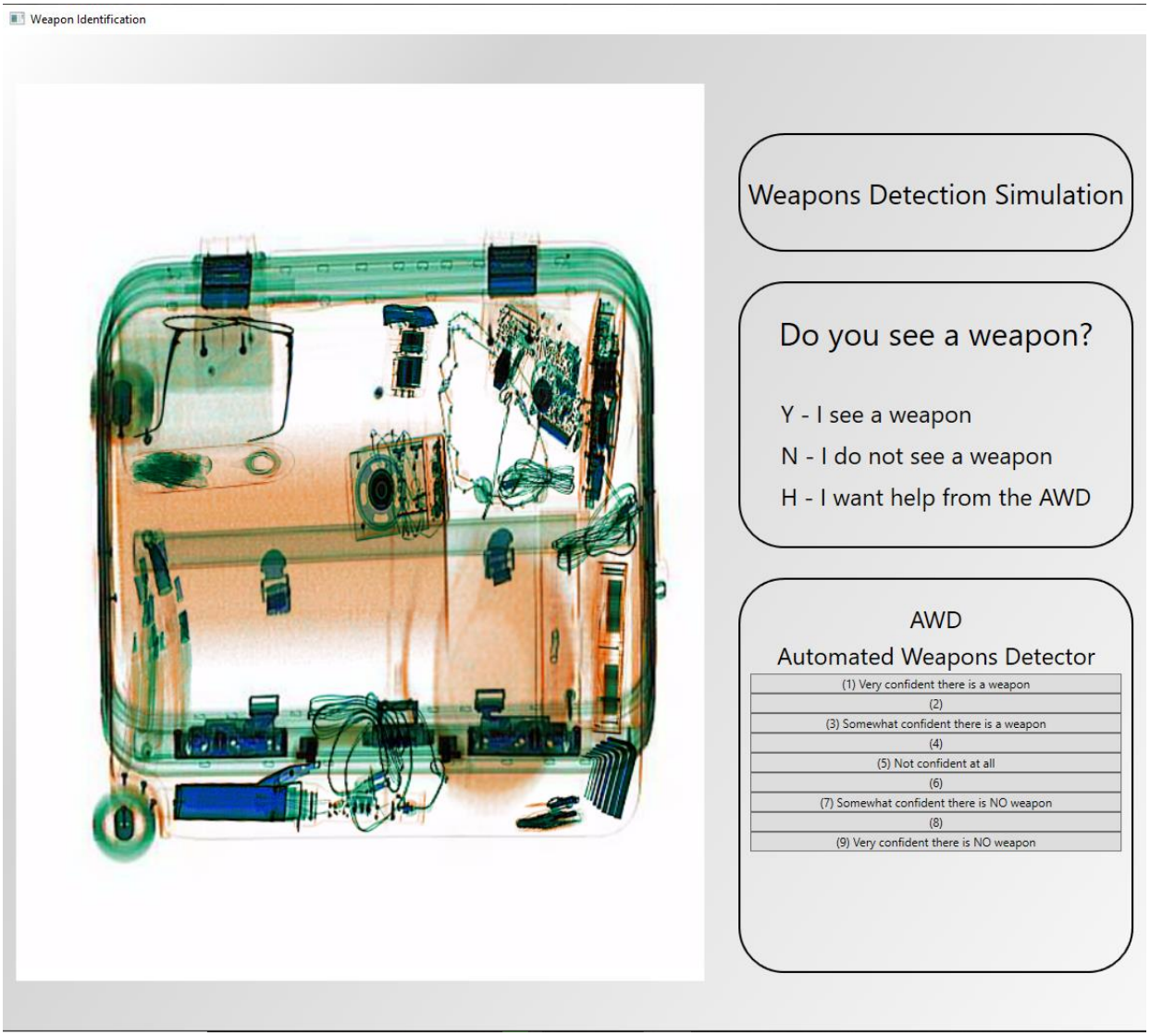
Weapon Search Simulation



If a participant requested help, s/he was first required to answer the question "How confident are you in your current answer?" Participants responded using a 9-point scale with endpoints of "(1) Very confident there is a weapon" to "(9) Very confident there is no weapon" with a midpoint of "(5) Not confident either way." See Figure 10 for an example of the pre-help confidence indication.

Figure 10

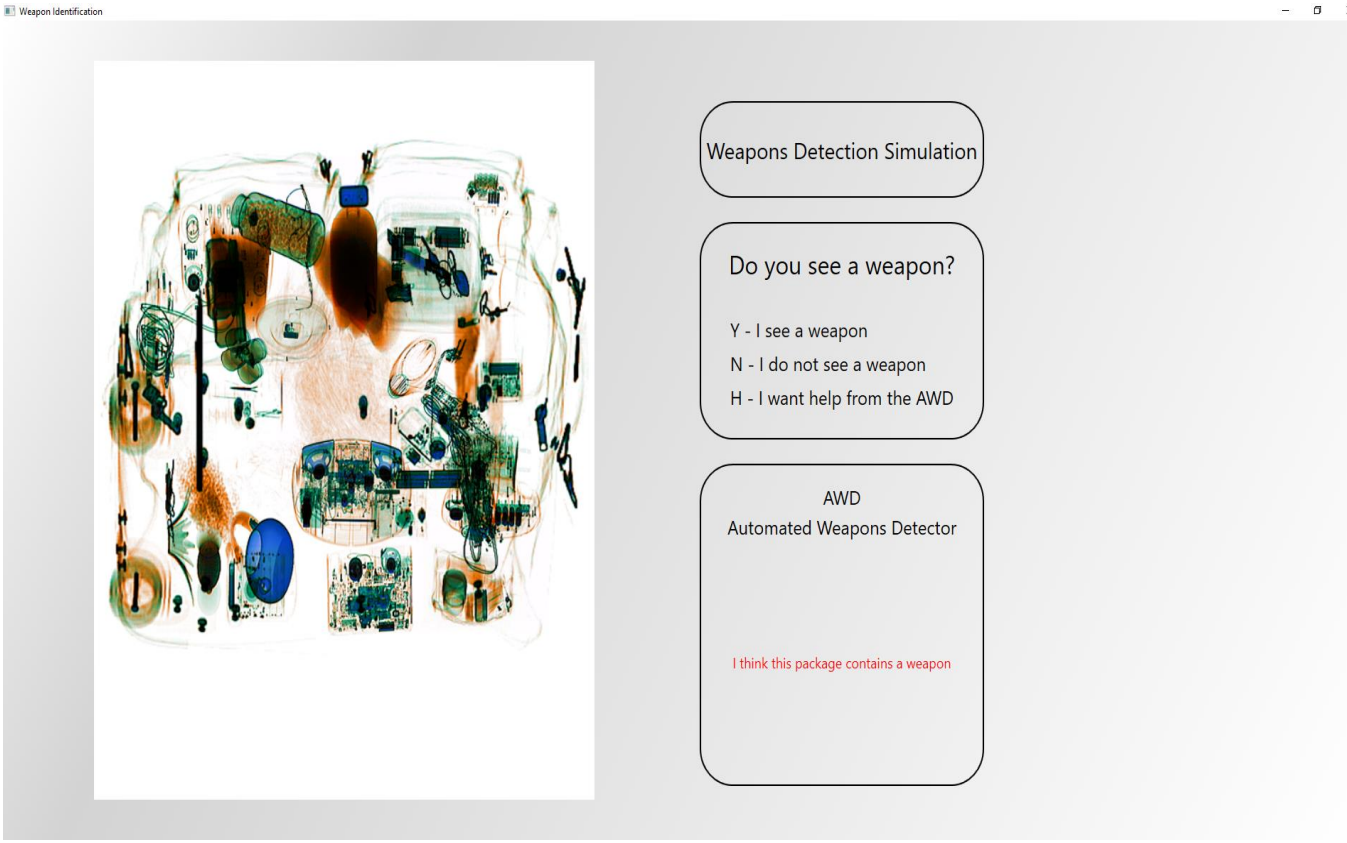
Confidence Indication



After indicating their current confidence level, a 2-second progress bar was shown and then the help from the automated weapons detector was displayed (see Figure 11). This advice stated that “I think this package contains a weapon” in red or “I think this package does not contain a weapon” in green. The cost of requesting help was threefold– it cost 2 seconds of time, it was the only input requiring the use of the mouse or other scrolling tool, and it required cognitive effort to answer the confidence question before help can be requested.

Figure 11

Automated Weapons Detector Advice – Weapon Detected



After every fifth trial, participants received feedback on the accuracy of their responses and told how much money they had received as a result of their performance

(i.e., earning 4 cents per correct answer and losing 15 cents per incorrect answer). I structured feedback this way to increase external validity and prevent perfect measurement by the participant of the accuracy of the automated system (i.e., the accuracy of responses and/or the automated system will not be known on a per-stimulus basis).

Participant Motivation

To increase participant motivation, I gave each participant 50 cents as a reward for participation in this study. In addition, to increase goal-directed behavior, all participants were able to gain by correctly identifying weapons in the images and lose money for errors. Participants earned 4 cents for each image s/he correctly identified as reflecting the presence or absence of a weapon. Participants lost 15 cents for each image s/he incorrectly identified as reflecting the presence or absence of a weapon. Thus, a participant with no correct answers on the 150 trials would earn 50 cents, and a participant with all correct answers would earn $6.50. If the participant asked for and followed the automated advice on each trial, the participant would receive 50 cents (the initial reward) plus $4.80 (= 4 cents * 120 trial) minus $4.50 (-15 cents * 30 trials) for a total of 80 cents.

Measures and Variables

Personality Measures

As part of the pre-task survey, I administered a set of personality measures. These measures were not required to support any of the established hypotheses. I used these measures to assess whether assignment to experimental conditions resulted in groups

similar in relation to gender, race, and the personality traits assessed. Also, I used these measures in exploratory analyses and tests of potential alternative explanations.

Dispositional Trust. To assess dispositional trust for exploratory analyses, I administered the Automation-Induced Complacency Rating Scale (Merritt, Brew, Bryant, Stanley, McKenna, Leone, & Shirase, 2019; see Appendix D). Merritt et al. reported Cronbach's alphas ranging from .77 to .84. Merritt et. al assessed two dimensions of automation-induced complacency: alleviating workload (5 items, $\alpha = .84$) and monitoring (5 items, $\alpha = .77$). Given the nature of the experimental task, the Automation-Induced Complacency Rating Scale was appropriate to use with no revised wording. An example alleviating workload item is "When I have a lot to do, it makes sense to delegate a task to automation." An example monitoring item is "Constantly monitoring an automated system's performance is a waste of time." Participants responded using a five-point scale from 1 (disagree strongly) to 5 (agree strongly), and I averaged these responses for each participant within each scale. I kept these independent scales separate throughout the analysis.

Openness. I administered the 10-item NEO Openness measure from the International Personality Item Pool (Goldberb, 1999; see Appendix E). The reported Cronbach's alpha value is .81. An example item is "I have a vivid imagination." Participants responded using a five-point scale from 1 (disagree strongly) to 5 (agree strongly). I averaged item responses, and a higher score indicated a higher level of openness.

Conscientiousness. I administered the 10-item NEO Conscientiousness measure from the International Personality Item Pool (Goldberg, 1999; see Appendix F). The

reported Cronbach's alpha value was .81. An example item is "I get chores done right away." Participants responded using a five-point scale from 1 (disagree strongly) to 5 (agree strongly). I averaged item responses, and a higher score indicated a higher level of conscientiousness.

Goal Orientation. I administered Vandewalle's (1997) Goal Orientation Measure (see Appendix G). Vandewalle reported Cronbach's alpha values ranging from .65 to .89 and test-retest reliabilities ranging from .57 to .66. Vandewalle assessed three dimensions of goal orientation: learning (5 items, $\alpha = .89$), prove performance (6 items, $\alpha = .85$) and avoid performance (5 items, $\alpha = .88$). Vandewalle validated this measure within a work domain. Given the nature of the experimental task, the Goal Orientation Measure was appropriate to use with no revised wording. An example learning item is "I often read materials related to my work to improve my ability." An example prove performance item is "I prefer to work in situations that require a high level of talent and ability." An example avoid performance item is "I would avoid taking on a new task if there was a chance that I would appear rather incompetent to others." Participants responded using a five-point scale from 1 (disagree strongly) to 5 (agree strongly), and I averaged these responses for each participant within each scale. I kept these independent scales separate throughout the analysis.

Feedback Orientation. I administered Linderbaum and Levy's (2010) Feedback Orientation Scale (see Appendix H). Linderbaum and Levy reported Cronbach's alphas for this scale ranging from .88 to .73, and the test-retest reliability was .69. Linderbaum and Levy assessed four dimensions of feedback orientation: utility (5 items, $\alpha = .88$), accountability (5 items, $\alpha = .73$), social awareness (5 items, $\alpha = .85$), and feedback self-

efficacy (5 items, $\alpha = .78$). Linderbaum and Levy validated this measure in a work domain. Given the nature of the experimental task, the Feedback Orientation Measure was appropriate to use with no revised wording. An example utility item is "feedback contributes to my success at work." An example accountability item is "I hold myself accountable to respond to feedback appropriately." An example social awareness item is "I try to be aware of what other people think of me." An example feedback self-efficacy item is "I know that I can handle the feedback that I receive." Participants responded using a five-point scale from 1 (disagree strongly) to 5 (agree strongly), and I averaged these responses for each participant within each scale. I kept these independent scales separate throughout the analysis.

