ABSTRACT

PEARSON, CARL JAMES. How Cognitive Risk Types Influence Trust and Reliance Across Automation Stages. (Under the direction of Dr. Christopher B. Mayhorn).

Automation is an important feature of everyday life and more safety-critical situations. One paradigm to measure and describe human interactions with automation is through trust and reliance, respectively, an attitude and subsequent behavior. Risk is an important factor that influences how trust develops between a human and automation (Hoff & Bashir, 2013). Existing literature fails to adequately explore risk and its effect on trust in automation. These studies explored analytical risk (logical, probabilistic judgements), but not affective risk (immediate feelings of fear) as indicated to be critical in how risk affects cognition (Loewenstein et al., 2001). The first experiment manipulated analytical risk and affective risk in an x-ray luggage scanning task where participants were supported by an automated tool. Results indicated that trust was far more predictive of reliance at higher analytical risk levels (affective risk manipulations were unsuccessful).

Experiment two was intended to further generalize findings by manipulating automation stages, a taxonomy of how the participant can be supported by different automation. In a similar x-ray task, results indicated that trust only influenced by affective risk (not analytical risk). Trust was most predictive of reliance in a complex interaction. Trust was predictive of reliance, generally, where the automated tool was most congruent with the cognitive style of the operator. For example, in high affective risk conditions that induced an information-seeking cognitive style, the automated stage gave more information to the participant showed the strongest trust-reliance relationship. Generally, results showed that both risk types are critical to determine when trust in an automated tool is most predictive of behavioral reliance.

How Cognitive Risk Types Influence Trust and Reliance Across Automation Stages

by
Carl James Pearson

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina

2019

APPROVED BY:

Christopher B. Mayhorn Chair of Advisory Committee	Douglas Gillan
Anne. C McLaughlin	William Boettcher

BIOGRAPHY

Carl Pearson is a doctoral student an North Carolina State University in the Human Factors & Applied Cognition program of the Psychology department. He received his undergraduate degree from the University of Minnesota Duluth before moving to Raleigh, NC to achieve an MS in the same program as his doctoral program. In his academic work, his focus has consistently explored factors that influence the development of trust with automated aids, as well as how predictive that trust attitude is of eventual behavioral reliance on an automated aid. He has also worked in the industry at Red Hat software, focusing on quantitative measurement of usability benchmarks in enterprise programs, as well as strategic implementation of user experience research methods.

ACKNOLWEDGEMENTS

Many thanks to my advisor, Dr. Chris Mayhorn, for his consistent support and encouragement through the last four years. Your guidance was a critical reason I was able to achieve was I set out to do. Thanks also to my committee members, Dr. Anne McLaughlin, Dr. Douglas Gillan, and Dr. William Boettcher. Each helped to influence key areas of my dissertation research and my development more broadly as a psychologist over the past four years. Thanks to so many fellow students, in particular, Michael Geden for teaching me that statistics can be enjoyable in its own right. Thanks to the Crew, Bobby Sall and Steve Cauffman. Graduate school felt often like a group endeavor and I'm thankful you were my group. Thanks to my parents, Ron and Tricia, for instilling curiosity as a virtue since the beginning of my life and for the continued support. Finally, thanks to my fiancé, Laurel Oswald. You were my rock in this entire process, from application to graduation. It means the world to have you by my side.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Automation is an increasingly prevalent aspect of work in modern society. Automated tools exist in tandem with humans as a complement to the success (or failure) of an operator's task (Parasuraman & Riley, 1997). This human-automation relationship can involve a high cost of failure or a large benefit of success. This is easy to imagine in common examples such as air traffic control, x-ray baggage screening, or medical diagnostics. Therefore, the concept of risk is critical when understanding operators' attitudes and behaviors towards automated tools. The operationalization of trust and risk, as it was mainly adopted by human factors researchers, began within the field of organizational management.

Mayer, Davis, and Schoorman defined *trust* in their seminal work as how willing a person is to be vulnerable to the decisions or opinions of another person, without the trustor having control over the trustee (1995). This is delineated from *reliance* in their definition of trust: reliance is the behavior that ultimately puts the trustor in the potential way of harm. Another key aspect of contexts involving trust is that the trustor stands to lose something and may or may not fail to perform adequately. This is often described as *risk* by Mayer and colleagues (1995).

This definition of trust has had a major influence in the field of human factors, and how operators use trust to guide behavior related to specific technology like online shopping (Gefen, Karahanna, and Straub, 2003). Human-to-human definitions of trust have been adapted so that they apply to interactions with agentive technologies, such as automation and computers.

Trust in Automation

Automation describes a broad set of technologies. Common definitions include the "execution by a machine agent (usually a computer) of a function that was previously carried out

by a human" (Parasuraman & Riley, p. 231, 1997), and "technology that actively selects data, transforms, information, makes decisions, or controls processes" (Lee & See, p. 50, 2004). Given that humans often treat computers as social actors (Nass, Steuer, & Tauber, 1994), the social process of trust is an important determinant of how humans ultimately behave with automation.

A recent and comprehensive model of human-automation trust was put forth by Hoff and Bashir (2013). The model separated three layers to conceptualize trust variability: dispositional trust, situational trust, and learned trust. Dispositional trust includes personal operator characteristics such as culture, age, gender, and personality. Learned trust involves evaluations of the operator drawn from past or current interactions with the automation. Situational trust involves a subdivision that becomes critical for our questions in this study.

Situational trust, as conceptualized by Hoff and Bashir (2013), involves external aspects (e.g., system complexity, task type, etc.) and internal variability (e.g., mood, attention, self-confidence, etc.). These factors do not merely point to an increase or decrease in trust, but also a change in the relationship between how predictive trust can be of behavioral reliance. Of the many influencing factors to situational trust, the conceptualization of *risk* may have been underserved by extant human factors experimental literature.

Risk and Automation Trust: What has been missing?

Some human factors laboratory research has investigated how risk affects trust and reliance with automation by manipulating risk. Risk has traditionally been characterized as a situational factor across all the existing automation-trust/risk research, which likely led to the characterization of risk as such in Hoff and Bashir's model of influencing factors: risk is an external variable of the environmental factor of trust influences (2013). One of the earliest experiments used a GPS route task where risk was defined as "a situation in which there is a

probability of harm or loss" (Perkins, et al, 2010, p. 2131). Subsequent studies also defined risk as a situational factor centered around the probability of negative or positive outcomes (Lyons & Stokes, 2012; Satterfield, et al., 2017). Outcome probability is certainly a critical aspect of risk, as shown in models such as expected-utility theory (EU). EU theory posits that choices in risky situations can be predictably assessed by decision-makers who weigh probabilistic outcomes, though sometimes with bias (Mongin, 1997).

While some theories are not explicit about underlying processes within an EU-type framework, those that explain the cognitive processes do so through explicit algebraic processes around probability (Loewenstein et al., 2001). Human factors research has found confirmatory results to this idea (Ezer, Fisk, & Rogers, 2008). However, significant lines of related research have explored another critical factor in risk that is distinct from logical, probabilistic judgements: *risk-as-feelings* or affective risk.

While the probabilistic definition of risk is not wrong, it only presents a partial scope of how cognitive aspects of risk could influence operator trust and reliance with automation. Extensive work exists that shows how risk exists not only as a cognitive evaluation of probability, but also as an influence of emotional components such as fear or dread (Loewenstein et al., 2001; Slovic et al., 2007). And while affect is present in cognitive evaluations through *anticipated emotions* like expectation of loss or gain, Loewenstein and colleagues (2001) posit that these anticipated emotions are distinct in their effect on decision-making compared to *anticipatory emotions*, such as immediate fear and dread. Thus, it is not entirely novel to tie in risk as affect to automation trust/reliance, given that affect has already been shown to be an influence in trust attitudes and reliance behavior outside of risk (Merritt, 2011). In essence, while Hoff and Bashir (2013) present risk as a situational/external factor tangentially related to

cognitive evaluations of probability, multiple dual-process approaches demonstrate that risk exists both as an analytical (probabilistic) evaluation of situational outcomes and an immediate affective (emotional) response to situational stimuli.

The affect heuristic (Slovic et al., 2007) has shown how affective risk influences are not only linked to final behavioral outcomes directly, but also influence the concurrent cognitive evaluations of probabilistic judgements. Further, research on the related risk-as-feelings hypothesis has shown that affective risk is critical to decision-making behavior, and factors that influence analytical risk assessment may be semi-related or unrelated to factors that influence affective risk. It is important to consider both affective and analytical risk, so existing empirical literature on risk will be considered through this extended viewpoint and may shed new light on previously reported results.

Before empirical research is examined, here we are explicit about the two constructs we have discussed. First, analytical risk will use a construct drawing again from outside literature. Boettcher (2004) provided a robust examination of how risk has been used across domains. In the end of this examination of constructs, he defined an integrated construct of risk that retains the phenomenological aspect sometimes ignored in past human factors work while maintaining key situational features that are present in cognitive evaluations. Higher analytical risk has three main characteristics, (a) a larger frequency/presence of divergent outcomes, (b) a perception of the possibility of extremely negative outcomes, and (c) an understanding that the estimates of probability surrounding those outcomes could be flawed or incorrect. This works theoretically in tandem with affective risk. Affective risk is defined conceptually as *anticipatory* emotions, such as fear, that are present in the decision-maker's immediate affective state leading up to and during the decision moment (Loewenstein et al., 2001).

Empirical risk research related to trust in automation

To our knowledge, the earliest study manipulating risk and automation reliance involved a representative design study that gave participants an automated tool to support their task in a geometric guessing game (Ezer et al., 2008). Risk was manipulated as the cost of an error to the participant's overall points, as well as cost of verification of the automation's recommendation. The results showed a significant effect in cost of error, such that reliance decreased as the cost (risk) was higher. Conversely, cost of verification was related to an increase in reliance. These risk factors were essentially cognitive, and emotional components would have been anticipated, not anticipatory. The game did not prime any affectively fearful context. In addition, this study did not measure trust. Despite this, the supported patterns of reliance can inform the larger picture of risk and automation trust/reliance.

Another previously mentioned study investigated trust and reliance on automation, in a situation that could be considered more affectively risky (Perkins et al., 2010). In general, they found that trust was most related to reliance at the high risk conditions. In addition, they found an interesting trend related to affective risk. Within a single condition involving no manipulation in analytical risk, there was more reliance involving a hazard with a mortality aspect (car crash) than without (traffic jam). They did not note any dual process risk theories; however, it seems the sudden increase in affective risk may have overtaken the otherwise consistent analytical risk considerations. It is important to note that this differed from a higher risk condition result, where higher mortality hazards resulted in less reliance. However, this higher affective component in the higher risk condition covaried with analytical risk increase. Therefore, it is challenging to know what ultimately contributed to the results.