Cognitive Ability. I administered a subset of items from the Shipley Institute of Living Scale to ensure that cognitive ability was relatively similar across experimental groups (Shipley, 1940). The Shipley Institute of Living Scale has two subscales. The first section has 20 items and contains a fill-in-the-blank abstraction section in which participants are instructed to complete a pattern. The second section is a 40-item multiple-choice section testing vocabulary that instructs the user to select amongst potential synonyms for an uncommon word. I administered five items (i.e., Items 4, 8, 12, 16, and 20) from the abstraction section and five items (i.e., Items 8, 16, 24, 32, and 40) from the vocabulary section (see Appendix I). Bowers and Pantle (1998) found that the correlations between this cognitive test and other cognitive tests ranged from .77 to .83. An example item from the first section gives the prompt "RENOWN," with the answers of "length, head, fame, loyalty." An example item from the second section is "Z, Y, X,

W, V, U, - .” I summed the number of correct answers for each participant and compared the sums across the two experimental conditions to check for homogeneity.

Predictor Variables

Certain/Uncertain Conditions. As described in the task description above, I manipulated certainty versus uncertainty, administering the induction in the online task tutorial. Specifically, participants in the certain condition were told “The AWD’s estimated accuracy is 80%.” Participants in the uncertain condition were told “The AWD’s accuracy is constant and is much better than chance”.

Trial. All participants completed 150 trials of the task, divided into 30 blocks of 5 trials each. Participants received accuracy feedback, i.e., how many stimuli in the previous 5 trials they answered correctly, after every fifth trial. Different trials were used as described below to test different hypotheses.

Stimulus difficulty. As described in the task description above, each stimulus had an identified difficulty. The difficulty of each stimulus was defined as the percentage of participants in responded to that stimulus correctly, i.e., the percentage of participants who correctly stated that a given stimulus did or did not have a weapon present.

Behavioral Measures

Behavioral Trust. I calculated the rates of use, misuse, disuse, and abuse from the data of each participant. See Figure 12 for a mapping of each variable. I described the calculation of each below although I used only use and misuse in my calculation of behavioral trust. This variable of behavioral trust can be generally thought of as all the times in which the user agreed with advice from the automated system regardless as to whether the system was correct. Disuse and abuse, whereas not included in the behavioral

trust measure, were calculated for exploratory purposes. The current measure, behavioral trust, used data from all 150 trials. Note that this is distinct from "later behavioral trust," discussed below, which instead only measures trust from the final 100 trials.

Figure 12

Operational Variable Mapping

	"This package seems to contain a weapon" signal given		"This package does not seem to contain a weapon" signal given	
	Weapon Present	No Weapon Present	Weapon Present	No Weapon Present
Participant selects "I see a weapon"	**Use** (Appropriate Compliance)	**Misuse** (Inappropriate Compliance)	**Disuse**	**Abuse**
Participant selects "I do not see a weapon"	**Abuse**	**Disuse**	**Misuse** (Inappropriate Reliance)	**Use** (Appropriate Reliance)

Specifically, to calculate trust, I summed examples of use and examples of misuse across all trials and then divided by the total number of help requests. Examples of use were calculated as the sum of the number of cases in which, when presented with a stimulus that contained a weapon, the participant responded with "this package contains a weapon" after being advised by the automated assistant that a weapon was present. Added to this sum were the cases in which, when presented with a stimulus that did not contain a weapon, the participant responded with "this package does not contain a weapon" after being advised by the automated assistant that a weapon was not present. Examples of misuse were calculated as the sum of the number of cases in which, when

presented with a stimulus that contained a weapon, the participant responded with “this package does not contain a weapon” after being advised by the automated assistant that a weapon was absent. Added to this sum were the cases in which, when presented with a stimulus that did not contain a weapon, the participant responded with “this package contains a weapon” after being advised by the automated assistant that a weapon was present. Then, I divided the sum of examples of use and example of misuse by the total number of times in which a participant requested help. Note that the total number of help requests could also reflect disuse and abuse.

Below, I describe how I calculated ratios for each of the four responses to help advice (use, misuse, disuse, and abuse) for possible use in exploratory analyses.

First, I calculated use alone as the sum of the number of cases in which, when presented with a stimulus that contained a weapon, the participant responded with “this package contains a weapon” after being advised by the automated assistant that a weapon was present. Added to this sum were the cases in which, when presented with a stimulus that did not contain a weapon, the participant responded with “this package does not contain a weapon” after being advised by the automated assistant that a weapon was not present. Then, I divided this sum by the total number of times in which a participant requested help. I repeated this process for the other three operational variables: misuse, disuse, and abuse.

I calculated misuse as the sum of the number of cases in which, when presented with a stimulus that contained a weapon, the participant responded with “this package does not contain a weapon” after being advised by the automated assistant that a weapon was absent. Added to this sum were the cases in which, when presented with a stimulus

that did not contain a weapon, the participant responded with “this package contains a weapon” after being advised by the automated assistant that a weapon was present. Then, I divided this sum by the total number of times in which a participant requested help.

I calculated disuse as the number of cases in which, when presented with a stimulus that contained a weapon, the participant responded with “this package contains a weapon” after being advised by the automated assistant that a weapon was absent. Added to this sum were the cases in which, when presented with a stimulus that did not contain a weapon, the participant responded with “this package does not contain a weapon” after being advised by the automated assistant that a weapon was present. Then, I divided this sum by the total number of times in which a participant requested help.

I calculated abuse as the number of cases in which, when presented with a stimulus that contained a weapon, the participant responded with “this package does not contain a weapon” after being advised by the automated assistant that a weapon was present. Added to this sum were the cases in which, when presented with a stimulus that did not contain a weapon, the participant responded with “this package contains a weapon” after being advised by the automated assistant that a weapon was absent. Then, I divided this sum by the total number of times in which a participant requested help.

Later Behavioral Trust. Some analyses (specifically, those related to Hypothesis 5) required the differentiation between behavioral trust and later behavioral trust. This variable, later behavioral trust, used only data gathered from the final 100 trials. It was in all other ways identical to behavioral trust.

Intentional Tests. I counted each intentional test conducted by each user across all trials. The current measure, intentional tests, used data from all 150 trials. Note that

this is variable was distinct from "later intentional tests," discussed below, which instead only measured intentional tests conducted during the final 100 trials. An intentional test is defined as an instance in which a user requested help despite already being extremely confident in the presence or absence of a weapon. Selecting a 1 or 9 indicated that the participant was extremely confident in their initial assessment of either the presence or absence of a weapon. All trials in which a participant requested help and selected a 1 or 9 were defined as intentional tests. Then, I subdivided these tests into passed and failed tests for exploratory analyses and as a component to the early system accuracy variable (discussed later). Passed tests referred to all cases in which the system gave advice agreeing with the user's initial confident guess. Failed tests referred to all cases in which the system disagreed with the user's initial confident guess. All trials in which a participant did not conduct an intentional test were referred to as genuine requests.

Early Intentional Tests. Some analyses (specifically, those related to Hypothesis 5) required the differentiation between intentional tests and early intentional tests. This variable, early intentional tests, only used data gathered from the first 50 trials. It was in all other ways identical to the variable "intentional tests."

Response Time. I recorded response times across all trials. This included time spent prior to requesting help, time between requesting help and declaring a response, and time spent analyzing a stimulus in cases in which participants did not request help. Whereas I only used the second of these times in verifying hypotheses, the others will be used in exploratory analysis.

Task Performance. I recorded the overall performance of each user. Whereas I did not use task performance in tests of my hypotheses in the current study, I assessed

task performance for use in exploratory analyses. I defined task performance as the number of trials in which a final correct answer was given by a particular user.

Early System Accuracy Measure

Whereas the overall accuracy of the automated weapons detection system was 80%, a single user may have tested that accuracy via an intentional test only a handful of times throughout the study. A user conducting only 3 intentional tests in the first 50 trials may have seen a system accuracy of 33.3% during those three trials. This accuracy was "early system accuracy during intentional tests." It was calculated by counting the number of passed intentional tests within the first 50 trials and dividing by the total number of intentional tests conducted.

I also calculated the accuracy of the system during all genuine requests (i.e., non-intentional tests) within the first 50 trials. I did this by first counting up all genuine requests within the first 50 trials (expected to be close to 50). Then I counted the number of genuine requests for help that resulted in accurate help being given. I divided the number of accurate responses by the total number of genuine requests for help within the first 50 trials. This final variable was called "early system accuracy during genuine requests." I expected this to be very, very close to 80% for all participants.

Results

Data Cleaning. A total of 6701 participants attempted the study that was posted on Amazon's Mechanical Turk. Of those, 6401 did not complete the study due to difficulty running the experimental program, elimination due to insufficient effort responding, or a failure to submit a final results file. Because I did not receive any information from participants who failed to submit a final results file, there was no way

to differentiate between participants that simply could not get the program to run, participants who failed the IER check, and participants who were unable to submit their final data file for some other reason. In order to be retained, participants had to pass the following insufficient effort responding (IER) checks administered by the experimental program. The first IER check immediately ended the program for any participant who answered in less than 1 second per trial for 20 trials in a row. Given the need to respond to three feedback panels fairly quickly, this was judged to be impossible to achieve during any normal usage behavior. The second insufficient effort responding check immediately ended the program if a participant responded with the same answer for 30 consecutive trials and had a task performance for those trials of less than 60%. In both cases, the program would leave a file on the user's system that prevented them from ever opening the experimental program again. Twelve (4.56%) participants answered incorrectly at least once in the forced-choice answers on the survey, but I did not exclude them from the analysis because they did not have any unusual patterns in their behavioral data. I suspected that the insufficient effort responding checks in the experimental program caught the most problematic participants. Out of the 6,701 participants to complete the survey, 1641 (24.5%) missed at least one forced-choice question. I did not eliminate any participants due to response times on the survey for two reasons. First, the IER present in the experimental program seemed strenuous enough to eliminate most suspicious behaviors. Participants who failed either the forced choice or the one-second-rule did not have any suspicious patterns of response in their later use of the experimental program. Second, the relatively high level of survey-taking expertise present in many mTurk workers suggested that removing participants due to survey item response times

may remove valid data from experienced mTurk workers (Deetlefs, Chylinski, & Ortmann, 2015).

A total of 300 participants remained after removing 6401 participants who did not complete the study due to difficulty running the experimental program, elimination due to insufficient effort responding, or a failure to submit a final results file. Because all three categories of participant appeared identical (i.e., I simply did not receive a file from them), there was no way to subdivide this group further. Next, I removed 37 from the remaining 300 participants. First, I removed 29 participants who requested help from the system three or fewer times. Because I derived all trust and testing variables from the instances in which the user requests help, I judged three data points insufficient for analysis. Also, I removed six participants who answered "9" or "1" on the confidence indication question every time they requested help. I suspected this behavior resulted from an attempt by these participants to minimize mouse movements, and thus I decided that their data regarding their intentional test behavior reflected inattentive behavior. These participants had very high numbers of intentional tests, with a range of 80-130. Then, I examined patterns of outliers using the outlier labeling method (Hoaglin & Iglewicz, 1987; Hoaglin, Iglewicz, & Tukey, 1986; Tukey, 1977). I used the outlier labeling method because it keeps cases that a two standard deviations rule would otherwise incorrectly remove (Hoaglin et. al, 1986). This analysis revealed a number of participants were outliers on a relatively small number of behavioral variables. I chose to retain these participants with two exceptions. That is, conducting this analysis revealed that one participant had not been assigned a participant number, so it was impossible to ensure that their behavioral data could be correctly paired with their survey data. Also,

this analysis revealed a second participant who had stepped away from the computer for several hours, i.e., response times of over 40,000 seconds. Finally, after constructing the survey measures, I looked for outliers again using the same outlier-labeling method and observed none. Following these procedures, I had 263 participants available for analyses.

Demographics

I examined the demographics of my 263 remaining participants. The average age was 31.09 years (*SD* = 8.70). Of the participants, 61.51% were non-Hispanic Caucasian, 21.89% Asian, 6.79% Hispanic, 3.02% Black, 2.64% participants of mixed race, and 4.15% undeclared. Due to an error, I did not collect gender for 77 participants. Of the remaining 188, 75% were male, 23.4% female, and 1.59% were nonbinary. Participants were from diverse countries. The greatest percent (38.02%) of participants were from the United States of America; 14.06% were from India; 9.88% from Brazil; 9.50% from Italy; 8.74% from the United Kingdom; 3.42% from Spain; 2.28% from France; 1.90% from Germany; 1.90% from Canada; 1.14% from Mexico; 1.14% from Romania; 0.76% from Ireland; 0.76% from Scotland; 0.76% from Turkey; and a single participant (0.38%) each from Argentina, Columbia, Ecuador, Georgia, Honduras, Hong Kong, Iran, Jamaica, Jordan, Poland, Sri Lanka, Sweden, the Netherlands, and Trinidad.

Behavioral Variable and Scale Construction

Behavioral variable. As described in the method section, I calculated (per participant) use, misuse, disuse, abuse, and trust rates using all 150 trials, as well as using only the final 100 and final 130 trials. As a reminder for the reader, use refers to cases in which a user accepted correct advice. Misuse refers to cases in which a user accepted incorrect advice. Disuse refers to cases in which a user ignores incorrect advice. Abuse

refers to cases in which a user ignores correct advice. Trust refers to the percentage of all help requests that resulted in the help being accepted (use plus misuse, divided by help requests). Also, I calculated (per participant) the number of intentional tests, the test accuracy of the system, the genuine request accuracy of the system, and variants of each using only the first 20 and 50 trials. As a reminder for the reader, intentional tests refer to help requests accompanied by a pre-request guess of 1 ("certain I do not see a weapon") or a 9 ("certain I do not see a weapon"). Test accuracy refers to the system's percentage of correct answers during those tests. Genuine request accuracy refers to the system's percentage of correct answers given during all non-intentional test interactions.

Scale construction. I constructed the scales as described in the method section. For the dispositional trust measure, the observed internal consistency reliabilities for the two subscales were: $\alpha = .63$ (alleviating workload), .61 (automation-induced monitoring. The observed internal consistency was .73 for openness and .82 for conscientiousness. The observed internal consistency was .79 for learning, .59 for prove performance, and .78 for avoid performance goal orientation. The observed internal consistency reliabilities for the three feedback orientation subscales were: .79 (utility), .66 (accountability), and .72 (feedback self-efficacy). Four of the 10 observed reliabilities were below .70, and most of the observed reliabilities were below those reported in prior research.

Descriptive Statistics

Table 1 shows the correlations between all behavioral variables except those collected from the pre-task survey. Table 2 removes the participants who failed the forced-response items on the pre-task survey and shows the correlations between select behavioral variables and all survey variables. The participants who failed the forced-

response items on the pre-task survey did not exhibit any patterns of faking behaviors when using the experimental program. Those participants were removed for all tables and analyses that involve the pre-task survey. I reported the means, standard deviations, and correlations for study variables in both tables.

One early concern in the design of this study concerned the frequency of intentional tests that would be conducted by participants. Figure 13 shows a histogram of the distribution of intentional tests conducted amongst the participants.

Table 1 mostly displays the expected relationships between variables. For example, trust, help requests, and intentional tests all predict task performance. However, Table 2 presents an unusual pattern. The relationship within survey variables and within task variables appear as expected. The same is not true of the relationships between them. Even after eliminating all participants who had failed the insufficient effort responding checks on the survey, the relationship between the survey and the task looks largely like noise. Manually matching times the survey and study were taken was not conclusive. Generating fake participants illustrated that the data was being matched correctly for the fake cases.

To provide further evidence that the data was (or was not) being matched correctly, I tested some expected correlations with unassisted performance (i.e., "percentage correct on all trials in which they did not ask for help"). Unassisted performance significantly correlates with cognitive ability ($r = .26$, $df = 258$, $p < .001$). A histogram for cognitive ability can be seen in figure 14. Unassisted performance does not significantly correlate with conscientiousness ($r = -.07$, $df = 258$, $p = .40$).

Conscientiousness was normally distributed. Unassisted performance does not correlate with any other survey variables.

One potential explanation for why conscientiousness would not predict performance is that the task itself is not one that allows for learning or skill acquisition within such a short period. I generated a skill acquisition curve using a multilevel model. First, I examined the ICC using percentage correct (task performance) as an outcome, examining only the cases in which a user did not ask for help. The ICC of .64 was greater than .1 and thus sufficient in moving forward with the MLM analysis because 64% of the variance was explained by the individual, i.e., between person variance. The deviance score for the random intercept model (deviance = 15819.69) was significantly different from the deviance score for the random intercept/slope model (deviance = 15424.71), X^2_{diff} (2) = 398.98, $p < .0001$. Allowing intercepts and slopes to vary fit significantly better than the model allowing only intercepts to vary. I tested whether time (i.e., 15 blocks of 10 trials each) accounted for significant variance in intercepts ($ß_{0j}$) and slopes ($ß_{1j}$). Time accounted for significant variance in intercepts ($ß_{0j}$ = 2.76, *SE* = 0.13, *df* = 3681, *t* = 20.45, $p < .0001$) but did not significantly predict performance slopes ($ß_{1j}$ = 0.23, *SE* = -0.009, *df* = 3681, *t* = -0.87, *p* = 0.37). These results indicated that participant performance improved with time (intercept) at similar rates for different participants (slope). Thus, the MLM analysis revealed the presence of skill acquisition, an effect that masked in the between person analysis approach used to test my predictions by the very large between person differences in task performance.

To summarize, the results of exploratory analysis to determine the accuracy of my link across survey and behavioral data was inconclusive. However, other exploratory

analyses exposed the presence of the expected relationship between cognitive ability and performance and the presence of skill acquisition, providing me with confidence in drawing conclusions from my results.

Condition Homogeneity

A computer randomly generated each participant's experimental condition. Because the random number generator operated independently with each participant, I needed to check for homogeneity across the conditions. To examine homogeneity across my experimental conditions, I ran a set of t-tests. The two experimental conditions did not differ by gender (t = -0.97, df = 161.35, p = 0.34), age (t = 0.92, df = 246.94, p = 0.36), dispositional trust (t = 0.40, df = 212.78, p = 0.69), or cognitive ability (t = 0.82, df = 236.42, p = 0.41).

Table 1

Means, Standard Deviations, and Correlations Between Behavioral Variables

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Age	31.13	8.72																
2. Task Performance	119.3	15.09	-.07															
3. Help Requests	91.25	50.11	-.01	**.79***														
4. Intentional Tests	14.28	20.87	-.07	**.23***	**.30***													
5. T1-20 I Tests	2.37	3.42	-.05	.01	**.13**	**.63***												
6. Test Accuracy	0.89	0.14	-.04	**.14**	.03	.03	-.03											
7. T1-50 Test Acc	0.86	0.23	-.05	.13	.05	.13	.09	**.74***										
8. Gen Request Acc	0.88	0.08	-.03	**.34***	**.26***	.03	.04	.06	.01									
9. T1-50 GR Acc	0.88	0.09	-.09	**.19***	.11	-.01	-.02	.09	.01	**.66***								
10. Trust	0.93	0.08	.09	**.25***	**.16***	.02	**-.14**	**.15**	.06	**.18***	.11							
11. T51-150 Trust	0.67	0.21	.02	**.54***	**.54***	**.15**	**-.14**	.12	.06	**.16***	**.17***	**.38***						
12. Use	0.83	0.10	.01	**.45***	**.31***	.10	-.05	**.26***	**.16**	**.70***	**.46***	**.75***	**.43***					
13. Misuse	0.10	0.07	.10	-.38	**-.29***	**-.13**	-.09	**-.25***	**-.21***	**-.84***	**-.55***	.04	**-.20***	**-.63***				
14. Disuse	0.02	0.02	-.11	.02	.01	.07	**.14**	**-.17**	-.05	-.07	.06	**-.59***	**-.17***	**-.31***	**-.21***			
15. Abuse	0.01	0.07	-.07	**-.29***	**-.19***	-.05	.11	-.11	-.06	**-.19***	**-.15**	**-.96***	**-.38***	**-.76***	.03	**-.34***		
16. Pre-Request RT	7554	5492	-.02	**-.20***	**-.36***	-.07	.05	.08	**.16**	**-.14**	-.09	**-.28***	**-.29***	**-.27***	.09	.10	**.29***	
17. Post-Request RT	5567	3085	-.01	**-.17***	**-.20***	-.01	.07	.00	.09	**-.24***	**-.21***	**-.49***	**-.31***	**-.40***	.04	**.33***	**.46***	**.49***

Note. Bolded Correlations are significant at $p < .05$. Bolded correlations with an * indicate $p < .01$. T1-20 I Tests = Intentional Tests from Trials 1-20. T1-50 Test Acc = Test Accuracy from Trials 1-50. Gen Request Acc = Genuine Request Accuracy. T1-50 GR Acc = Genuine Request Accuracy from Trials 1-50. T51-150 Trust = Trust from Trials 51-150. N = 263.

Table 2

Means, Standard Deviations, and Correlations Between Select Behavioral Variables and All Survey Variables

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Age	30.90	8.96																
2. Task Performance	120.5	14.58	-.09															
3. Help Requests	95.28	47.51	-.04	**.74***														
4. Intentional Tests	17.06	21.63	-.10	**.24***	**.34***													
5. Test Accuracy	0.89	0.14	-.04	**.14**	.03	.04												
6. Genuine Request	0.89	0.07	-.06	**.26***	.18/	-.03	.07											
7. Trust	0.93	0.08	.09	**.32***	**.21***	.04	.14/	**.30***										
8. Pre-Request RT	7910	5670	-.01	**-.23***	**-.39***	-.13	-.09	**-.22***	**-.29***									
9. Post-Request RT	5689	2995	-.02	**-.17**	**-.19***	-.05	.00	**-.24***	**-.50***	**-.49***								
10. Cognitive Ability	7.70	0.40	.13	.05	.00	-.13	-.08	-.02	.02	.07	.04							
11. DT – Alleviating	3.67	0.71	.11	.05	.02	.01	.04	.01	.04	-.01	.03	**.35***						
12. DT – Monitoring	2.93	0.79	.06	.06	.07	-.02	.01	.01	.06	-.07	.05	**.23***	**.65***					
13. Goal O. - Learning	3.97	0.83	.02	.00	-.02	.03	-.06	.04	-.01	.02	.07	**.38***	**.51***	**.27***				
14. Goal O. – Prove	3.49	0.75	.01	-.05	-.05	-.01	.04	.02	.04	.02	.01	**.36***	**.48***	**.25***	**.64***			
15. Goal O. – Avoid	3.55	1.17	-.07	.05	.05	**-.25***	.11	.08	.07	-.01	-.06	**.17**	**.34***	**.24***	-.02	**.46***		
16. Feedback – Utility	3.92	0.95	.09	.01	.00	.05	.03	-.04	.07	-.01	.01	**.46***	**.49***	**.29***	**.57***	**.49***	**.18***	
17. Feedback – Accnt.	3.76	0.88	.09	-.01	-.02	.08	-.02	.01	.03	.04	.07	**.50***	**.44***	**.24***	**.58***	**.51***	**.21***	***.83****
18. Feedback – Social	3.83	0.90	.10	.01	.02	.01	.00	-.02	.04	-.05	.03	**.50***	**.53***	**.32***	**.52***	**.59***	**.34***	***.82****
19. Feedback – Self-E	3.57	1.02	.07	.02	-.02	.11	.04	.05	-.02	-.06	-.01	**.41***	**.43***	**.27***	**.60***	**.36***	-.07	***.68****
20. Openness	3.78	0.79	.05	.03	-.04	-.09	-.04	.05	.04	.10	.05	**.41***	**.52***	**.37***	**.59***	**.41***	.09	***.45****
21. Conscientiousness	3.49	0.83	**.17**	.06	.01	-.01	-.03	.07	.11	-.03	-.13	**.36***	**.38***	**.17**	**.66***	**.47***	-.03	**.46***

Note. Bolded correlations are significant at $p < .05$. Bolded correlations with an * indicate $p < .01$. DT – Alleviating = The Alleviating Workload facet of the Dispositional Trust scale. DT - Monitoring = the Monitoring facet of the Dispositional Trust scale. Goal O. – Learning = the Learning facet of the Goal Orientation scale. Goal O. – Prove = the Prove Performance facet of the Goal Orientation scale. Goal O. – Avoid = the Avoid Performance facet of the Goal Orientation scale. Feedback – Acnt = the Accountability facet of the Feedback scale. Feedback – Social = the Social Awareness facet of the Feedback scale. Feedback – Self-E = the Self-efficacy facet of the Feedback scale. N = 242.

Table 2 (cont)

Means, Standard Deviations, and Correlations Between Select Behavioral Variables and All Survey Variables

	M	*SD*	17	18	19	20	5	6	7	8	9	10	11	12	13	14	15	16
1. Age	30.90	8.96																
2. Task Performance	120.5	14.58																
3. Help Requests	95.28	47.51																
4. Intentional Tests	17.06	21.63																
5. Test Accuracy	0.89	0.14																
6. Genuine Request	0.89	0.07																
7. Trust	0.93	0.08																
8. Pre-Request RT	7910	5670																
9. Post-Request RT	5689	2995																
10. Cognitive Ability	7.70	0.40																
11. DT – Alleviating	3.67	0.71																
12. DT – Monitoring	2.93	0.79																
13. Goal O. - Learning	3.97	0.83																
14. Goal O. – Prove	3.49	0.75																
15. Goal O. – Avoid	3.55	1.17																
16. Feedback – Utility	3.92	0.95																
17. Feedback – Accnt.	3.76	0.88																
18. Feedback – Social	3.83	0.90	**.81***															
19. Feedback – Self-E	3.57	1.02	**.67***	**.59***														
20. Openness	3.78	0.79	**.46***	**.45***	**.36***													
21. Conscientiousness	3.49	0.83	**.46***	**.37***	**.56***	**.42***												

Note. Bolded correlations are significant at $p < .05$. Bolded correlations with an * indicate $p < .01$. DT – Alleviating = The Alleviating Workload facet of the Dispositional Trust scale. DT - Monitoring = the Monitoring facet of the Dispositional Trust scale. Goal O. – Learning = the Learning facet of the Goal Orientation scale. Goal O. – Prove = the Prove Performance facet of the Goal Orientation scale. Goal O. – Avoid = the Avoid Performance facet of the Goal Orientation scale. Feedback – Acnt = the Accountability facet of the Feedback scale. Feedback – Social = the Social Awareness facet of the Feedback scale. Feedback – Self-E = the Self-efficacy facet of the Feedback scale. N = 251.