A later study manipulated risk to compare trust and reliance between a human and automated aid in a military convoy route choice simulation (Lyons & Stokes, 2012). They did not measure reliance directly on the automation, but a decrease in reliance on the human in high risk situations has been interpreted as an increase in reliance on the automation in the same condition (Hoff and Bashir, 2013). Reliance intentions (a close analog to trust) did not differ across risk in either the human or automation. This study's risk manipulation shows the opposite effect of risk on reliance compared to previous findings that greater risk decreases automation reliance overall (Ezer et al., 2008; Perkins et al., 2010). Lyons and Stokes' (2012) results indicated greater reliance on the automation in high risk scenarios.

Given this pattern of conflicting results, Satterfield and colleagues (2017) identified existing challenges in studying risk within a laboratory framework and presented a novel methodology to manipulate risk in a trust-in-automation context. They adapted a behavioral economic framework of an ultimatum game, presenting greater monetary value to be lost with poor task outcomes. This method was motivated by the desire to increase the tangibility of analytical risk, something earlier studies may have lacked; other studies did not involve any personal impact to participants outside of intrinsic motivation. In this study, the automation was a largely independent teammate in a dynamic gaming task, and therefore, reliance was not as simple as choosing the same recommendation. Reliance could be characterized within their measure of how often a participant 'checked' on the automation, rather than allowing it to operate independently. Results showed that reliance decreased (or 'checking' increased) in higher risk conditions. It should also be noted that this behavior of checking was postulated (but not measured) to be 'behavioral trust' as opposed to reliance. Therefore, the exact construct was somewhat unclear in regards to participant behavior around the automation.

These four studies have all contributed to a growing base of knowledge regarding how risk is related to trust/reliance with automation, albeit with somewhat conflicting results. They have strongly focused on analytical risk, but the lack of control or consistency in affective risk could be a critical variable related to inconsistent findings across studies. Implementing tighter control on affective risk was the motivation for our first experiment.

H1a: Reliance is expected to lower as analytical risk increases

H1b: Reliance is expected to increase as affective risk increases

H2: Reliance is expected to be higher when both risk types are high than when both are low.

H3a: Trust is expected to increase as affective risk increases

H3b: Trust is expected to remain the same across analytical risk changes

H4: The relationship of trust and reliance is expected to strengthen as affective risk increases.

Automation Levels

In addition to affective risk inconsistencies, there is another feature of these past experiments that may be related to a lack of consensus on trust and reliance in automation when risk is involved: stages of automation. Automation is not all-or-none; it exists across a spectrum of support and autonomy (Parasuraman, Sheridan, & Wickens, 2000). This can range from a fully autonomous teammate that a human operator cannot control, or a simple tool that displays potential options on which a human operator must ultimately decide. The model put forth by Parasuraman, Sheridan, and Wickens (2000) builds on multiple earlier models to robustly describe automation in relationship to a human operator in analog to the stages of a human information processing model. These four stages include information acquisition (stage one),

information analysis (stage two), decision selection (stage three), and action implementation (stage four). A meta-analysis by Onnasch and colleagues (2013) has shown that the degree that automation supports an operator is directly related to operator reliance and performance on a task with that automation. Hoff and Bashir (2013) have also pointed out in their literature review that automation complexity and an operator's decisional freedom might be related to the relationship of trust and reliance in an automated tool, especially in risky contexts.

Research in risk and automation trust/reliance is scattered across the spectrum of automation stages. Lyons & Stokes' (2012) automated tool existed at stage one, information acquisition: it did not recommend a route explicitly but did show what options existed. Perkins and colleagues' (2010) automation was a similar map tool but did give a recommendation. It existed partially as stage one when it showed possible streets and hazards, and at stage three (decision selection) when it gave its recommended route. The automation in the experiment by Ezer and colleagues (2008) gave support at stage three, providing a single decision recommendation. In the recent experiment by Satterfield and colleagues (2017), the automation was at the latest stage of automation, stage four. As an autonomous teammate, it had the expectation (with varying reliability) to be able to fend off enemies by controlling multiple defensive vehicles without human input on decision selection.

It was unclear from existing research in what direction and to what degree automation stages are related automation trust and reliance in risky situations. Not only do results vary across stages between experiments, the experimental risk paradigms are varied enough to only be able to strongly connect the results to the automation stage used within a study. These issues motivated our second experiment on how automation stages are related to trust and reliance in risky conditions.

H5: Reliance is expected to be higher in higher risk conditions

H6a: Trust is expected to be higher in high risk conditions

H6b: Trust is expected to be higher in late stage automation conditions

H7a: The relationship of trust and reliance is expected to be weaker in early stage automation conditions

H7b: The impact of high risk conditions is expected to intensify the impact of automation stage's effect on the trust-reliance relationship

Experiment One

This experiment placed participants pseudo-randomly in a 2x2 design (analytical risk level by affective risk level) via block randomization. High versus low analytical risk was manipulated by how much money participants could gain ($10 or $1). High versus low affective risk was manipulated by how fear-inducing an initial priming video is, and by how fear-inducing the images that accompany decision points were (negative valence/high arousal or neutral valence/low arousal). This was intended to differentiate how different types of cognitive risk (analytical and affective) would change the relationship of trust and reliance on automation.

Participants

We recruited sixty-five participants via Amazon's Mechanical Turk (MTurk) website. This sample was roughly double that of the study that first applied the behavioral economic paradigm to risk and automation trust (Satterfield et al., 2017). Mturk has been shown across multiple studies to be reliable and valid in collecting experimental data, compared to more traditional sampling methods (Buhrmester et al., 2017; Lenz et al., 2014). To avoid cross-cultural differences, geographic location was constrained to the United States. Participants' base compensation was $2.00.

Materials

In a similar paradigm as other trust in automation work (Merritt, 2011; Merritt & Ilgen, 2008), an x-ray baggage scanning microworld task was used. Images were taken from the Kedlin Company's *Airport Scanner* app, which has been used in human performance experiments previously (Mitroff & Biggs, 2014). Participants performed a signal detection task for contraband in the x-ray images while being supported by an automated tool (example images in Appendix item 1). The base rate for present contraband was at approximately 30%, again based on previous research paradigms (Merritt, 2011; Merritt & Ilgen, 2008).

Emotion Elicitation

Two forms of emotional elicitation were used. First, participants were primed in conditions to induce a fearful or neutral affect. The pre-task video priming used the video database from Schaefer and colleagues for stimuli (2010). The fearful videos were taken from the climax moments in horror moving (*The Shining*, *The Blair Witch Project*). The neutral videos were taken from highly naturalistic scenes of daily life, also from a Hollywood film (*Blue*), depicting things such as a walking into a room to take a file out of a drawer. Tables are provided in appendix item 2 that give descriptions from the original study.

Specific emotional responses to video clips were measured in their database, and only those 15 images highest in fear responses were used in the high affect condition (appendix item 3). This was in congruence with the risk-as-feelings hypothesis construct of affective risk as fearful emotions. Showing affect-specific videos as a prime for automation-use experiments has been used previously by Merritt (2011). To tie affect temporally to the decision-making, images appeared at the time of decision-making input in the x-ray task. In high affective conditions, images taken from the OASIS image database were used (Kurdi, Lozano, & Banaji, 2016).

Images with high arousal and negative valence that depict explosions, fire, and aviation crashes were shown during decision- making input. In low affective conditions, images with neutral valence and low arousal were shown.

Microworld task

The microworld task was given through a series of dynamic web pages in *Qualtrics* and had point values assigned to various outcomes and actions. A baggage x-ray image was presented to participants for one second. The following page allowed for participants to clear it (no contraband), check it (contraband present), use the automation outright, or peek at the automated answer. If 'peek' was chosen, a secondary decision screen appeared. This screen contained the options to 'check' or 'clear', but one option was indicated to be recommended by the automation. Fifteen x-ray images and their subsequent decision pages were shown in random order with four blocks (totaling 60 x-ray images).

Each correct answer was worth 10 points. Each incorrect answer was worth 0 points. This varied somewhat depending on the use of the 'peek' function, as described in the next section.

Automation

The automation supported the participant by automatically choosing to 'check' or 'clear' the bag or letting the participant 'peek' at the automation recommendation first before relying on it (appendix item 4). This qualified as a stage four automation type when not checking, or stage three automation type when checking the recommendation first (Parasuraman, Sheridan, & Wickens, 2000). The automation featured a reliability level of approximately 75%, serving as moderate reliability that was halfway between the high and low conditions used by Merritt and Ilgen (2008). There was a slight variation of reliability by trial type of contraband-present compared to contraband-clear; the respective reliability levels were 72% and 76%.

To further validate Satterfield and colleagues' measurement of *behavioral trust* (2017), participants chose whether to rely on the automation without seeing the automation response first or costing some potential points and verifying what the automation's recommendation was. Checking automation responses before relying on them cost 10% of point potential. Overuse of automation blindly was controlled by (1) a 10% point cost to use automation or checking of automation and (2) the information told to participants that total reliance on automation would not yield the critical bonus score of 500 points.

Analytical manipulation reward values

The amounts involved in the high vs. low analytical risk conditions were $10 and $1, respectively. Previous research has shown these values (in euros) as being significantly different in their expected unpleasantness (Harinck et al., 2007). This unpleasantness expectation would fall within the construct of *anticipated* emotions, not anticipatory emotions, fitting our construct of analytical risk. This same study showed the importance of framing these denominations as a gain rather than loss. At these denominations of money, Harinck and colleagues found that gains were seen as more emotionally impactful than losses (2007).

Measures

Behavioral reliance was defined as when a participant used the automation outright or chose the recommendation of the automation after peeking; non-use was when the participant did not use the automation at all or did not choose the automation's recommendation after peeking.

Behavioral trust (BT) was defined by Satterfield and colleagues (2017) within the context of their specific automation task paradigm as a checking behavior when the participant would refocus her or his attention in a dynamic task to take stock of automation performance and override automation behavior. This was postulated to not include bias generally found in self-

report subjective measures, but no analysis was done to compare BT to subjective trust (ST) or reliance. This experiment, given its different automation type, was able to measure participant BT by the amount of times a participant peeked at the automation answer (like the checking behavior with stage four automation in the past experiment).

ST was measured using a self-report scale from Merritt (2011). This scale was given at multiple points throughout the task to measure subjective trust over time. While other scales for trust exist, this inventory has been used in similar x-ray baggage paradigms where it was given to participants multiple times across a single experimental session. At six items, it was also briefer than other scales that go above 10 items (Bisantz & Seong, 2001); this was advantageous when the scale was given multiple times.

Trust propensity was measured once by adapting a scale from Singh, Molloy, & Parasuraman (1993). The *trust* and *reliance* factor-loading questions were used resulting in six questions about general trust propensity.

Manipulation check

To ensure that participants felt differently across affect conditions, the *Positive and Negative Affect Schedule* (PANAS) was administered at multiple points (Watson, Clark, & Tellegen, 1988). This measure has been widely validated and allowed for multiple time-span instructions (Coan & Allen, 2007). Because in-the-moment assessments of emotion are often more accurate than retrospective assessments, the PANAS was given immediately after the video priming, before all of the subsequent x-ray blocks, and once after all the x-ray trials were complete (totaling five times).