Figure 13

Histogram of Intentional Tests

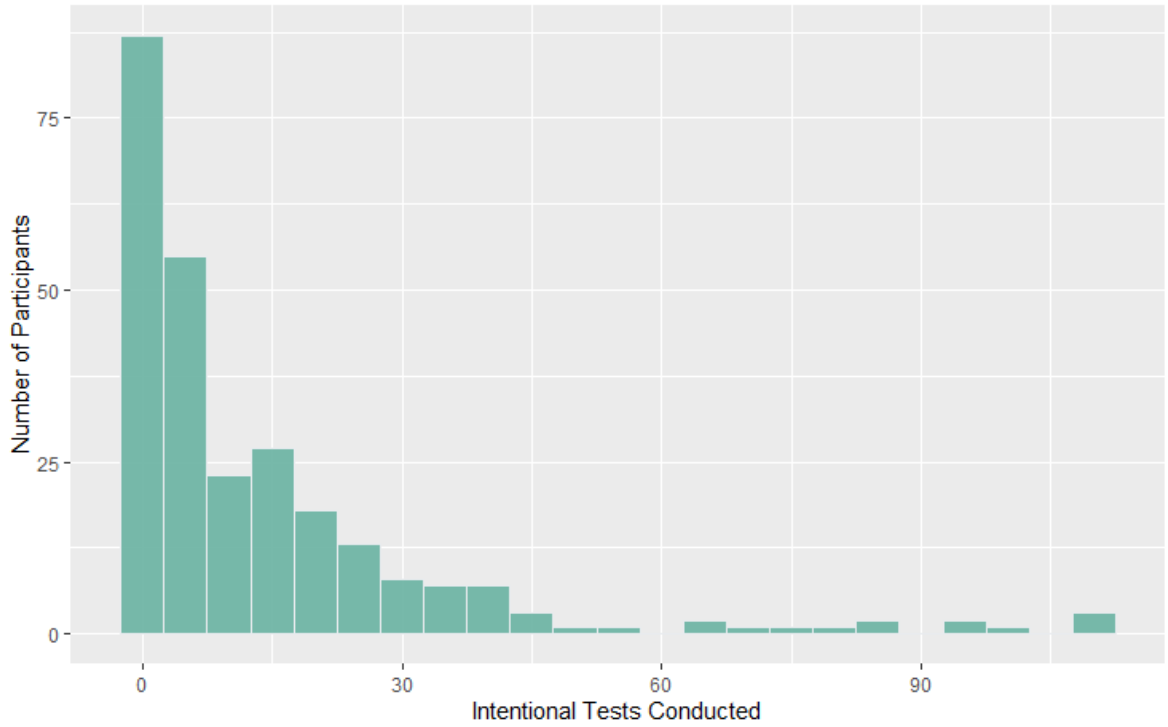
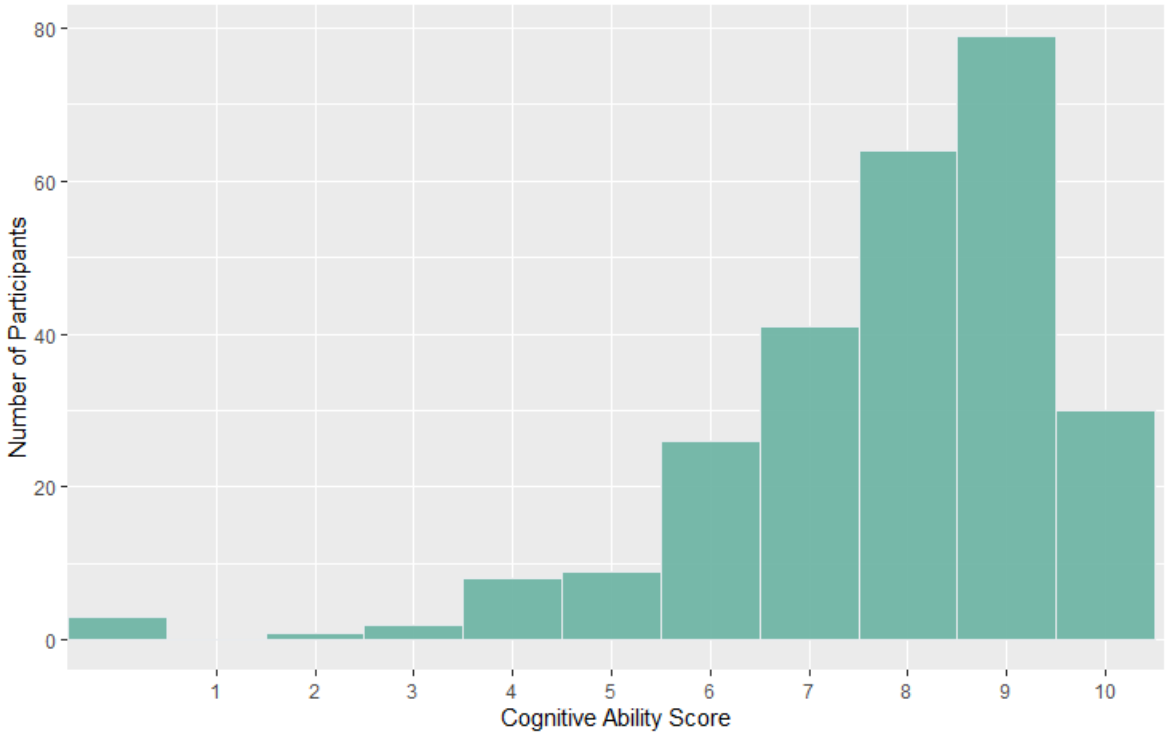


Figure 14

Histogram of Cognitive Ability



Hypothesis Testing

The effect of uncertainty on intentional tests (Hypothesis 1). Hypothesis 1 stated that individuals in the uncertain condition would conduct more intentional tests than individuals in the certain condition. To test this, I conducted a t-test comparing intentional tests between the high (M = 12.33) and low (M = 15.70) certainty conditions across all 150 trials. There was no significant difference between the two groups (t = 1.38, df = 258, p = 0.08). The distribution of intentional tests across uncertainty conditions indicated a potential floor effect. That is, intentional tests had a negative skew of 2.59 and a leptokurtic kurtosis of 10.39 (M = 14.34, SD = 20.86). Also, the distribution of each group individually indicated a potential floor effect. The high certainty condition had a negative skew of 2.02 and a leptokurtic kurtosis of 7.41. The low certainty condition had a negative skew of 2.47 and a leptokurtic kurtosis of 8.97. To account for the potential floor effect, I ran a Mann-Whitney test to determine whether a difference in medians existed and found nonsignificant results (w = 8545, p = .36). Hypothesis 1 was not supported.

I suspected that the induction (whether the system informed the participant of the system accuracy or not) would have the strongest effect in the early trials immediately after the participants' initial induction. So, in an exploratory analysis, I examined only the first 50 trials (rather than all 150 as hypothesized) and conducted another *t*-test comparing intentional tests between the high (M = 5.95) and low (M = 4.25) certainty conditions. The results of this test were significant (t = 1.98, df = 255.8, p = .02). To account for a potential floor effect, I also ran a Mann-Whitney test to determine whether a difference in medians existed and found nonsignificant results (w = 8645.5, p = .36). The number of intentional tests across the two conditions were similar.

The relationship between trial number and intentional test (Hypothesis 2).

Hypothesis 2 stated that the number of intentional tests would be negatively correlated with trial number. First, I calculated a simple Pearson correlation between trial number and number of intentional tests carried out by all participants on that trial. This correlation was significant (r = -.38, df = 148, p < .000001). These results supported Hypothesis 2. Trial number and number of intentional tests were negatively correlated.

I conducted exploratory analyses to examine further this effect. Possibly, this decrease is the result of fatigue. There is evidence of a fatigue or task practice effect. That is, results indicated a negative the relationship between the average time spent on each trial across all participants and trial numbers provides evidence for one of these effects (r = -.78, df = 148, p < .0001).I examined whether the relationship between time spent and trial number differed for the high (r = -.39, df = 148, p < .0001) versus low (r = -.15, df = 148, p = .07) certainty conditions, and I found a significant difference in relationship size (z = -2.28, p = .01). There was a larger negative relationship between time spent and trial number in the high certainty than in the low certain condition.

The relationship between stimulus difficulty and intentional tests (Hypothesis 3).

Hypothesis 3 stated that the number of intentional tests would be negatively correlated with stimulus difficulty (i.e., percent correct across participants). I tested this hypothesis using difficulty data from two different samples: Haskins (2018) and the current study. Both samples used the same 150 stimuli. Haskins' difficulty information had the advantage of stimuli difficulty calculated on an independent sample. However, the sample in the current study was larger and was exposed to more stringent insufficient effort response screening. The correlation between the number of intentional tests conducted for each stimulus and stimulus was nonsignificant (r = -

.07, *df* = 148, *p* = .38) using the Haskins difficulty information. The correlation between the number of intentional tests conducted for each stimulus and the stimulus difficulty was significant (*r* = -.18, *df* = 148, *p* = .03) using data from the current study sample. These results partially supported Hypothesis 3. The number of intentional tests was negatively related to stimulus difficulty.

The relationship between intentional tests and response times (Hypothesis 4). Hypothesis 4 stated that the post-request response time would be shorter for an intentional test relative to genuine requests for help. To test this, I conducted a paired t-test comparing participants' mean post-request response times during intentional tests (*M* = 5635.64ms, *SD* = 3291.62ms) to their mean post-request response times during genuine requests (*M* = 6298.42ms, *SD* = 4650.78ms). Response times were shorter for intentional tests compared to genuine requests for help (*t* = 1.94, *df* = 230, *p* = .03). These results supported Hypothesis 4.

The relationship between system accuracy during intentional tests versus genuine requests for help and trust (Hypothesis 5). Finally, I predicted that trust would be more strongly related to system accuracy during intentional tests than to system accuracy during all other interactions (i.e., genuine requests). The correlation between early intentional test system accuracy and later trust was not significant (*r* = .06, *df* = 186, *p* = 0.43). The correlation between early genuine request system accuracy and later trust was significant (*r* = .17, *df* = 256, *p* = .006). I conducted a Stiegler's test to compare correlations and found nonsignificant results (*z* = 0.74, *p* = .77). Thus, Hypothesis 5 was not supported.

I conducted exploratory analyses to examine further this effect. I suspected that the effect of the intentional test may be more localized and that a trust increase or decrease from a respective passed or failed intentional test may wear off over time. I did not have sufficient

power to address this issue using multilevel analysis. To test this exploratory explanation, I first examined only the sets of five trials immediately following passed intentional tests. I calculated participants' trust levels using only those trials that had been preceded (within five trials) by a passed intentional test. This created a variable called "post passed test trust". I compared participants' trust in trials immediately following the system passing an intentional test (M = 0.96, SD = .10) to participants' overall trust levels (M = 0.93 and SD = .08) using a paired one-directional t-test (t = 1.78, df = 201, p = .04). Then, I repeated this process using failed intentional tests. To test this, I first examined only the sets of five trials immediately following failed intentional tests. I calculated participants' trust levels using only those trials that had been preceded (within five trials) by a failed intentional test. I compared the trust shown in trials immediately following the system failing an intentional test (M = .89 and SD = .12) to the participants overall trust levels (M = 0.93 and SD = .08) using a paired one-directional t-test (t = -1.88, df = 104, p = .03). When a participant conducted an intentional test, their trust behaviors adjusted upwards (for a passed test) or downwards (for a failed test) for the following five trials.

Discussion

Overview

The purpose of the current study was to examine factors that influence the number and probability of intentional tests and distinguish intentional tests from genuine requests for help. I found evidence of the existence of intentional tests as purposeful acts intended to gather information about an unknown system. I found results suggesting that users are more likely to conduct intentional tests on earlier and easier trials and that users also spend less time considering system advice when conducting an intentional test. Finally, whereas I did not find evidence that intentional tests influenced later trust levels, I did find evidence that there is a more

localized effect on trust levels that persists for a relatively short number of trials after an intentional test. These results contributed to the literature by providing evidence of a new behavior, intentional testing, in which users ask for help they do not need to gauge the usefulness of an unfamiliar system. Furthermore, these results provided insight into feedback mechanisms that appear in some trust models (e.g., Ghazizzadeh, Lee, & Boyle, 2012; Mayer, Davis, & Schoorman, 1995). These results raised issues relating to 1) the implications of feedback-seeking research for an automation context, 2) elaborations of feedback mechanisms described in trust models, 3) intervention longevity, and 4) the design of future products.