Procedure

Participants arrived at the experiment page through the Mechanical Turk task assignment page and directed to a condition via block randomization. After obtaining consent, participants were given a brief overview of the tasks and instructions on what to look for in the x-rays. They were then given a brief practice opportunity. First, they chose between two isolated items to decide what was or was not contraband; there were five individual choice trials. Then, in a similar fashion to the eventual experimental trials, they completed a trial of two slides to see how the process worked with full x-ray images. The automation was not present in this practice. Before the automation description, participants were given the analytical manipulation by describing the point system and the bonus amount they could win.

Participants were instructed on characteristics of the automation, and how to use it. With 60 timed trials, in blocks of 15 to be consistent with the paradigm by Merritt and Ilgen (2008), the maximum amount of points was 600 points. To have participants strive for optimal performance under analytical risk conditions, participants were told that they would obtain their reward only if they get above 500 points in the trials, and that relying on automation entirely would not result in that point amount to encourage thoughtful automation engagement.

After instructions, participants completed a check to ensure proper audio function and then viewed the affect priming video. After the video, they rated their affect on the PANAS and completed the automation trust questionnaire. Participants completed the scored trials in four blocks of 15 slides. Before each block, participants completed the trust questionnaire and the PANAS. Once the trials were completed, there was a final trust and PANAS administration.

Participants were told that they would receive their bonus regardless of their score, upon completion of the experiment. All participants had the option to view a *happy* video (Schaefer et

al., 2010) to ameliorate any fear conditioning. After this, participants completed the trust propensity questionnaire, viewed the experimental debrief, and received their payment code.

Results

Data were collected over a period of four days. Sixty-five participants completed the experiment (see Table 1). We oversampled with the expectation of removing some participants who demonstrated low accuracy. However, all participants demonstrated sufficient accuracy (greater than 50% accurate), so no participants were removed. Gender was even overall (male, n = 33) but subgroup consisted of larger differences across participant genders. The average age was 34 and ranged from 21-61. While PANAS and trust measures were collected after the final x-ray block, these were not used in analysis as they did not predict any subsequent x-ray trial reliance outcomes. This means that only the first four of the five total subjective response data points were used.

Table 1.

Descriptives by conditions.

Condition	Gender (Female/Male)	Mean Age
High Analytical / High Affective	14 / 2	39
High Analytical / Low Affective	4 / 12	32
Low Analytical / High Affective	6 / 12	34
Low Analytical / Low Affective	8 / 7	33

Manipulation check

PANAS sub-questions that referred to negative emotions were averaged per participant at each block. A random intercepts model (similar to an RM ANOVA but allowing for varying group sizes) was conducted using the ‘lme4’ package in R to assess if the emotional elicitation

method was effective, as seen in figure 1 (Bates, Maechler, Bolker & Walker, 2014; Raudenbush & Bryk, 2002). *P* values to assess statistical significance were given via the 'lmerTest' package that is built upon the 'lme4' package (Kuznetsova, Brockhoff , Christensen, 2017); R^2 values were obtained using the 'sjstats' package that is also built upon 'lme4' (Lüdecke D, 2019). The dependent variable was the continuous measurement of PANAS average scores by block, and the predictor variable was the categorical variable of affective manipulation. This method of manipulation check assessment was used by Merritt (2011) across multiple points in time during their similar x-ray experiment that manipulated positive affect.

$$\text{Level 1: Negative\_Affect}_{it} = \beta_{0it} + r_{it}$$
$$\text{Level 2: } \beta_{0i} = \gamma_{00} + \gamma_{01}(\text{Affective}) + u_i$$

Figure 1. Model formulation for the manipulation check.

Assessing Model A detailed in table 2 (descriptives in appendix item 5 and diagnostics in appendix item 6), the results indicated that the manipulation was not effective in increasing negative affect for those in the high affective risk group compared to the low affective risk group. Unexpectedly, there was a nonsignificant increase in negative affect in the low affect group compared to the high affective group (γ_{01} = .19, t = -1.02, p = .371), as is generally seen in figure 2. The marginal effect size, considering fixed effects only, was very low with R^2 = .01.

Table 2.

Experiment one manipulation check results.

Fixed Effects	Model A
High affect intercept, γ_{00}	.544*** (0.15)
Low affect intercept, γ_{01}	0.190 (0.21)
Random Effects	**Value**
Between-person variance (τ_{00})	.66
Within-person variance (σ^2)	.23

*p<.05, **p<.01, ***p<.001

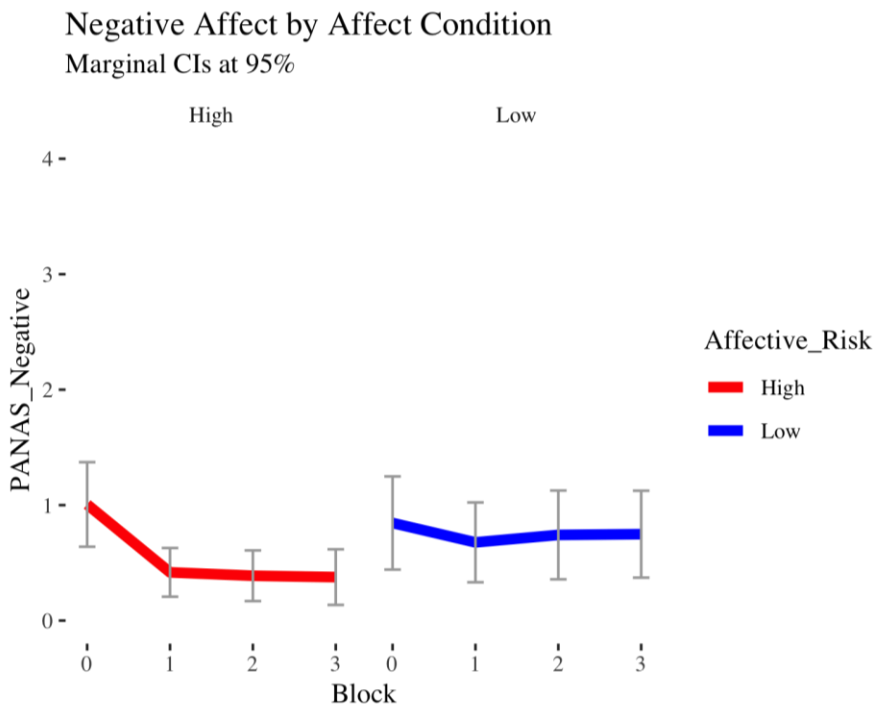


Figure 2. Negative PANAS questions across blocks by condition in experiment one.

Because the manipulation in this experiment was not demonstrated to be effective, subsequent analyses involved the affective predictor in the formulation but relegated its role to merely control for the related variance: interactions with analytical risk were not included in the analyses.

Behavioral Metrics

Descriptive metrics were analyzed to discern the difficulty of the x-ray task and the general behavioral trends. The table below (table 3) summarizes between-group metrics. Overall metrics are discussed subsequently.

Table 3.

Descriptive behavioral statistics with continuous standard deviations.

Metric	High affective risk			Low affective risk		
	Mean	SD	Range	Mean	SD	Range
Accuracy	75%	7%	58% - 85%	80%	6%	60% - 92%
Score	444	41	360 - 510	470	36	364 - 540
Reliance	23%	20%	0% - 65%	19%	24%	0% - 95%
Behavioral trust	6%	14%	0% - 67%	2%	6%	0% - 22%

On average, participants were 77% accurate (SD = 7%, range = 58% - 92%). Participants were given feedback of their ongoing total score during the experiment (out of 600 points), and only nine participants achieved the 500-point cutoff. The average score (the distribution seen in figure 3) was 457 (SD = 40, range = 360 - 540).

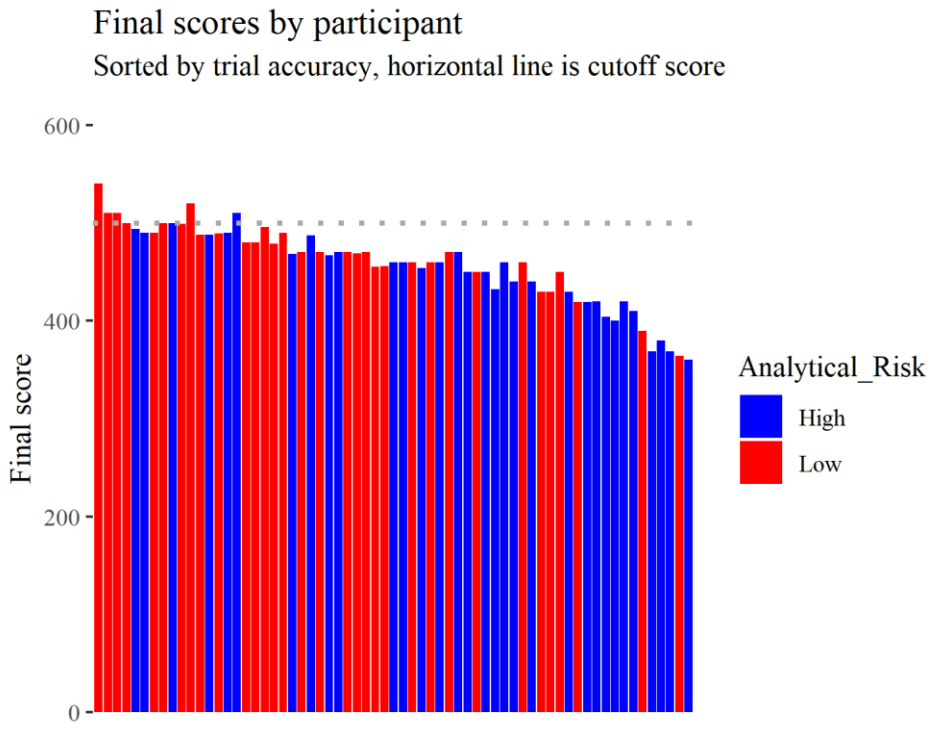


Figure 3. Final scores by participant, with bars sorted by accuracy proportions.

Reliance was measured across sixty x-ray slide trials per participant. On average, participants relied on the automation 21% of the time (SD = 22%, range = 0% - 95%). A simple non-parametric z-test conducted on the proportions between analytical risk groups, from the base '*stats*' package, indicated significantly more reliance in the high analytical group: $\chi^2 = 6.29$, $df = 1$, $p < .05$ (R Core Team, 2013). However, there was a large amount of individual variance between participants with a reliance proportion range from 0% to 95% (figure 4), so simplistic aggregated tests (like a z-test) may not properly account for such variation.

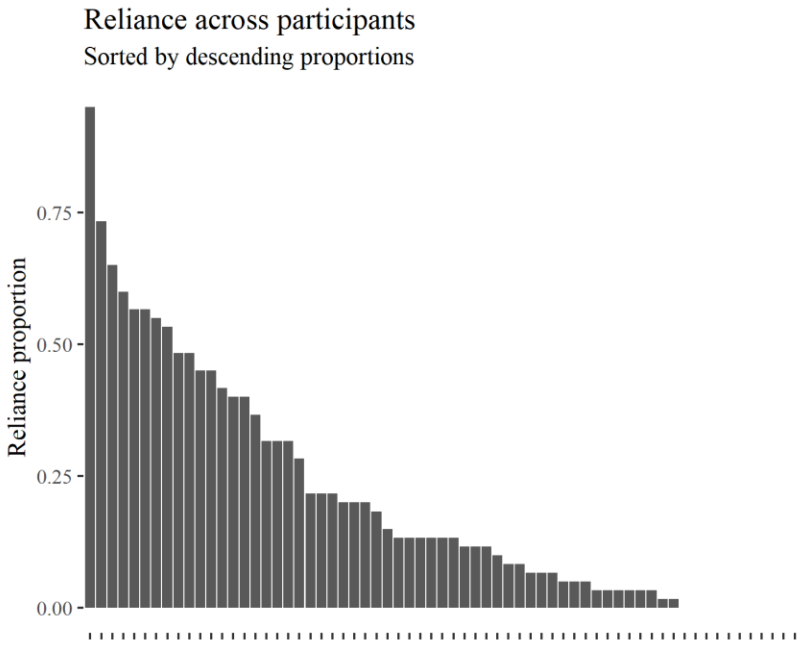


Figure 4. Experiment one reliance proportions across participants

Generally, behavioral trust (BT), or peeking, was rare with only 5% (SD = 11%, range = 0% - 67%) of trials resulting in peeking behaviors (figure 5).

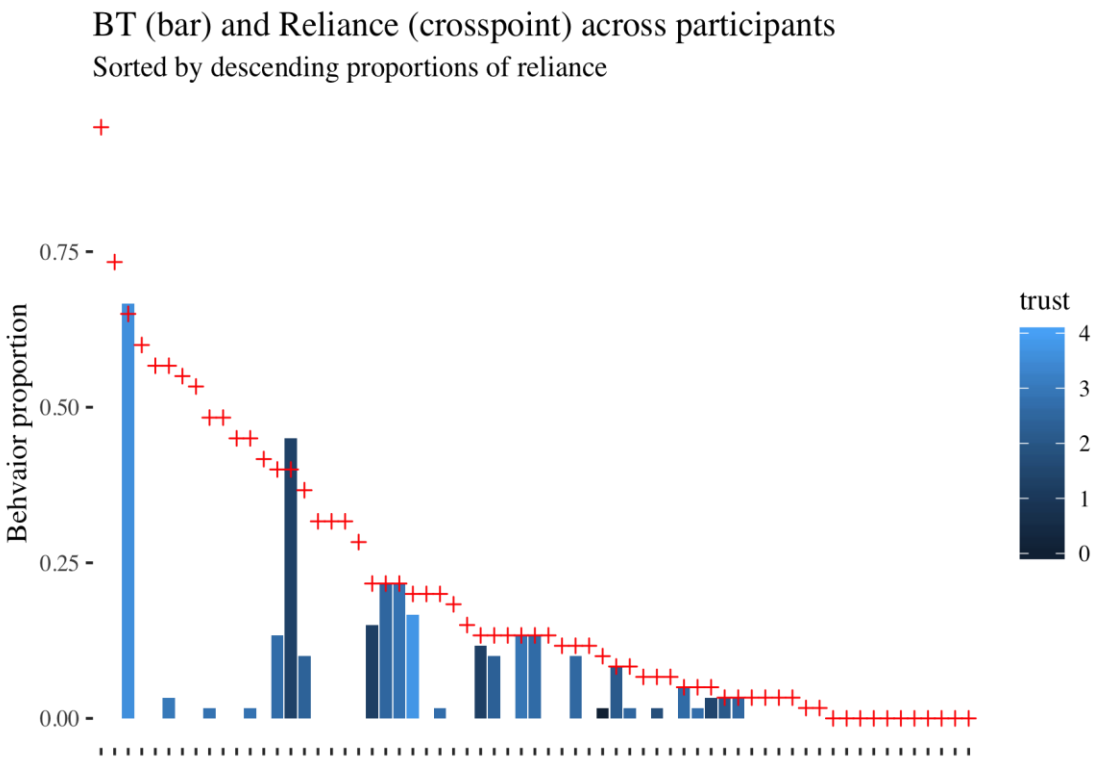


Figure 5. Reliance and BT proportions across participants with trust

Behavioral metrics indicated that individual strategies largely varied in usage of the automation, across both immediate reliance and peeking behaviors. Metrics of accuracy and score also indicated the task was sufficiently difficult, as the cutoff score was only reached in 14% of participant cases (though many more were close).

Trust

Trust questionnaires (Merritt, 2011) were averaged per participant at each block (descriptives in appendix item 7 and diagnostics in appendix item 8). A random intercepts model was conducted in R as formalized in figure 6 (Model B). The continuous dependent variable was average trust score at each block per participant. The categorical independent variables were risk manipulations (analytical and affective). The continuous predictor variable was block, or each time point of trust measurement before subsequent x-ray trials. This resulted in a two-level

model with trust measurements across time (level 1) and experimental manipulation grouping (level 2). The same R packages as the manipulation check model were used to create and analyze the models. A traditional motivation for this multilevel approach is accounting for large amount of variance between these levels, calculated with an intraclass correlation coefficient (Raudenbush & Bryk, 2002). This calculation indicated a moderate partitioning between levels with 72% of the variance related to individual variation in trust scores (level 1) and 28% of the variance between participants (level 2). This moderate partitioning of variance was a good motivation for use of our method of analysis, but we would have continued with it in any case as it allowed us to incorporate different group sizes and model time (block) as a continuous predictor (Jaeger, 2008).

Level 1: $Trust_{it} = \beta_{0it} + \beta_{1it}(Block) + r_{it}$

Level 2: $\beta_{0i} = \gamma_{00} + \gamma_{01}(Analytical\_high) + \gamma_{02}(Affective\_high) + u_i$

$\beta_{1i} = \gamma_{10}(Block) + \gamma_{11}(Analytical\_high)$

Figure 6. Model B formula for trust mixed effects model.

Results (table 4 and figure 7) indicated that trust did not differ significantly due to analytical risk manipulations when controlling for affective manipulation differences and change over time ($\gamma_{01} = .16$, $t = .41$, $p = .68$). However, there was a significant effect of time on trust, such that trust decreased slightly over time ($\gamma_{10} = -.07$, $t = -2.33$, $p < .05$). The marginal effect size was small with $R^2 = .05$.

Table 4.

Trust model results in experiment one

Fixed Effects	Intercept	Std. Err.	Intercept	Std. Err.
	Model B		Model C	
Intercept, γ_{00}	2.68***	.20	.255***	.19
Block, γ_{10}	-.09*	.04	-.07*	.03
Analytical * Block, γ_{11}	.05	.06		
Block * Trust propensity interaction, γ_{12}			.13*	.05
High analytical, γ_{01}	0.16	0.21	.11	.22
High affective, γ_{02}	-.38	0.21	-.15	.22
Trust propensity, γ_{03}			.39	.22
Random Effects	Value		Value	
Between-person variance (τ_{00})	.67		.25	
Within-person variance (σ^2)	.26		.59	

*p<.05, **p<.01, ***p<.001

Because of the affective manipulation components' failure to impact participants, these results cannot indicate any information around hypotheses 3a. As was expected in hypothesis 3b, trust remained relatively static across changes in analytical risk.

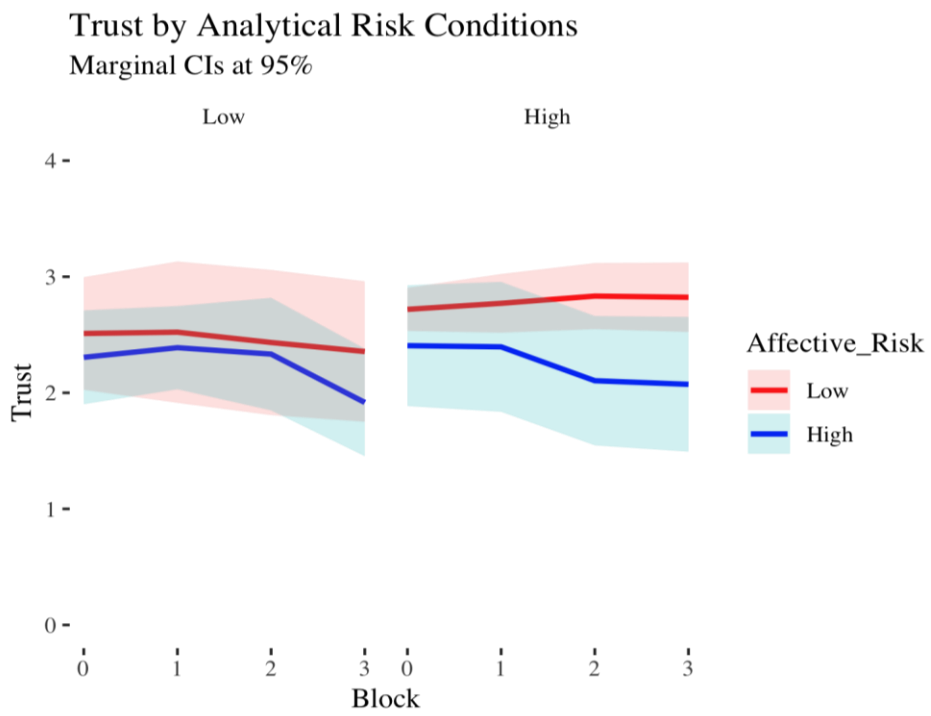


Figure 7. Trust across time and conditions in experiment one.

In addition to the planned analyses, an exploratory model (Model C, appendix item 9) was created by adding an additional predictor of trust propensity scores measured once per participant and an interaction of trust propensity by time block. Trust propensity was centered by overall score mean across participants. This model was markedly higher in its fixed-effects effect size with marginal R^2 = .15. No changes to significance results or beta directions were present in previously existing variables. Propensity on its own was (marginally) not significant (γ_{02} = .39, t =1.82, p = .07). However, there was an interaction present between block and trust propensity, as seen in figure 8 (γ_{11} = .13, t = 2.53, p < .05).