Theoretical Implications

Implications of feedback-seeking research within the context of automation. There are three main implications of this research for the feedback-seeking literature. First, replicating interpersonal research within automation opens the door to examining other potential generalizations of feedback-seeking literature to a domain of automation. Second, this research provides evidence supporting two of Ashford and Cumming's (1983) propositions, i.e., Propositions 5 and 6. Last, this research provides at least one example in which the feedback-seeking literature does not generalize to the domain of automation, confirming another example of a case in which interpersonal research does not fully generalize to the domain of automation.

First, this research replicates some existing feedback-seeking effects found in interpersonal domains. The feedback-seeking literature (e.g., Ashford, De Stobbeleir, & Nujella, 2016) has uncovered a variety of useful behaviors participants display in interpersonal situations. Now, researchers have justification to use this research stream, generalized to automation, to provide hints at potentially useful and untested behaviors in the automation domain. I found support for the well-researched relationship (e.g., Chapanis, 1964) between feedback and task

effectiveness in the correlation between intentional tests (i.e., feedback requests) and task performance seen in Table 1 ($r = .23$, $p < .001$). The results I found supported at least a partial application of interpersonal feedback-seeking research to the domain of automation. Participants took actions in the first 50 trials to learn about their situation and performance when they were placed in an uncertain situation (as per Hypothesis 1), and this participant learning enhanced performance. This relationship between feedback requests and task performance is the central relationship that started the feedback seeking literature and remains one of the most well-researched relationships within that research stream (Ashford & Cummings, 1983; Chapanis, 1964). In a general sense, the current study found another example of participants seeking useful information about their performance (i.e., feedback seeking behavior), and so at least some of the feedback seeking research may be applied within the automation domain.

Second, this research found support for two of the propositions put forth by Ashford and Cummings (1983). Ashford and Cummings proposed several scenarios in which researchers could expect individuals to engage in behaviors that seek more feedback about their environment. I did not design my study to address any of Ashford and Cummings' propositions specifically. However, upon review I note that some of my results may be interpreted as providing evidence in support of two of their propositions. The current study's Hypothesis 5 could be interpreted as support for Ashford and Cumming's (1983, p.387) fifth proposition that suggested that active feedback seeking behavior should be less common when using technology that is "more routine." My results indicated that when a participant conducted an intentional test, their trust behaviors adjusted upwards (for a passed test) or downwards (for a failed test) for the following five trials. If the reader accepts the premise that the slower times in the later trials were analogous to the use of "more routine" technology, then I have provided that support. In the

same way, the reader could elect to interpret Hypothesis 3 as providing support for their Ashford and Cumming's sixth proposition (1983, p.389) that suggested individuals tend to use strategies that require more effort less than easier to use strategies. My results indicated that the number of intentional tests was negatively related to stimulus difficulty. My results support the sixth proposition if the reader accepts the premise that more difficult stimuli require "more effort". My study did not provide evidence relevant to any of Ashford and Cummings' other propositions.

Third, my research provides an example of a case in which interpersonal research does not generalize to the domain of automation. The relatively short-term impact of the results of the feedback (i.e., the short-term effect of intentional tests on trust seen in hypothesis 5) is the main divergence from the feedback-seeking literature (e.g., Ashford, De Stobbeleir, & Nujella, 2016). That is, in general, the feedback-seeking literature has suggested that the effect of feedback lasts much longer than the matter of minutes I observed in my study (as seen in the exploratory analysis accompanying Hypothesis 5).

One possible explanation for the observed short-term effect of feedback in my study relates to the frequency and short duration of interactions with an automated system, compared to interpersonal interactions. Interactions with automation are notably distinct in that they lack some of the moderators that decrease feedback-seeking behavior in interpersonal cases. Specifically, the impression management motive of feedback-seeking behavior (Morrison & Bies, 1991) should limit the expression of feedback seeking behavior in interpersonal interactions. Some of the properties of feedback-seeking interactions that decrease feedback-seeking behavior are simply not present when dealing with an automated system. Four of these properties (publicness, dependency, performance, and dispositional factors) were discussed previously alongside Figure 4. These differences outlined between feedback seeking within the

domain of automation and interpersonal interaction imply that there should be much more feedback seeking behavior when dealing with automation. This provides one potential explanation for the relatively short-term effect on behavior that feedback seeking has within an automation domain. Because participants are much more frequently seeking new information, previous instances of feedback seeking may more easily be "overwritten" by more recent experiences.

As mentioned above, the short-term effects on trust by intentional tests (i.e., feedback seeking behavior) may be in part due to the relative frequency with which a user requests feedback from an automated assistant sometimes as frequently as multiple times per minute for some participants. The costs of impression management identified by Morrison and Bies (1991) lead to a much more severe limiting of the rate of feedback-seeking behaviors during interpersonal interactions. Whereas an individual may display a handful of instances of feedback seeking behavior over a matter of weeks, that same individual is able to seek out feedback at a rate of several times per minute without risking social repercussions.

Because researchers can expect individuals to perform feedback-seeking behavior much more frequently when interacting with an automated system that trigger impression management concerns, it makes sense that the effect any individual intentional test has on behavior is more limited in scope. A user may quickly follow up a failed intentional test by the system by conducting another test that the system immediately passes. Participants do not display the same quick repetition in behavior when asking others for feedback (Ashford, Blatt, & Dewalle, 2003). Researchers may expect each instance of interpersonal feedback to alter behavior over a longer time frame, as it will be a comparatively longer amount of time before a feedback seeker encounters conflicting information.

As an additional note, further conclusions could have been drawn regarding the application of feedback-seeking research. However, due to unexpected patterns existing between my survey and behavioral data, it is unclear whether these conclusions would be valid. Manually matching data beyond what had already been done automatically was impossible. However, fake participant data were correctly matched when used as a test. Cognitive ability and conscientiousness were both expected to predict unassisted performance. Cognitive ability did predict unassisted performance, but conscientiousness did not. As a note, I could have examined also the pre-help guesses a user submitted on trials in which they did ask for help, but I expected that either set would show the expected pattern of relationships. The simplest explanation for this relationship is that there still exists an error in matching the behavioral and exploratory survey data. Another potential explanation for this looks at the lack of a skill acquisition curve within this task. It is possible that this perceptual task differs enough from common I/O tasks that the conscientiousness/performance relationship is not detectable here. Not only is this a perceptual task, but the addition of an automated assistant potentially created a situation that may limit the expression of individual differences. Given the constraints of the system, one could argue that the only "skill" the participants learned was how to trust the system if their personal accuracy was less than 80%. However, given that there still exists the potential survey/behavioral data mismatch, I will refrain from further analysis of the survey data.

Expanding upon the feedback mechanisms described in trust models. Many models of trust (both interpersonal and in automation) include an arrow showing a feedback loop, suggesting that the results of previous interactions will inform future trust levels (e.g., Ghazizzadeh, Lee, & Boyle, 2012; Lee & See, 2004, Mayer, Davis, & Schoorman, 1995; Muir, 1994). The automation acceptance model in particular describes the feedback process as being a

dynamic, bidirectional process that results in users calibrating their trust levels over time (Ghazizzadeh, Lee, & Boyle, 2012). My research contributes to the trust literature by providing evidence that there may be a missing step in the trust feedback loop. Because users explicitly act to start at least the portion of this feedback loop by conducting intentional tests, I suspect that there are conditions that may increase or decrease the rate at which feedback occurs because a user is acting explicitly to seek feedback. Conditions that may increase the likelihood of feedback-seeking behavior may be based upon system properties such as easier stimuli. Indeed, I observed that stimuli difficulty related to feedback-seeking frequency in support of Hypothesis 3. Also, familiarity with or knowledge about the system may reflect conditions that influence the likelihood of feedback-seeking behavior. Indeed, my results revealed that situational uncertainty (Hypothesis 1) and trial number (Hypothesis 2) are both related to feedback seeking behavior (i.e., intentional tests). Note that my results did not support Hypothesis 1 across the entire study, but examining the difference in behavior within the first 50 trials alone showed that the uncertainty induction did influence behavior in a more temporary fashion. This provides justification for a set of moderators acting upon the feedback arrow drawn in many models, including Ghazizzadeh, Lee, and Boyle (2012), Lee and See (2004), Mayer, Davis, and Schoorman (1995), and Muir (1994). This feedback arrow, although present in all the listed models, is rarely expanded upon. I suggest adding the moderators of task difficulty and situational uncertainty, along with some indication that the feedback from the system may be increased or decreased due to specific behaviors on the part of the user.

Practical Implications

Intervention longevity. One of the difficulties identified by the automation trust literature is an ongoing difficulty in finding s long-lasting interventions (e.g., French, Duenser, &

Heathcote, 2018). Though researchers have identified interventions such as particular forms of training or warnings that artificially raise or lower trust, the effect of these interventions is often very short-lived (Skitka, Mosier, & Burdick, 1999). This has been a persistent problem throughout automation research, and it is a problem that leads to poorly calibrated user trust and costs lives. For example, pilots who trust their autopilot too much are responsible for at least a portion of airplane crashes (Parasuraman & Manzey, 2010). This problem likely will generalize to personal vehicles in the next decade as automated driving becomes more ubiquitous. Whereas the automation trust literature also identifies problematic situations that arise from trusting an automated system too *little* (i.e., Parasuraman & Riley, 1997), these cases are harder (though not impossible) to address with the use of intentional tests.

One problem might be that researchers/practitioners need to administer trust interventions more frequently, but more frequent administrations might not be feasible due to limitations of cost or time. Training, the most common intervention, is impractical to repeat at rates necessary to have an impact on behavior, due to the relatively short-lived effect this training has on interactions with automated systems (Bisantz & Seong, 2001). Planned failures are impractical but only because arbitrarily causing a system to fail randomly for the sake of decreasing user trust is a cure that is as bad as the disease. Originally, I suspected that the results of intentional tests would be long-lasting enough to circumvent this problem. Instead, I now believe that intentional tests are frequent enough that a regular intervention worked into the very system design itself may be possible: planned failures.

Planned failures, specifically introduced during interactions that we identify as a likely intentional test, have two main advantages. First, they are more likely to have a stronger effect than randomly introduced failures. Second, they are safer to introduce due to user vigilance and

preprocessing (see Hypothesis 4). As seen in Hypothesis 4, participants spent less time after the system advice had been given during intentional tests–consistent with the expected behavior that they were merely confirming or denying a previously made judgment. The cost of intentionally failing in these times is lower because it should not influence the previously-made judgments of users. The benefit of intentionally failing during these times is also higher than simply implementing random failures, as failing during an intentional test will affect trust levels more than a random failure (see Hypothesis 5 exploratory analyses). Engineers and scientists unfortunately cannot use this same mechanism to artificially inflate trust levels. The concept of a "planned success" that engineers may call upon at strategic times is not feasible from an design standpoint, as there is no way to eliminate true system failures in these cases.

As a practical example, if engineers were able to accurately identify cases in which a driver is watching his or her system as closely as if they themselves were driving to ensure it is operating correctly (i.e., an intentional test), then causing a minor failure such as a noticeable wobble that stays within the lane in exactly that moment will both reduce the trust of a user (perhaps even on-demand if the system judges trust levels to be too high) and do so without risking life. Even that level of failure may be too much. Perhaps a wobble within the lanes might lead to an over-corrective steer and a catastrophic failure. In high-stakes environments such as driving, perhaps a failure so small as flipping on a turn-signal at an incorrect time could serve to correct for over-trusting users. Obviously any interventions that may impact human life would require extensive testing before enactment, but a careful implementation could save countless lives. Engineers could introduce these failures at a much higher rate due to their relative safety, and as such may be the beginning of a real solution to the problem of unsuccessful long-lasting trust-decreasing interventions.

Designing future products with intentional tests in mind. If an engineer wants to have more high-resolution control over their user's trust level moment-to-moment via repetitive intervention during intentional tests, then the mechanisms to detect the presence of an intentional test should be part of the very design of the system. For example, a baseline requirement for this kind of system functionality would be some ability to store user profiles (assuming multiple users). If a system cannot differentiate between users, users swapping out on a system would not be able to reset to a new baseline trust level.

An engineer could even reduce the precision of the instructions presented during training as suggested in Hypothesis 1 to decrease user certainty and thus increase early trust malleability through an increased number of opportunities for interventions during early intentional tests. An engineer could build in the capability for a system to change its answer from an incorrect first answer to a later correct suggestion. This would result in the user acknowledging the system as failing a greater number of intentional tests, but less risk from an acceptance of those failed suggestions (as per Hypothesis 4). Also, an engineer could build in specific cases that attempt to elicit an intentional test (e.g., by showing easy tasks earlier in an interaction with a new user), and then voluntarily fail or attempt to pass those tests as needed to adjust user trust levels.