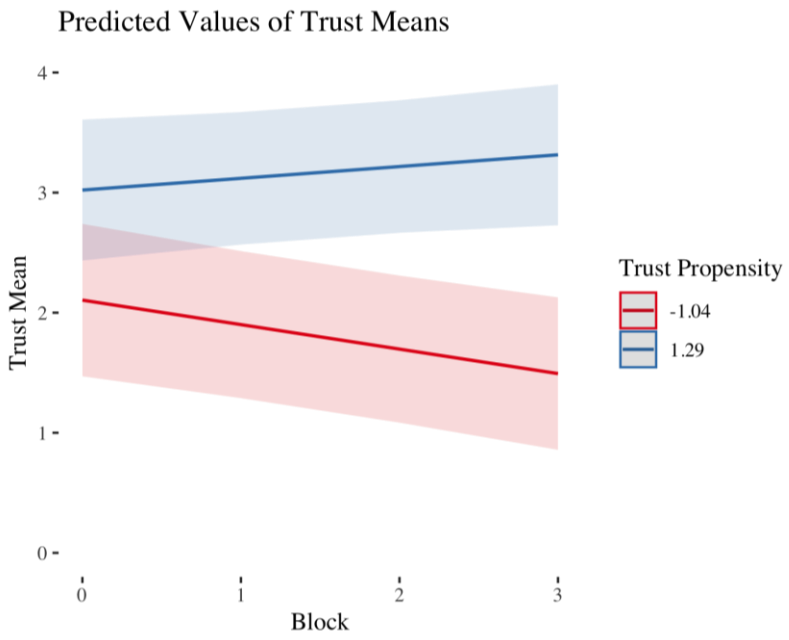


Figure 8. Trust across time by propensity differences in experiment one.

While trust decreased over all, this was conditionally dependent on propensity scores. Those who scored higher in trust propensity reported increasing trust attitudes, an inversion of the overall trend and those who score low on trust propensity.

Reliance

A logit mixed-effects (or multi-level) model was created in R using the '*lme4'* package as in previous models, but this time using the *glmer* function to model dichotomous outcomes (Model D). The same packages used in previous analyses were also used to achieve *p* values and pseudo-R^2. While termed as a *structural equation model*, Merritt (2011) used a similarly multi-leveled approach in their study to investigate multiple reliance outcomes. Another major motivation for this type of model is related to the binomial nature of our outcome variable where the participant either relied on the automation in a trial (1) or did not rely on the automation in the same trial (0). Past papers related to the current research have aggregated binomial reliance outcomes into proportions that are treated as a single continuous outcome per participant

(Satterfield et al., 2017). We chose to use mixed-effects models instead for a few reasons outlined by Jaeger (2008). Unlike ANOVAs, mixed models do not require homogeneity of variance; our reliance models were in significant violation based on visual inspection of diagnostic plots (appendix item 10). A second major advantage for our use of mixed models over ANOVAs is appropriate power; due to sample size limitations related to the expense of manipulations, it was critical to maximize statistical power. In complement to that, ANOVAs with aggregated binomial data can produce spurious significance. The third major advantage of mixed models over ANOVAs was the possibility of modeling continuous predictors: trust measured as a continuous scale was a critical theoretical predictor for the present research questions.

Our model included trust measurements predicting reliance trials across analytical risk conditions, along with the trust by analytical interaction (appendix item 11). Affective risk was included without an interaction term due to the manipulation check failure. This resulted in a three-level model (see figure 9) where reliance (level 1) was nested in trust measurements (level 2) and all blocks were crossed between experimental manipulation groupings (level 3).

$$\begin{aligned}
&\text{Level 1: Reliance}_{ibt} && = \square_{0ibt} + e_{ibt} \\
&\text{Level 2: } \square_{0bt} && = \beta_{00bt} + \beta_{10bt}(\text{Trust}) \\
&\text{Level 3: } \beta_{00t} && = \gamma_{000t} + \gamma_{001}(\text{Analytical_high}) + \gamma_{02}(\text{Affective_high}) + u_{00t} \\
&\qquad \beta_{01t} && = \gamma_{010t}(\text{Analytical_high})
\end{aligned}$$

Figure 9. Model formula for reliance generalized mixed effects model

Initially, our reliance model was fit with an explicit random effects structure of each block at level 2 (Model D, appendix items 12 and 13). The results of this model indicated extremely low variance attributed to this level two variance with an intraclass correlation coefficient (ICC) .05, compared to .49 of level one and .54 of level three. Therefore, the fit of this model was further investigated using a Hosmer-Lemeshow goodness-of-fit test for binary

outcomes via the '*sjstats*' package (Hosmer & Lemeshow, 2000; Lüdecke, 2019). This indicated a poor overall model fit: with the null hypothesis being that the model was well-fit, we rejected that null hypothesis (p = .007). Because of the poor fit, we considered formulating the model without explicit specification of level two variance. While no hard rule exists for how small an ICC measurement must be to not explicitly model a level of random effects, existing research based on data simulations of general linear models has shown at if an ICC is less than .1, it is suitable to ignore the nesting level (Chen et al., 2012). Our level two ICC of .05 is moderately lower than Chen and colleagues' recommended cutoff value. Therefore, we fit the same reliance model (Model E) without explicit specification of level two variance. Model E was shown to be a better fitted model, where the Hosmer-Lemeshow test failed to reject the null hypothesis that the model was well-fitting (p = .20). Considering the two model fits, our results and hypothesis tests were based on the better fit model (Model E, table 5).

Table 5.

Experiment one reliance Model E results with values in Odds-Ratios (Confidence intervals).

Fixed Effects	Odds-Ratio	*Confidence intervals*
Reference Intercept, γ_{000}	.05***	.01-.14
Trust, γ_{010}	**1.34***	**1.00-1.79**
High analytical, γ_{001}	.47	.11-1.94
High affective, γ_{002}	1.16	.45-2.97
High analytical*Trust, γ_{011}	**1.60***	**1.04-2.45**
Random Effects	Value	
Between-person variance (τ_{00})	3.29	
Within-person variance (σ^2)	3.35	

*p<.05, **p<.01, ***p<.001

There was no significant difference in reliance behaviors comparing the low analytical risk group to the high analytical risk group (γ_{001} = .47, z = -1.05, p = .30). We failed to reject the null hypothesis of H1a; there was no significant difference in reliance due to analytical risk

alone. While affective manipulations were included in the model, these were just for control purposes as the manipulation itself failed. We failed to reject the null hypotheses of H1b and H4 because the variable could not be investigated: we found no evidence for more reliance in high affective conditions or more prediction of trust on reliance at high levels of affective risk.

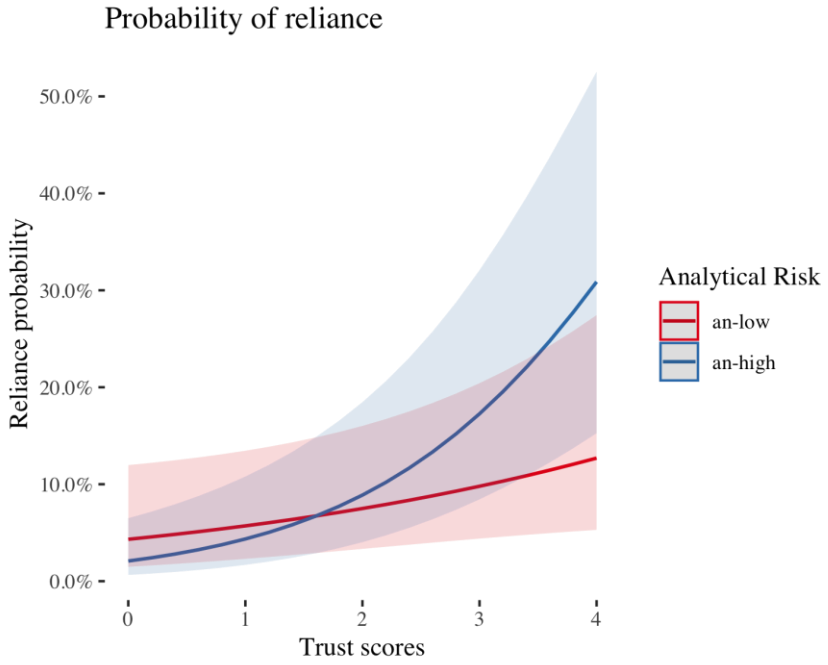


Figure 10. Reliance as predicted by trust across analytical risk groups in experiment one.

The results indicated a significant effect of trust on reliance, such that higher trust increased the odds of reliance behaviors (γ_{010} = 1.34, z =1.99, p < .05). Further considering the interaction of trust and analytical risk (seen in figure 10), a significant interaction was found such that participants within the high analytical risk group relied more on the automation when trust scores were high compared to those in the lower analytical risk group whose trust scores were high (γ_{011} = 1.60, z =2.15, p < .05).

Behavioral Trust

Building from the work by Satterfield and colleagues (2017), we conducted another logit mixed-effects model (figure 9) using the same packages as the reliance model to investigate the relationship of subjective trust (ST) to behavioral trust (BT). ST was measured by a questionnaire at each block (Merritt, 2011) and BT was measured by if a participant engaged with the 'peeking' behavior at each trial (before choosing to rely or not rely).

$$\text{Level 1: } BT_{ibt} = \square_{0ibt} + e_{ibt}$$
$$\text{Level 2: } \square_{0bt} = \beta_{00bt} + \beta_{10bt}(ST) + r_{0bt}$$
$$\text{Level 3: } \beta_{00t} = \gamma_{000t} + \gamma_{001}(Analytical\_high) + \gamma_{002}(Affective\_high) + u_{00t}$$
$$\beta_{10t} = \gamma_{010t}$$

Figure 11. Model formula for BT generalized mixed effects model

Our BT model (Model F, figure 11) was fit with three levels (trial, block, individual treatment) with respective ICCs of .68, .02, and .20. While the level 2 ICC was extremely low, a Hosmer-Lemeshow goodness-of-fit test indicated the model was well-fit (p=.67), so the full level structure was used in the model.

Table 6.

BT Model F results with values in Odds-Ratios (Confidence intervals)

Fixed Effects	Odds Ratio	Confidence Interval
Reference Intercept, γ_{000}	.00***	.00-.01
Trust, γ_{010}	1.43	.97-2.11
High analytical, γ_{001}	2.68	.53-13.53
High affective, γ_{002}	2.24	.44-11.39
Random Effects	**Value**	
Between-person variance (τ_{00})	3.29	
Within-person block variance (e_{00})	.26	
Within-person variance (σ^2)	7.42	

*p<.05, **p<.01, ***p<.001

Results, as seen in table 6, indicated that BT was not significantly predicted by ST (γ_{010} = 1.43, z = 1.81, p = .07), albeit marginally. Further, there was no significant effect of analytical risk manipulations on BT (γ_{001} = 2.68, z = 1.19, p = .23). These results would indicate the BT and ST are somewhat distinct.

To more deeply investigate how BT and ST are related, we conducted an exploratory model with BT and ST predicting reliance (Model G can be found in appendix item 13). This model was not well-fit based on statistical analysis (Hosmer-Lemeshow with p = .03). Additionally, the odds-ratio values were extreme, (e.g., γ = 23141.51 with confidence intervals from 836.69-640053.35). This was indicative of quasi-complete separation (Albert & Anderson, 1984). Quasi-complete separation can be seen in our data by crosstabulation descriptives (table 7): some cells of the binary predictor BT were extremely small, specifically when participants peaked (BT=1) and did not rely (reliance=0); only 0.02% of cases were in this category.

Table 7.