If engineers seek domain-specific conditions for intentional tests in the future for specific high-risk activities (i.e., driving and piloting), then the usefulness and criticality of intentional tests increase dramatically. If, for example, researchers find that eye movement is critical to identifying intentional testing behavior during driving with an automated vehicle, then engineers can incorporate eye-tracking systems into the vehicle for the sake of acknowledging times that the car may safely fail to lower trust levels.

Limitations

I must acknowledge several limitations to the current study. First, I collected this data entirely during the COVID-19 pandemic and quarantine, which likely had an impact on the pool of participants participating in mTurk. Much prior research using MTurk participants has focused on survey research. As such, previous research identifying properties unique to mTurk workers might be relevant to my sample because I focused my study on performance on a task rather than survey responses (Burnham & Piedmont, 2018). For example, previous research has uncovered a tendency for mTurk workers to have higher levels of expertise in survey-taking (Hauser & Schwarz, 2016). However, the current study may not display that particular relationship given that the demographics of mTurk may have changed substantially during the COVID-19 pandemic due to the rise in unemployment.

Another limitation related to the strength of the uncertainty intervention. That is, my manipulation involved changing a single line of three pages of training text. The training pages gave this induction line extra focus by separating it into its own final paragraph, but this may have been insufficient. Also, I relied on the hypothesis test to confirm that the manipulation had the intended psychological effect. If the only difference between Condition 1 and Condition 2 was that specific change in wording, then any significant difference in behavior between the two groups should result from that change. I did not use a manipulation check because I was concerned that including an uncertainty induction manipulation check might influence the analyses for Hypotheses 2 through 5. For example, asking participants to report their knowledge about system accuracy before completing training could have an effect on intentional tests.

A third limitation relates to a particular property of the automated assistant. The automated assistant failed at completely random times, and this does not reflect the results a participant can expect when using system like this in the real world. For example, “true” AI

powered by machine learning may have particular issues when trying to identify one of the targets from a particular angle, such as when viewing guns from above. The system used in this study had no pattern to its responses, and this may have influenced participant behavior. One of the participants even mentioned that they had spent a great deal of time trying to learn the elements in a particular photo that might cause the automated assistant to fail. This participant created completely incorrect rules they believed might cause the automation to fail more frequently. Due to the limitations of the MTurk platform, I was unable to ask other participants whether they had a similar experience.

Finally, another potential limitation is that I might have had range restriction relating to participants' cautiousness (or riskiness), which might be relevant to my trust measures. Recall that, upon identifying insufficient effort responding by the user, the experimental program first ended the program and then saved a hidden file to the user's computer (without the user's knowledge) that prevented the experimental program from ever opening in the future. A program that saves files in this manner requires a security certificate from Windows (that I was unable to acquire) or else some antivirus programs will flag the program as problematic. As a result, I received some reports of users quitting the study due to their antivirus programs alerting them to a (not present, but understandable) danger. I identified only two antivirus programs as problematic but this could have created a bias in my sample. Potentially, I had a smaller number of participants who were more cautious with technology, for example, participants who run Norton or AVG antivirus programs.

Future Research

I addressed some specific suggestions for future research above, and here I will address more general suggestions. Whereas I found evidence in support of two of the hypotheses

proposed by Ashford and Cummings (1983), I was unable to address twelve of their fourteen hypotheses. As a result, it is difficult to determine what parts of interpersonal feedback-seeking behavior might not function within the domain of automation trust and intentional tests. Examining more of these properties en masse may shed some light on the limitations of the comparison between interpersonal and automation feedback seeking.

I did not structure this study to capture the relatively short-term impact on trust that resulted from a passed or failed intentional test. More specifically, I did not structure this study to identify whether there was any kind of more complex cumulative effect that was more persistent over time. Given that we as researchers have yet to produce an intervention that can have long-lasting effects on user trust levels, this seems like a potentially very valuable avenue of research. Revealing the details of the unexpectedly short-term nature of intentional test effects on user behavior is a valid path for future research.

Also, my study focused primarily on ensuring that internal validity was maintained. As a result, an attempt to identify the properties of intentional tests in a more realistic environment may uncover different results. For example, I did not examine the very common user expectations of repeated interactions and increasing expertise. Specifically, a pilot or a driver may have many interactions with our system throughout their lifetime. Do they tend to conduct a greater number of intentional tests at the beginning of a flight or drive? Is there a way to induce uncertainty such that the users feel the need to retest regularly? These questions are critical to answer within the critical domains of aircraft and ground autopilots. Because normal usage behavior for many systems involves repeated interactions with a single system over time, an examination of how or whether users alter their use of intentional tests over multiple sessions over a greater length of time would be useful. Learning how repeated interactions over time and

altered (i.e., more realistic) automated assistant behavior change the properties of intentional tests is a valuable future research direction.

Another potential question for future research has to do with the role of stakes. The payment structure was set up to somewhat reflect the stakes of a real TSA agent. However, the stakes in my study were much lower than for a real TSA agent. No serious consequences occurred in my study if a participant got the answer wrong. Rather, participants only lost a small amount of compensation for errors. In addition, the presence of a weapon in half of the trial stimuli images further divorced the study from reality. Future research should focus on the use of intentional tests in environments with higher stakes.

Furthermore, it is possible that the domains of aircraft and ground autopiloting have a large number of unique properties. Though many self-report measures of dispositional trust in automation do not display these domain-based effects (Merrit & Ilgen, 2008), the same is not necessarily true of the behaviors I observed here. Given their relative importance, finding whether intentional tests are tied to unique user behaviors within these important domains would dramatically increase the usefulness of intentional tests.

Conclusion

The current study contributes to the existing literature by using research on feedback-seeking behavior from interpersonal interactions to inform research examining learned trust in automation. I found evidence that feedback-seeking behavior occurs within interactions with automation, and that due to the lack of impression management required, feedback-seeking behavior occurs far more frequently and with far less long-lasting impact on behavior in automation. Engineers may use this research to begin addressing a long-standing problem in automation trust: the unavailability of interventions that have long-term effects on user behavior.

Feedback-seeking behavior in relation to automation use (i.e., intentional tests) may be the beginning of a solution if applied in regular and safe interventions. Future research should investigate the extent to which researchers can generalize interpersonal research into the domain of automation trust, along with the specific properties of intentional tests that are relevant in the domains of automated driving and piloting. Engineers should consider how to incorporate the ability to detect intentional tests if they wish to directly and safely alter a user's trust levels. Overall, my results demonstrated that intentional tests exist, can be a useful tool, may be identifiable using automation, and have at least some unintuitive properties that merit further study.

References

Arntz, M., Gregory, T., & Zierahn, U. (2017). Revisiting the risk of automation. *Economics Letters*, *159*, 157-160.

Ashford, S. J., Blatt, R., & VandeWalle, D. (2003). Reflections on the looking glass: A review of research on feedback-seeking behavior in organizations. *Journal of management*, *29*(6), 773-799.

Ashford, S. J., & Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies of creating information. *Organizational behavior and human performance*, *32*(3), 370-398.

Ashford, S. J., De Stobbeleir, K., & Nujella, M. (2016). To seek or not to seek: Is that the only question? Recent developments in feedback-seeking literature. *Annual Review of Organizational Psychology and Organizational Behavior*, *3*, 213-239.

Audibert, J. Y., Munos, R., & Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, *410*(19), 1876-1902.

Bahner, J. E., Hüper, A. D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, *66*(9), 688-699.

Berry, D. A., & Fristedt, B. (1985). Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability). *London: Chapman and Hall*, *5*, 71-87.

Bisantz, A. M., & Seong, Y. (2001). Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. *International Journal of Industrial Ergonomics*, *28*(2), 85-97.

Bowden, E. M., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in cognitive sciences*, *9*(7), 322-328.

Bowers, T. L., & Pantle, M. L. (1998). Shipley institute for living scale and the Kaufman Brief Intelligence Test as screening instruments for intelligence. *Assessment*, *5*(2), 187-195.

Brynjolfsson, E., & McAfee, A. (2013). The great decoupling. *New Perspectives Quarterly*, *30*(1), 61-63.

Burnham, M. J., Le, Y. K., & Piedmont, R. L. (2018). Who is Mturk? Personal characteristics and sample consistency of these online workers. *Mental Health, Religion & Culture*, *21*(9-10), 934-944.

Chapanis, A. (1964). Knowledge of performance as an incentive in repetitive, monotonous tasks. *Journal of Applied Psychology*, *48*(4), 263.

David, H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of economic perspectives*, *29*(3), 3-30.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management science*, *35*(8), 982-1003.

Deetlefs, J., Chylinski, M., & Ortmann, A. (2015). MTurk 'Unscrubbed': Exploring the good, the 'Super', and the unreliable on Amazon's Mechanical Turk. *UNSW Business School Research Paper*, (2015-20A).

De Stobbeleir, K. E., Ashford, S. J., & Buyens, D. (2011). Self-regulation of creativity at work: The role of feedback-seeking behavior in creative performance. *Academy of management journal*, *54*(4), 811-831.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013, March). Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction* (pp. 251-258). IEEE Press.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human factors*, *48*(3), 474-486.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International journal of human-computer studies*, *58*(6), 697-718.

Farias, Vivek F., and Ritesh Madan. "The irrevocable multiarmed bandit problem." *Operations Research* 59.2 (2011): 383-399.

French, B., Duenser, A., & Heathcote, A. (2018). Trust in Automation - A Literature Review. *CSIRO report EP184082*. CSIRO, Australia.

Frey, C. B., Berger, T., & Chen, C. (2017). Political machinery: Automation anxiety and the 2016 US presidential election. *University of Oxford*.

Frey, C. B., & Osborne, M. (2013). The future of employment.

Fridland, E. (2017). Automatically minded. *Synthese*, *194*(11), 4337-4363.

Geels-Blair, K., Rice, S., & Schwark, J. (2013). Using system-wide trust theory to reveal the contagion effects of automation false alarms and misses on compliance and reliance in a

simulated aviation task. *The International Journal of Aviation Psychology*, *23*(3), 245-266.

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the Technology Acceptance Model to assess automation. *Cognition, Technology & Work*, *14*(1), 39-49.

Goldberg, K. (2011). What is automation?. *IEEE Transactions on Automation Science and Engineering*, *9*(1), 1-2.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*, 84-96.

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

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, *48*(1), 400-407.

Hémous, D., & Olsen, M. (2014). The rise of the machines: Automation, horizontal innovation and income inequality.

Hoaglin, D. C., Iglewicz, B., & Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *81*(396), 991-999.

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American statistical Association*, *82*(400), 1147-1149.

Hoaglin, D. C., Iglewicz, B., & Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *81*(396), 991-999.

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

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of applied psychology*, *64*(4), 349.

Juvina, I., Collins, M.G., Larue, O., Kennedy, W., de Visser, E., & de Melo, C. (2018 conditionally accepted). Toward a unified theory of learned trust in interpersonal and human-machine interactions. *ACM Transactions in Interactive Intelligent Systems*.

King, W. R., & He, J. (2006). A meta-analysis of the technology acceptance model. *Information & management*, *43*(6), 740-755.

Krasman, J. (2010). The feedback-seeking personality: Big five and feedback-seeking behavior. *Journal of Leadership & Organizational Studies*, *17*(1), 18-32.

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

Linderbaum, B. A., & Levy, P. E. (2010). The development and validation of the Feedback Orientation Scale (FOS). *Journal of Management*, *36*(6), 1372-1405.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277-301.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human factors*, *48*(2), 241-256.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194-210.

Morrison, E. W., & Bies, R. J. (1991). Impression management in the feedback-seeking process: A literature review and research agenda. *Academy of Management Review*, *16*(3), 522-541.

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *The International Journal of aviation psychology*, *8*(1), 47-63.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905-1922.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human factors*, *52*(3), 381-410.

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51-55.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-inducedcomplacency. *The International Journal of Aviation Psychology*, *3*(1), 1-23.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230-253.

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

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology*, *9*(2), 371-377.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological review*, *84*(1), 1.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making?. *International Journal of Human-Computer Studies*, *51*(5), 991-1006.

Smith, E. E. (1968). Choice reaction time: An analysis of the major theoretical positions. *Psychological Bulletin*, *69*(2), 77.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245-251.

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: the mobilization-minimization hypothesis. *Psychological Bulletin*, *110*(1), 67.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2, pp. 131-160).

van der Rijt, J., van de Wiel, M. W., Van den Bossche, P., Segers, M. S., & Gijselaers, W. H. (2012). Contextual antecedents of informal feedback in the workplace. *Human Resource Development Quarterly*, *23*(2), 233-257.

VandeWalle, D. (1997). Development and validation of a work domain goal orientation instrument. *Educational and psychological measurement*, *57*(6), 995-1015.

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*(3), 475-482.

Wong, A. L., Haith, A. M., & Krakauer, J. W. (2015). Motor planning. *The Neuroscientist*, *21*(4), 385-398.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes*, *83*(2), 260-281.

Appendix A

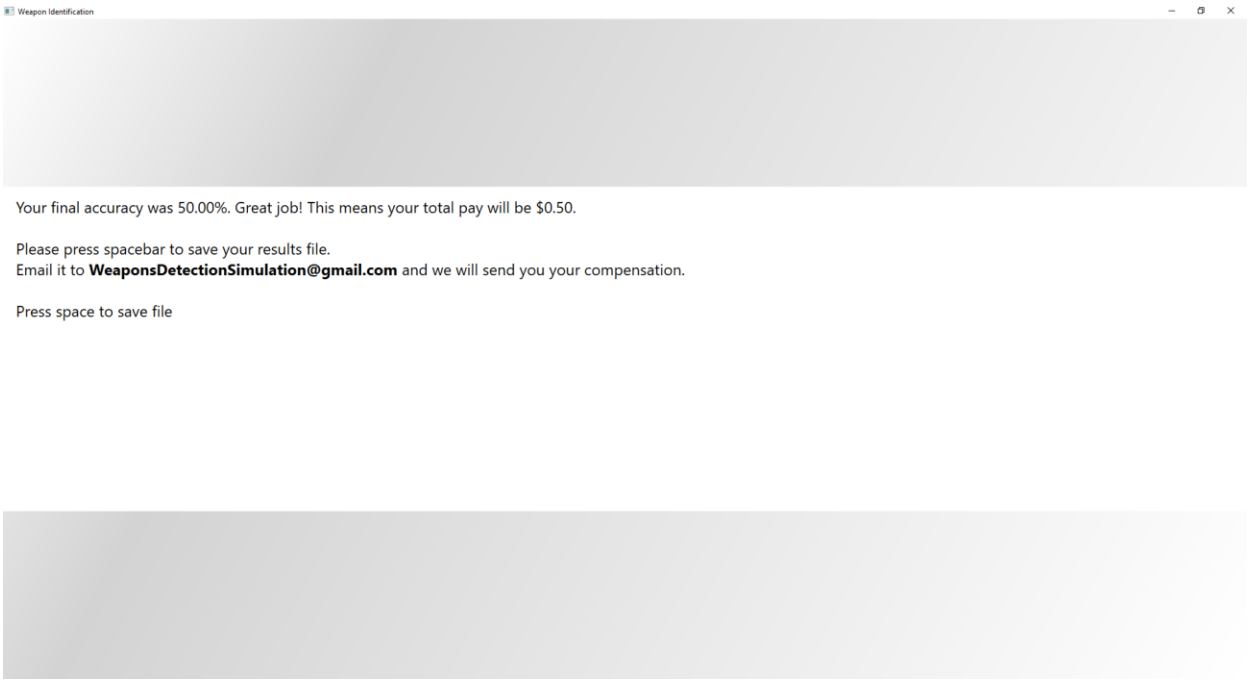
Consent Form

INSTRUCTIONS: Please input your anonymized MTurk Worker ID to indicate that you agree with the following statement.

> This study is anonymous. No information on your identity will be collected. Only aggregate (summarized data that does not identify individual answers) data will be presented or published. You are free to refuse to participate in this study or to terminate your participation at any time.
>
> Completion and submission of the survey implies your consent to participate. If you have any questions about this research study, you may contact me at weapondetectionsimulation@gmail.com. If you have general questions about giving consent or your rights as a research participant in this research study, you can call the Wright State University Institutional Review Board at 937-775-3336.

Appendix B

Debriefing Form



Appendix C

Task Tutorial

The following shows each screen participants will see in the tutorial. The only screen that differs is for the uncertainty/certain manipulation. In the certain condition, a participant will see a screen indicating that the automated assistant is accurate 80% of the time. In the uncertain condition, participants will see a screen indicating that the automated assistant is accurate most of the time.

Figure C1

Tutorial Page 1: First Screen of Weapons Detection Simulation

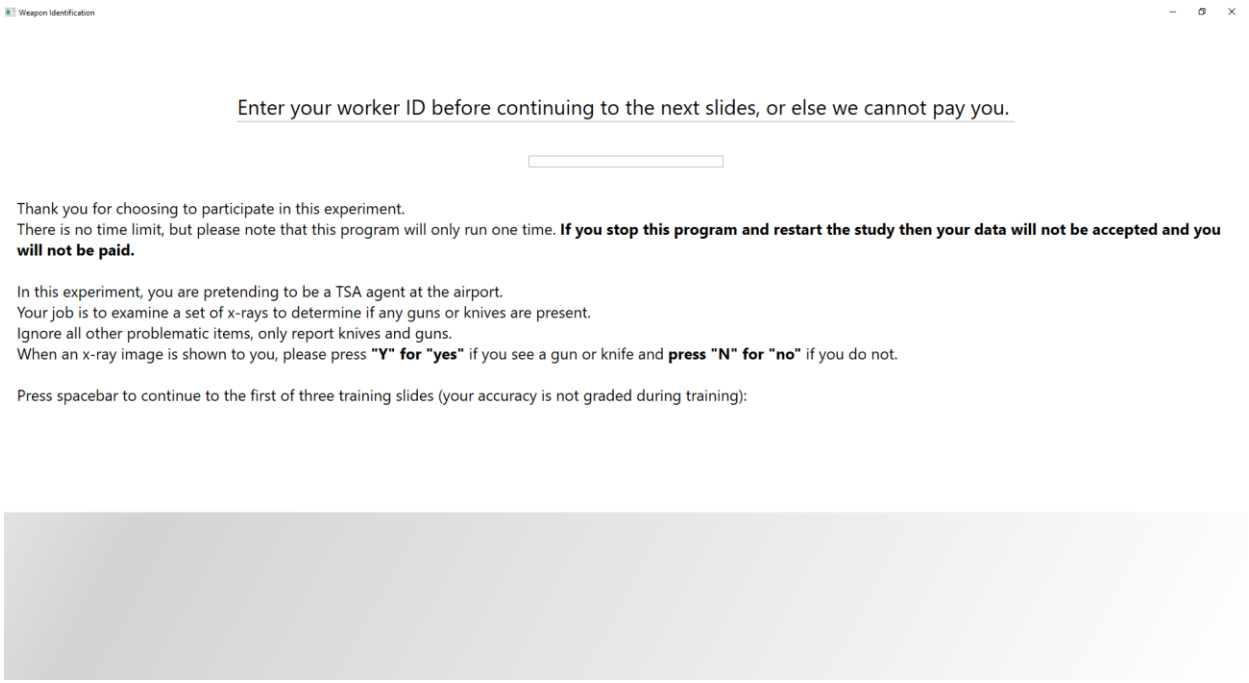
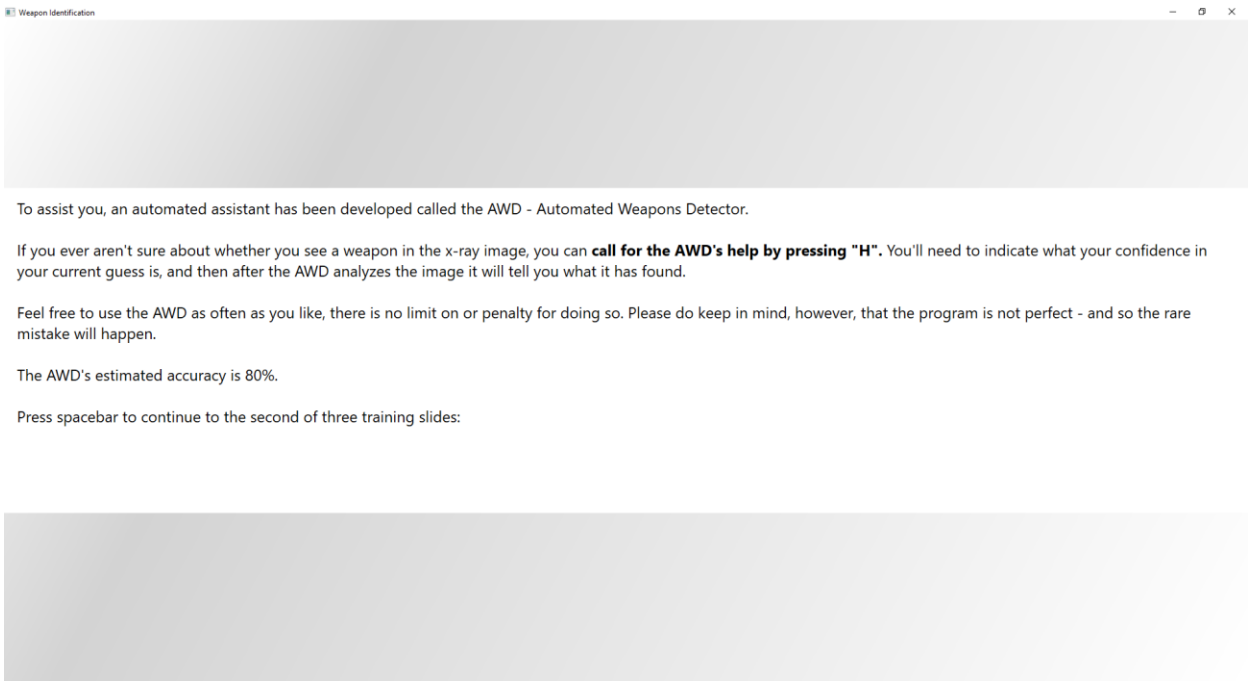


Figure C2

Tutorial Page 2: First Practice Stimulus

Figure C3

Tutorial Page 3a: Certain Condition Induction

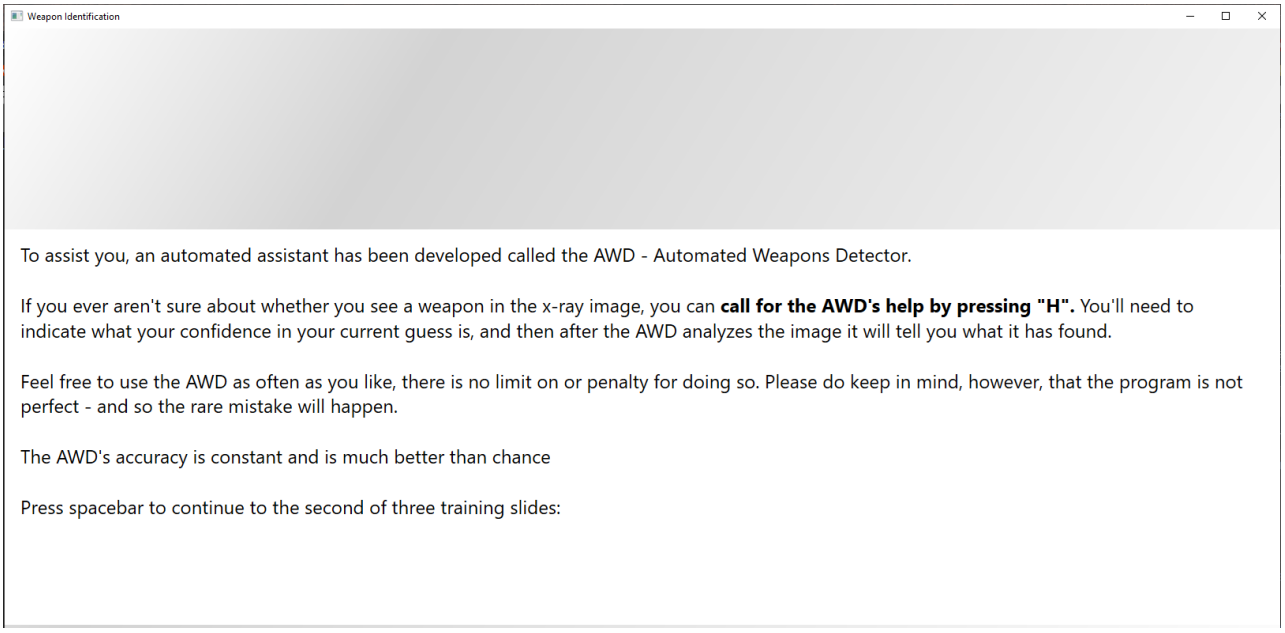


Note. This slide is only presented in the **low uncertainty condition**.

Figure C4

Tutorial Page 3b: Uncertain Condition Induction

**



Note. This slide is only presented in the **high uncertainty condition**.