Crosstabulation of BT and reliance, Count / Overall proportion

	No BT (0)	BT (1)
No Reliance (0)	3086 .791	9 .002
Reliance (1)	630 .162	175 .047

This quasi-complete separation was problematic for meaningful interpretation due to the extreme odds-ratio values. While the data were theoretically binomial based on the experimental design (fixed trial amount with binary outcomes), the extremely low count in the proportions has been shown to be better fit by approximations of the binomial with Poisson distributions (Berenson, Levine, Szabat, & Krehbiel, 2012). Therefore, Model H (figure 12) was conducted

with the same formula as Model G, but with a Poisson regression function to better account for the small crosstabulation cell counts in the model.

$$
\begin{aligned}
\text{Level 1: Reliance}_{ibt} &= \square_{0ibt} + \square_{1ibt}(BT) + e_{ibt} \\
\text{Level 2: } \square_{0bt} &= \beta_{00bt} + \beta_{01bt}(ST) + r_{0bt} \\
\square_{1bt} &= \beta_{10bt} + \beta_{11bt}(ST) \\
\text{Level 3: } \beta_{00t} &= \gamma_{000t} + u_{00t} \\
\beta_{01t} &= \gamma_{010t} \\
\beta_{10t} &= \gamma_{100t} \\
\beta_{11t} &= \gamma_{110t}
\end{aligned}
$$

Figure 12. Poisson approximation of binomial Model H

While there are some concerns over model specification issues with Poisson approximations of binomial distributions, those are relegated to models with concerning levels of overdispersion and small trial amounts (Ferrari & Comelli, 2016). Model H showed a low and non-significant amount overdispersion, where the dispersion measured at .79 via the '*AER'* package (Cameron & Trivedi, 1990; Kleiber & Achim, 2008). Further, our data featured 60 trials per participant, far larger than Ferrari and Comelli's largest trial size of 24. With this evidence, our model fit with a Poisson regression was not problematic based on existing findings. The ICCs of level one, two, and three were, respectively, 41%, <1%, and 57%. Despite the extremely low variance in level two, we maintained the three-level variance structure because the model fit well based on dispersion analyses (table 8).

Table 8.

Experiment one reliance Model H results with values in Odds-Ratios (Confidence intervals)

Fixed Effects	Odds-Ratio	Confidence Interval
Reference Intercept, γ_{000}	.03***	.02-.06
BT, γ_{010}	26.50***	12.41-56.61
ST, γ_{100}	1.40***	1.17-1.68
BT*ST, γ_{110}	.71*	.54-.94
Random Effects	**Value**	
Between-person variance (τ_{00})	1.71	
Within-person block variance (e_{00})	.01	
Within-person variance (σ^2)	2.39	

*p<.05, **p<.01, ***p<.001

Individually, both BT (γ_{010} = 26.50, z = 8.46, p < .001) and ST (γ_{010} = 1.40, z = 3.66, p < .001) were highly predictive of reliance. Our odds-ratio for BT was still high and therefore was interpreted cautiously. However, considering descriptive views of our data, it did appear that it was incredibly rare for participants to engage in peeking behavior and then ignore the automated recommendation. There was a significant interaction effect between the two predictors (γ_{010} = .71, z = -2.39, p < .05), such that BT behaviors were less predictive of reliance trust was high, as seen in figure 13.

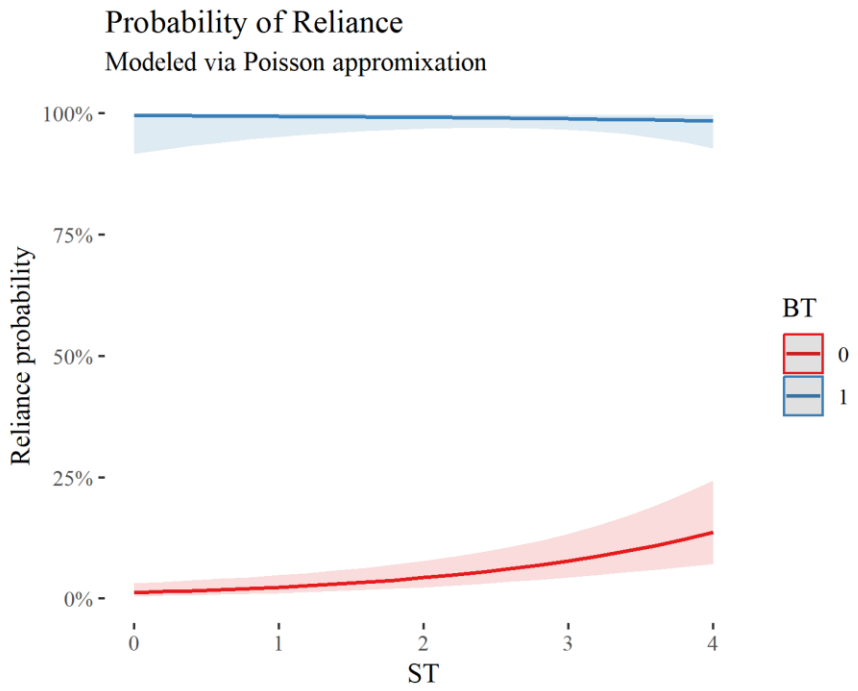


Figure 13. BT and ST predicting reliance in experiment one.

Overall, BT was a rare behavior observed in the experiment. Still, the results indicated important interactions with ST and how both predictors related to reliance.

Discussion

Subjective Trust

Our planned hypotheses around affective risk were not able to be investigated due to manipulation check results; these effects and hypotheses were revisited in experiment two. Therefore, hypothesis H3a was not investigated, regarding how affective risk influenced trust. We found support for hypothesis H3b; trust did not differ across analytical risk levels. This is in line with what past research has found (Satterfield et al., 2017). Other research on risk and automation use has either not measured trust (Ezer et al., 2008) or not explicitly compared trust across risk levels (Lyons & Stokes, 2012; Perkins et al., 2010).

Reliance

Our results for reliance related to risk main effects were similar to trust. Hypotheses H1b and H2 could not be investigated due our manipulation check. We also found no differences in reliance across analytical risk manipulations in our main model (Model E). This was similar to findings by Perkins and colleagues (2010), where no general effect of risk conditions on reliance was found. However, they did find increased reliance within conditions when uncommon hazards were encountered. Other studies have found significant differences on reliance due to risk, in contrast to our findings. Ezer and colleagues (2008) found that as the cost of errors (risk) increased, there was significantly less reliance on the automated tool. Other studies are harder to compare directly when considering their constructs of reliance and measurement.

Lyons and Stokes (2012) found significantly less reliance on a human adviser, indicating more reliance on the opposing automated tool; reliance on the automation was not measured directly. Satterfield and colleagues (2017) measured a behavior that could be considered 'behavioral trust' but could also be seen as a reliance behavior and be compared directly; we fully explored both options. If we first compare their measured behavior directly as the construct of reliance, their results indicated reliance was lower in situations of high risk. These studies provide some evidence for lower reliance at higher risk, but suffer from indirect constructs of reliance, unlike the present study.

Another concern that our study was able to overcome is dealing with discrete data and aggregation (through generalized mixed effects models). Satterfield and colleagues (2017) analyzed their outcome data with a t-test on aggregated proportions of count data to find less reliance in high risk situations. Jaeger (2008) has shown how aggregation of non-continuous data can produce spurious correlations; indeed we also found significant effects (albeit in the opposite

direction of effect) by aggregating our data reliance outcomes between groups in a non-parametric z-test. Therefore, Satterfield and colleagues' results should be interpreted with some caution. Their manipulation itself did involve a different sum ($50 compared to our $10, in high risk conditions), so that aspect of the manipulation could also contribute to the differences between our findings.

Subjective Trust and Reliance

A major motivation for measuring trust is in its ability to predict reliance. Again, because of the manipulation check failure, we could not investigate hypothesis H4 related to trust's prediction of reliance when affective risk was high. We did find when analytical risk was high, participants relied on the automation significantly more when their trust was also high, compared to low-trust participants. This is in line with past findings that gave evidence for a stronger correlation of trust factors and reliance in high risk conditions (Perkins et al, 2010). Other studies either did not measure trust (Ezer et al., 2008), or did not specifically compare trust and reliance (Satterfield et al., 2017).

Our results also indicated that trust declined slightly over time, depending on the level of trust propensity. In this way, it appears that the stable propensity attitudes participants already had impacted new attitudes, at least when the reliability of an automated tool was constant between groups. This fits with previous findings that high trust propensity only led to lower immediate trust ratings when reliability was varied to be low (Merritt & Ilgen, 2008).

Behavioral Trust, Subjective Trust, and Reliance

Behavioral trust is an operator action related to use of an automated tool that precedes ultimate reliance on the output of the automated tool (Satterfield et al, 2017). We have investigated the behavioral results of Satterfield and colleagues (2017) through the lens of

reliance; this section considers their findings against ours through the lens of ‘behavioral trust’. We believe this dual consideration is warranted as they did not explicitly compare behavioral trust with reliance in their paper. Our experiment was designed to clearly delineate the two constructs. Their delineation of subjective and behavioral trust was supported by our findings that behavioral trust (BT) did not significantly predict subjective trust (ST). This also suggests that Satterfield correctly established their measure as BT instead of reliance, as they also found no relationship of BT to ST.

In terms of BT’s relationship to reliance, participants in our study with low subjective trust almost always peaked (low behavioral trust) when they relied and almost never relied if they were not peeking; participants with high trust almost always relied after peeking but would sometimes rely without peeking. BT may be similar to reliance, but not a surrogate for trust, given the interaction effect. When trust was low, BT essentially occurred along reliance every time; this was a necessary step to reliance in the unique strategy of low-trust participants. High-trust individuals were willing to use the automation without peeking. In this sense, BT is extremely close as a construct to measuring reliance behaviors, yet still worthy of distinction because of its dependence upon ST levels of individuals. BT could be useful when reliance measurement is not feasible due to the automated tool’s design, but ST must also be considered if reliance cannot be measured as BT’s relationship to reliance depends on ST.

Affective manipulation changes from experiment one to experiment two

Throughout the results and discussion of experiment one, some hypotheses were unable to be addressed because of the failed manipulation check. Despite expectations based on previous research findings (Kurdi et al., 2017), the affective manipulation could not be considered effective because the high affective group did not show higher levels of negative

affect than the low affective group. For this reason, we decided to alter experiment two. The first change was to maintain the separation of manipulations between affective and analytical risk, rather than collapse them into high affective/analytical and low affective/analytical, so that we could still investigate the subcomponents of risk. This change necessitated double the sample size for experiment two from 60 participants in a 2 (collapsed risk type) x 2 (automation type) design to at least 120 participants in a 2 (affective risk type) x 2 (analytical risk type) x 2 (automation type).

The second change included differences in the affective stimuli and their presentation. Because of the initially high level in negative affect immediately following the fearful video, a second video was included in experiment 2 between x-ray blocks 2 and 3 (the halfway point) in both the high and low affective risk groups.

Third, the images that were used in each trial were changed to those of a different picture set, the GAPED image database (Dan-Glauser & Scherer, 2011). This database was more clearly delineated for the present purpose of fear-induction. Compared to OASIS (Kurdi et al., 2017) that had four categories within each all having a range of negative to neutral valences, GAPED had categories specifically for negative emotions (e.g., spiders, human suffering, snakes, etc.) and for neutral emotions. This allowed for clearer delineation between affective stimuli based on their qualitative categorization. More image stimuli were used (60 per affective condition) so that none were repeated within any trials. This may be critical due to the rapid habituation of negative stimuli (Bradley, Lang, & Cuthbert, 1993; Wright et al., 2001).

Lastly, the placement of the negative stimuli was also changed within the trials. In experiment one, images appeared consistently and reliably at every decision point; in experiment two, images appeared only when the participant failed to properly classify an x-ray image. This

was intended to increase the impact of negative stimuli by making images less expected and more task-relevant. Herwig & colleagues (2007) showed evidence that expectation of negative stimuli decreases activation in emotional processing brain regions, compared to when there is no expectation possible. Wilson & colleagues (2016) found that negative images' impact on behavior was more dependent on the relevance of the images than valence alone (although mere valence did have singular effects).

All changes to experiment two were intended to increase negative affective ratings in the high affective groups compared to low effective groups. The present study paradigm was novel (negative affective manipulations and automation reliance behaviors), so design changes did not have direct empirical evidence. Some past studies similarly showed failed manipulation checks for negative mood (not affect) induction with the PANAS (Stokes et al., 2010). However, the theoretical background literature indicated that image presentations in the modified experiment two design should have increased the impact of negative visual stimuli on subjective ratings and behaviors because the images (a) could not be habituated, (b) were not presented in an explicitly expected timing, and (c) were task-relevant. Further, considering the near floor effects of negative affect ratings, these changes were not expected to have any serious detrimental effect on the results over and above the implementation of image stimuli in experiment one.

Experiment Two

Design

This experiment pseudo-randomly assigned participants in a 2x2x2 design (affective risk by analytical risk by stage of automation in tool) via blocked randomization. Affective risk was manipulated with different images and videos placed at different times, compared to experiment one. Analytical risk manipulations were identical to experiment one. Early automation involved

an information analysis stage of automation in the automated tool (stage two/three), and late automation involved an action implementation tool (stage four).

Participants

Experiment two used the same exact sampling methods from experiment one to recruit participants from Mechanical Turk, with 148 participants. Participants' base compensation was $2.00.

Materials

The affective videos (appendix item 2) were kept the same as experiment one, but each was shown once (one at the beginning and the other at the middle of the trials) instead of one being shown alone at the beginning (Schaefer et al., 2010). Images for emotional elicitation were used from GAPED (Glauser & Scherer, 2011). Sixty images of the highest arousal and most negative valence were used from fearful image categories in the high affective risk condition (examples in appendix item 14). Sixty images nearest to 0 in valence and arousal from the neutral image category were used for the low affective risk condition.

The x-ray slide images were the same as in experiment one but were shown in a bisected format: .85 seconds for the first viewing and .75 seconds for the second viewing.

Automation

Because the automation stage was manipulated (between stage two/three and stage four), it differed from experiment one. Both automation types featured a reliability level of approximately 75% accuracy, serving as moderate reliability that is halfway between the high and low conditions used by Merritt and Ilgen (2008). Consistent with experiment one, correct responses were worth 10 points, and incorrect gave 0 points. Using automation exclusively

would have resulted in a suboptimal score; participants were told this information. Unlike experiment one, no option to *peek* was used in this experiment.

Automation A (appendix item 15) helped to analyze information (stage two) and had an implicit decision recommendation (stage three), though it mainly existed within stage two. This was in the form of a visual marker for high risk areas in a baggage slide (via a bright rectangle outline of the high likelihood area) or no visual marker at all if there were no high risk areas in a slide. The implicit recommendation to 'check' was the presence of the marker, and the implicit recommendation to 'clear' was the lack of the marker. Automation B (appendix item 16) completed the trial automatically (stage four). It did not show its recommendation for override, it simply made the decision once the participant decided to rely on it by automatically choosing to check or clear the bag.

Measures

This experiment generally used the same materials as experiment one but did not include behavioral trust measures. This was because of practical differences of measuring behavioral trust. While higher levels of automation have explicit behavioral checking in stage three (e.g.; seeing a recommendation explicitly or not), low levels of automation may involve more cognitive strategies for checking (e.g.; mentally assessing accuracy of analyzed information in a stage two automated tool). Because measurement of these cognitive checking strategies would have been subjective in nature (through a questionnaire), there would not be an inherent advantage over traditional trust scales. This experiment utilized stage two/three and four automation types to fully delineate automation with no overlap in automation stages between conditions. For this reason, only behavioral reliance and subjective trust (Merritt, 2011) were measured in experiment two.

For Automation A, viewing of the trial slide was bisected by the option to enable the tool (for a 10% cost in point potential if the final decision is correct) or look again without automation; when activated, the second duration included the visual components of the automation. Reliance was measured when a user engaged the automation, an operationalization that closely resembles how reliance was measured in earlier work in x-ray microworld tasks (Merritt, 2011; Merritt & Ilgen, 2008). For Automation B, viewing of the slides was bisected for consistency across conditions; participants could look again for no cost, or use the automation for a 10% cost if the automation is correct. Reliance here was measured as when a participant used the automated decision option.

Subjective trust used the same scale as experiment one (Merritt, 2011), given at the same point in time before a trial block, and after the final trial block. This was used as a predictor variable for reliance behaviors. Trust propensity was again measured once by adapting a scale from Singh, Molloy, & Parasuraman (1993) in an identical fashion as experiment one.

Procedure

As in experiment one, participants followed the same procedure with MTurk, consent, practice, microworld task, trust measurement, and manipulation check. The differences included that there was no ‘checking’ point penalty, as there was no option to peek in either automation condition. Automation instructions and use were different, depending on the automation stage condition each participant was randomly assigned to. Finally, the affective images used a different database and all affective stimuli were placed differently: videos were now viewed before and after the second trial block; images only appeared when participants made an incorrect decision after an x-ray trial.

Results

Data were collected over a period of five days. One hundred forty-eight participants completed the experiment (table 9). We oversampled with the expectation of removing some participants who demonstrated low accuracy. Six participants fit this performance category (less than or equal to 50% accurate) and were removed; two other participants were removed for contamination (awareness of alternative manipulations). Our final sample size was 140 participants with 76 being female and 64 being male. The average age was 39 and ranged from 20-72. While PANAS and trust measures were collected after the final x-ray block, these were not used in analysis as they did not predict any subsequent x-ray trial reliance outcomes. This means that only the first four of the five total subjective response data points were used per participant.

Table 9.

Descriptives by between-subjects conditions.

Condition	Gender (Female / Male)	Mean Age
Automation A / High Analytical / High Affective	10 / 6	41
Automation A / High Analytical / Low Affective	7 / 11	45
Automation A / Low Analytical / High Affective	11 / 8	39
Automation A / Low Analytical / Low Affective	9 / 9	38
Automation B / High Analytical / High Affective	10 / 9	38
Automation B / High Analytical / Low Affective	12 / 5	42
Automation B / Low Analytical / High Affective	9 / 8	36
Automation B / Low Analytical / Low Affective	8 / 8	36

Manipulation Check

The manipulation check analyses were conducted identically as in experiment one. PANAS sub-questions that referred to negative emotions were averaged per participant at each block. A random intercepts model was conducted using the 'lme4' package in R to assess if the emotional elicitation method was effective, as seen in figure 14 (Bates et al., 2014; Raudenbush & Bryk, 2002). *P* values to assess statistical significance were given via the 'lmerTest' package that is built upon the 'lme4' package (Kuznetsova, Brockhoff, & Christensen, 2017); R^2 values were obtained using the 'sjstats' package that is also built upon 'lme4' (Lüdecke D, 2019). The dependent variable was the continuous measurement of negative PANAS average scores by block and the predictor variable the categorical variable of affective manipulation (descriptives in appendix item 17 and diagnostics in appendix item 18).

Level 1: $\text{Negative\_Affect}_{it} = \beta_{0it} + r_{it}$
Level 2: $\beta_{0i} = \gamma_{00} + \gamma_{01}(\text{Affective}) + u_i$

Figure 14. Model I formulation for the manipulation check.

Assessing Model I detailed in table 10, the results indicated that the manipulation was effective in increasing negative affect for those in the high affective risk group compared to the low affective risk group. The high affective risk group reported negative affect scores significantly higher than zero ($\gamma_{00} = 1.07$, $t = 10.172$, $p < .001$). The low affective group reported significantly less negative affect than the high affective group ($\gamma_{01} = -.54$, $t = -3.64$, $p < .001$), as is generally seen in figure 14. The marginal effect size, considering fixed effects only, was small with $R^2 = .07$.

Table 10.

Experiment two Manipulation check results for Model I.

Fixed Effects	Intercept	Std. Err.
High affect intercept, γ_{00}	1.07***	0.10
Low affective Intercept, γ_{01}	-.54***	0.15
Random Effects	**Value**	
Between-person variance (τ_{00})	.73	
Within-person variance (σ^2)	.19	

*p<.05, **p<.01, ***p<.001

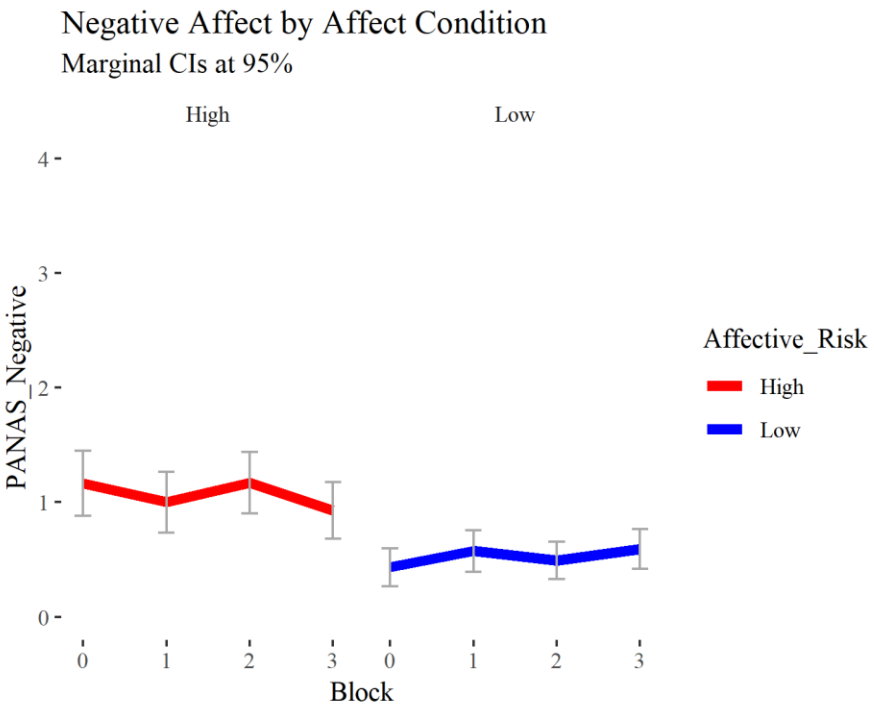


Figure 15. Experiment two negative PANAS questions across blocks by condition.