Figure C5

Tutorial Page 4: Second Practice Stimulus

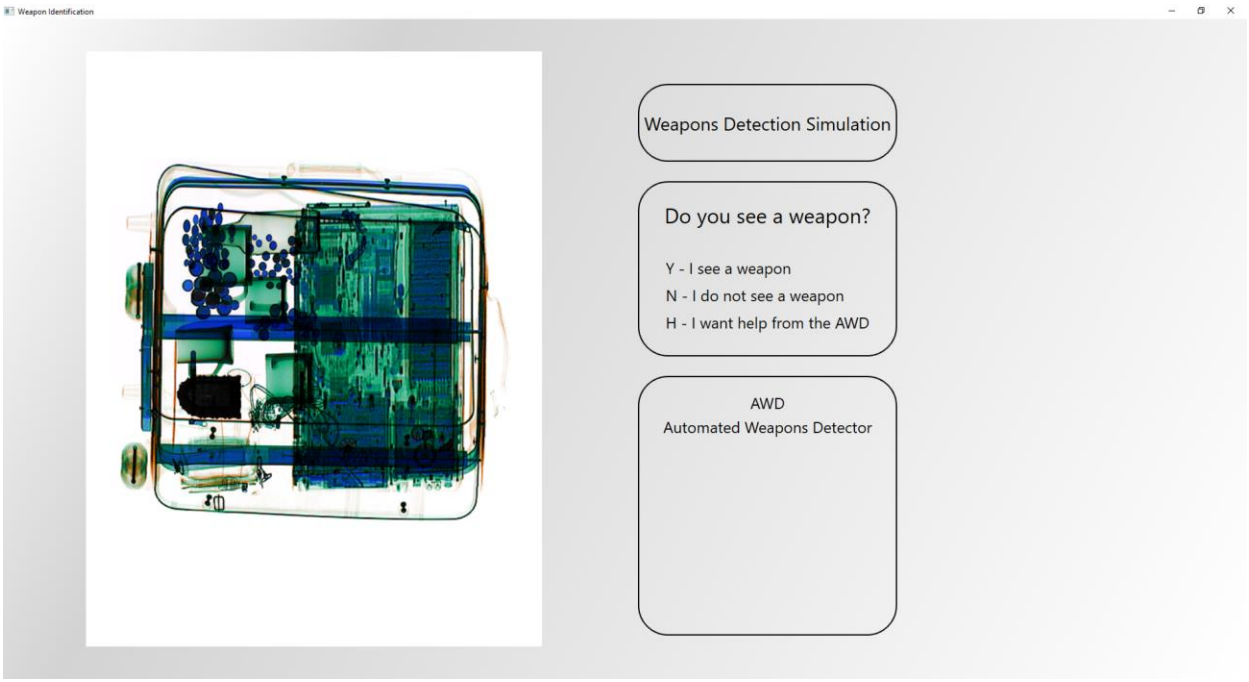


Figure C6

Tutorial Page 5: Reward System and Feedback Information

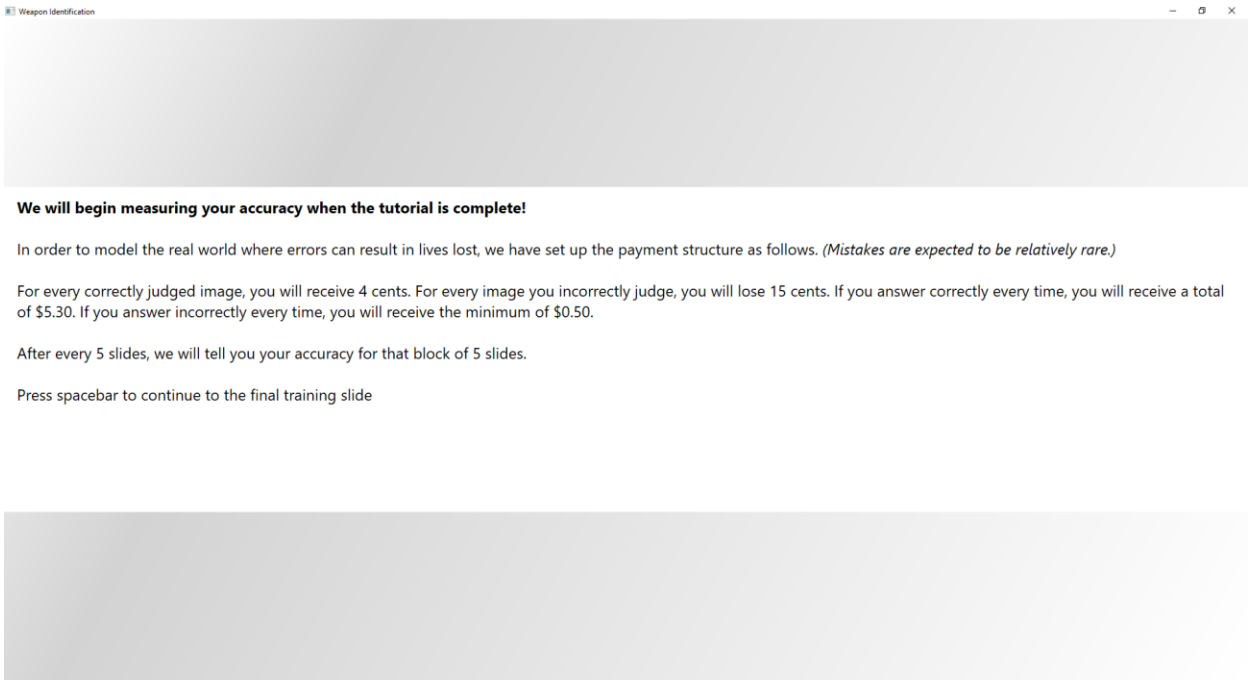
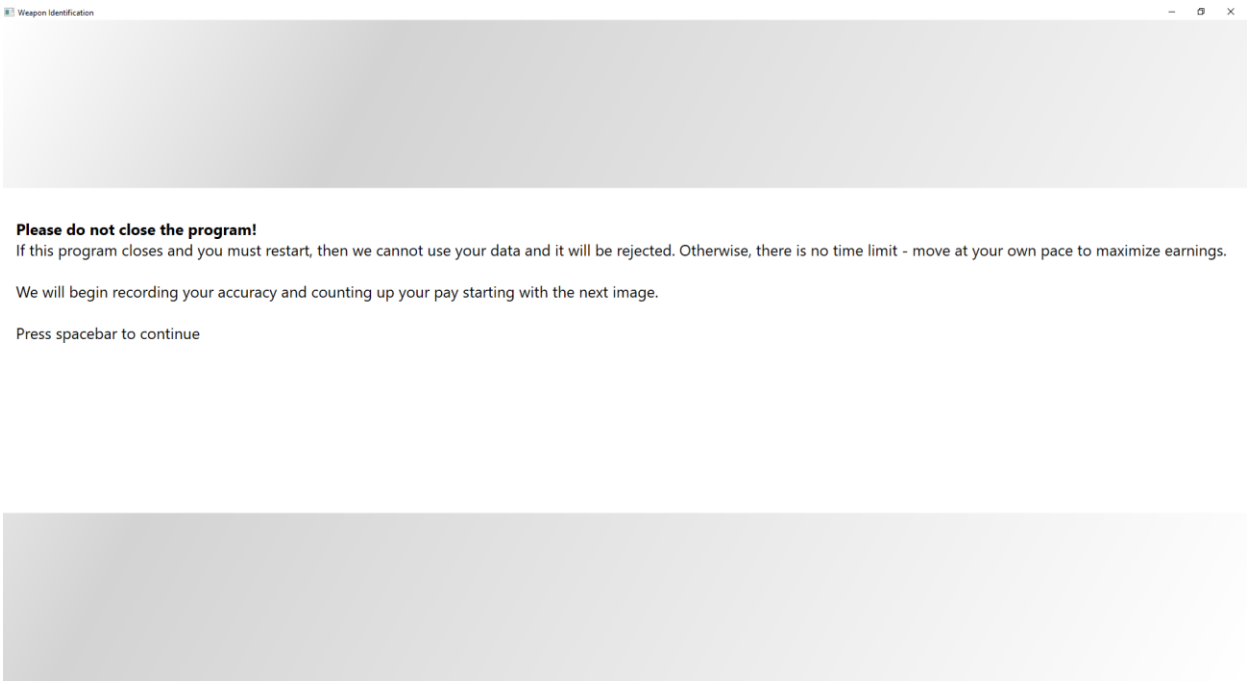


Figure C7

Tutorial Page 6: Final Practice Stimulus

Figure C8

Tutorial Page 7: Final Page of Tutorial



Appendix D

Automation-Induced Complacency Rating Scale

INSTRUCTIONS: Please indicate how much you agree or disagree with the following statements

1 (Disagree Strongly)……..……5 (Agree Strongly)

1. When I have a lot to do, it makes sense to delegate a task to automation.
2. If life were busy, I would let an automated system handle some tasks for me.
3. Automation should be used to ease people's workload.
4. If automation is available to help me with something, it makes sense for me to pay more attention to my other tasks.
5. [R] Even if an automated aid can help me with a task, I should pay attention to its performance.
6. Distractions and interruptions are less of a problem for me when I have an automated system to cover some of the work.
7. Constantly monitoring an automated system's performance is a waste of time.
8. [R] Even when I have a lot to do, I am likely to watch automation carefully for errors.
9. It's not usually necessary to pay much attention to automation when it is running.
10. Carefully watching automation takes away time from more important or interesting things.

Note. Items 1-5 reflect the alleviating workload facet of automation-induced complacency. Items 6-10 reflect the monitoring facet of automation-induced complacency. Items 5 and 8 are reverse-coded.

Appendix E

International Personality Item Pool 10-item NEO Openness Scale

INSTRUCTIONS: Please indicate how much you agree or disagree with the following statements

1 (Disagree Strongly)……..……5 (Agree Strongly)

"I…"

1. Believe in the importance of art.
2. Have a vivid imagination.
3. Tend to vote for liberal political candidates.
4. Carry the conversation to a higher level.
5. Enjoy hearing new ideas.
6. Am not interested in abstract ideas.
7. Do not like art.
8. Avoid philosophical discussions.
9. Do not enjoy going to art museums.
10. Tend to vote for conservative political candidates.

Note. Items 6-10 are reverse coded.

Appendix F

International Personality Item Pool 10-item NEO Conscientiousness Scale

INSTRUCTIONS: Please indicate how much you agree or disagree with the following statements

1 (Disagree Strongly)……..……5 (Agree Strongly)

"I…"

1. Am always prepared.
2. Pay attention to details.
3. Get chores done right away.
4. Carry out my plans.
5. Make plans and stick to them.
6. Waste my time.
7. Find it difficult to get down to work.
8. Do just enough work to get by.
9. Don't see things through.
10. Shirk my duties.

Note. Items 6-10 are reverse coded.

Appendix G

Goal Orientation

INSTRUCTIONS: Please indicate how much you agree or disagree with the following statements

1 (Disagree Strongly)……..……5 (Agree Strongly)

1. I often read materials related to my work to improve my ability.
2. I am willing to select a challenging work assignment that I can learn a lot from.
3. I often look for opportunities to develop new skills and knowledge.
4. I enjoy challenging and difficult tasks at work where I'll learn new skills.
5. For me, development of my work ability is important enough to take risks.
6. I prefer to work in situations that require a high level of ability and talent.
7. I would rather prove my ability on a task that I can do well at than try a new task.
8. I'm concerned with showing that I can perform better than my coworkers.
9. I try to figure out what it takes to prove my ability to others at work.
10. I enjoy it when others at work are aware of how well I am doing.
11. I prefer to work on projects where I can prove my ability to others.
12. I would avoid taking on a new task if there was a chance that I would appear rather incompetent to others.
13. Avoiding a show of low ability is more important to me than learning a new skill.
14. I'm concerned about taking on a task at work if my performance would reveal that I had low ability.
15. I prefer to avoid situations at work where I might perform poorly.
16. When I don't understand something at work, I prefer to avoid asking what might appear to others to be "dumb questions" that I should already know the answer to already.

Note. Items 1-5 reflect the learning facet of goal orientation. Items 6-11 reflect the prove performance facet of goal orientation. Items 12-16 reflect the avoid performance facet of goal orientation.

Appendix H

Feedback Orientation

INSTRUCTIONS: Please indicate how much you agree or disagree with the following statements

1 (Disagree Strongly)……..……5 (Agree Strongly)

1. Feedback contributes to my success at work
2. To develop skills at work, I rely on feedback
3. Feedback is critical for improving performance
4. Feedback from supervisors can help me advance in a company
5. I find that feedback is critical for reaching my goals
6. It is my responsibility to apply feedback to improve my performance
7. I hold myself accountable to respond to feedback appropriately
8. I don't feel a sense of closure until I respond to feedback
9. If my supervisor gives me feedback, it is my responsibility to respond to it
10. I feel obligated to make changes based on feedback
11. I try to be aware of what other people think of me
12. Using feedback, I am more aware of what people think of me
13. Feedback helps me manage the impression I make on others
14. Feedback lets me know how I am perceived by others
15. I rely on feedback to help me make a good impression
16. I feel self-assured when dealing with feedback
17. Compared to others, I am more competent at handling feedback
18. I believe that I have the ability to deal with feedback effectively
19. I feel confident when responding to both positive and negative feedback
20. I know that I can handle the feedback that I receive

Note. Items 1-5 reflect the utility facet of feedback orientation. Items 6-10 reflect the accountability facet of feedback orientation. Items 11-15 reflect the social awareness facet of feedback orientation. Items 16-20 reflect the feedback self-efficacy facet of feedback orientation.

Appendix I

Shipley Institute of Living Scale

INSTRUCTIONS: Complete the following. Each dash (-) calls for either a number or a letter to be filled in.

1. Z Y X W V U –
2. oh ho rat tar mood ----
3. 57326 73265 32657 26573 -----
4. tam tan rib rid rat raw hip ---
5. two w four r one o three -

INSTRUCTIONS: In the items below, the first word in each line is printed in capital letters. Each is followed by four other words. Indicate which word means the same thing, or most nearly the same thing, as the first word. If you don't know, guess.

6. CORDIAL swift muddy leafy hearty
7. RENOWN length head fame loyalty
8. JOCOSE humorous paltry fervid plain
9. LISSOM moldy loose supple convex
10. PRISTINE vain sound first level