The changes to the experimental protocol from experiment one to experiment two appeared to be effective in manipulating the affect of participants across conditions. While the effect size was small, the difference was significant, so affective manipulations were included in all relevant subsequent analyses.

Behavioral Metrics

Descriptive metrics were analyzed to discern the difficulty of the x-ray task and the general behavioral trends using the same outcomes as in experiment one (except BT/peeking behavior). The table below (table 11) summarizes between-group metrics. Overall metrics are discussed subsequently.

Table 11.

Behavioral descriptive metrics of experiment two between groups.

	Mean	SD	Range	Mean	SD	Range
	Auto. A / High Ana. / High Aff.			Auto. A / High Ana. / Low Aff.		
Accuracy	80%	9%	60% - 90%	79%	9%	52% - 90%
Score	475	57	346 - 538	468	55	289 - 540
Reliance	15%	16%	0% - 45%	20%	22%	0% - 70%
	Auto. A / Low Ana. / High Aff.			Auto. A / Low Ana. / Low Aff.		
Accuracy	77%	20%	52% - 90%	76%	9%	52% - 87%
Score	443	57	292 - 533	477	57	302 - 518
Reliance	37%	28%	2% - 85%	16%	21%	0% - 77%
	Auto. B / High Ana. / High Aff.			Auto. B / High Ana. / Low Aff.		
Accuracy	80%	8%	53% - 88%	78%	9%	60% - 90%
Score	478	47	309 - 525	460	54	356 - 539
Reliance	14%	13%	0% - 43%	13%	14%	0% - 60%
	Auto. B / Low Ana. / High Aff.			Auto. B / Low Ana. / Low Aff.		
Accuracy	79%	6%	7% - 92%	75%	8%	56% - 83%
Score	461	42	397 - 550	439	48	339 - 498
Reliance	23%	24%	0% - 75%	27%	26%	0% - 83%

On average, participants were 78% accurate (SD = 9%, range = 52% - 92%). Participants were given feedback of their ongoing total score during the experiment (out of 600 points), and

only thirty-two participants (23%) achieved the 500-point cutoff. The average score (the distribution seen in figure 15) was 459 (SD = 53, range = 289 - 550).

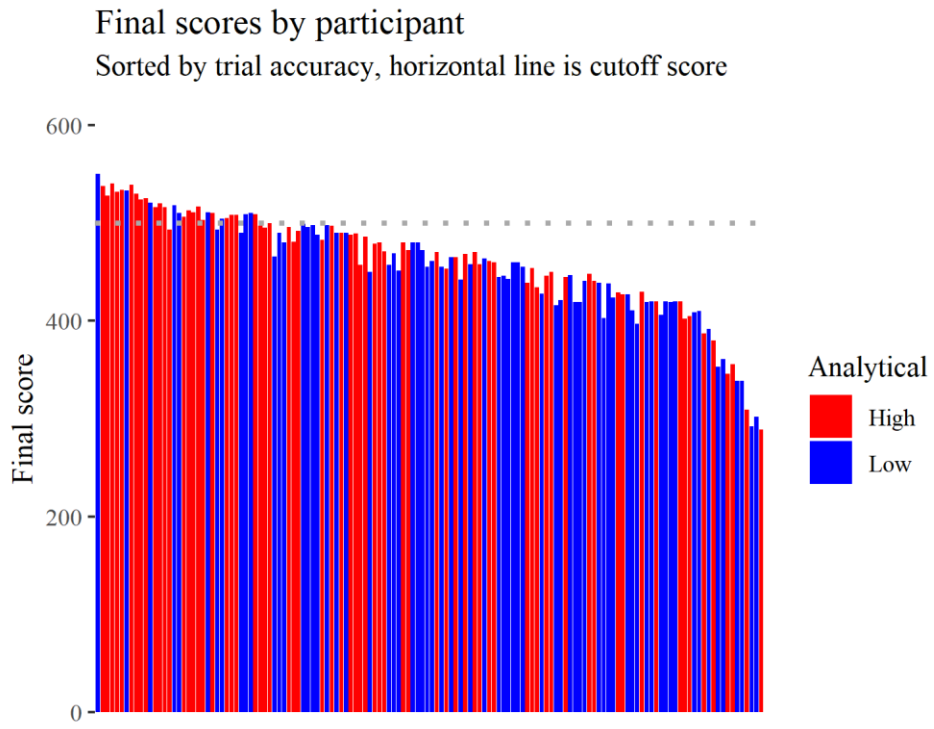


Figure 16. Final scores by participant, with bars sorted by accuracy proportions.

Reliance was measured across sixty x-ray slide trials per participant. On average, participants relied on the automation 21% of the time (SD = 22%, range = 0% - 85). There was a large amount of variance between participants, as seen in figure 16.

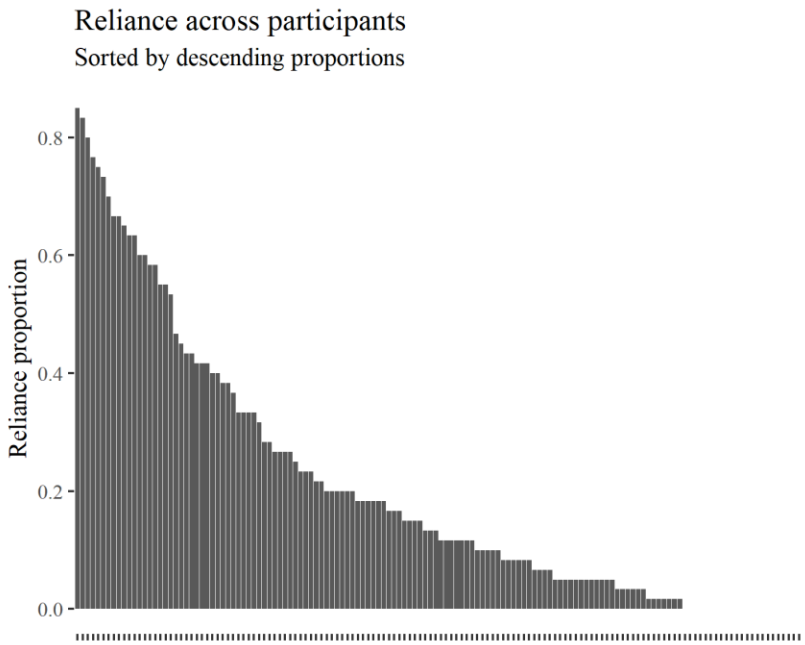


Figure 17. Reliance proportions across participants in experiment two

Like experiment one, behavioral metrics indicated that individual strategies of reliance on the automation largely varied. Metrics of accuracy and score also indicated the task was sufficiently difficult, because the cutoff score was only reached in 23% of participant cases. This proportion was slightly higher than experiment one results.

Trust

Trust questionnaires (Merritt, 2011) were averaged per participant at each block. A random intercepts model was conducted in R, as formalized in figure 17 (Model J), using the same packages as Models B and C in experiment one. The continuous dependent variable was average trust scores at each block per participant. The categorical independent variables were risk manipulations (analytical and affective) and automation type. The continuous predictor variable was block, or each time point of trust measurement before subsequent x-ray trials. This resulted in a two-level model with trust measurements across time (level 1) and experimental manipulation groupings (level 2). Descriptives are shown in appendix item 19 and diagnostics in

appendix item 20. The ICC indicated a moderate partitioning between levels with 65% of the variance related to individual variation in trust scores (level 1) and 35% of the variance between participants (level 2).

$$\text{Level 1:} \quad Trust_{it} = \beta_{0it} + \beta_{1it}(Block) + r_{it}$$

$$\text{Level 2:} \quad \beta_{0i} = \gamma_{00} + \gamma_{01}(Analytical\_high) + \gamma_{02}(Affective\_high) + \gamma_{03}(Automation\ type) + \gamma_{04}(Analytical\_high*Affective\_high) + u_i$$

$$\beta_{1i} = \gamma_{10}(Block) + \gamma_{12}(Analytical\_high) + \gamma_{13}(Affective\_high) + \gamma_{14}(Automation\ type) + \gamma_{15}(Risk\ interaction\ type)$$

Figure 18. Model J formula for trust mixed effects model.

Results, as detailed in table 12 and visualized in figure 18, indicated that trust did not differ significantly due to analytical risk manipulations alone (γ_{01} = .17, t = .82, p = .415), affective risk manipulations alone (γ_{02} = .23, t = 1.28, p = .20), or their interaction (γ_{04} = -.27, t = -.90, p = .37). Automation type did not have an effect on trust (γ_{03} = .08, t = .55, p = .59). There was a significant effect of time on trust, such that trust decreased slightly over time (γ_{10} = -.14, t = -2.91, p < .01). Time also interacted with automation type on trust, such that trust did not show its main effect of decreasing when participants interacted with automation type b (γ_{12} = .09, t = 2.17, p < .05). Time also interacted with affective risk on trust, such that trust decreased more steeply when participants had a high negative affect (γ_{13} = -.13, t = -2.15, p < .05). The marginal effect size was small with R^2 = .06.

Table 12.

Trust model results in experiment two with Intercepts (Standard Errors)

Fixed Effects	Intercept	Std. Err.	Intercept	Std. Err.
	Model J		Model K	
Intercept, γ_{00}	**2.23*** **	**.17**	**2.23*** **	**.16**
Block, γ_{10}	**-.17****	**0.02**	**-.14****	**.05**
Block by Analytical, γ_{11}	-.08	.06	-.07	.06
Block by Affective, γ_{12}	**-.13***	**.06**	**-.12***	**.06**
Block by Automation, γ_{13}	**.09***	**0.04**	**.09***	**.04**
Block by Risk interaction, γ_{14}	.10	.08	.10	.08
Block by Propensity, γ_{15}			-.08	.05
Block by Automation by Propensity, γ_{16}			.13	.08
High analytical, γ_{01}	.17	0.12	.19	.21
High affective, γ_{02}	.27	0.21	.27	.21
Automation, γ_{03}	.08	0.15	.08	.15
Risk interaction, γ_{04}	-.27	.30	-.29	.29
Trust propensity, γ_{05}			-.14	.19
Random Effects	Value		Value	
Between-person variance (τ_{00})	.55		.55	
Within-person variance (σ^2)	.30		.30	

*p<.05, **p<.01, ***p<.001

Revisiting hypotheses from experiment one that could not be investigated, we did not find strong support for H3a, there was no difference in trust due to affective risk changes; conversely, trust decreased over time when high affective risk was present. We again found support for H3b, there was no significant change in trust due to analytical risk manipulations, even when considering time's interaction. In experiment two hypotheses, we did not find support for H6a, trust was not significantly different across high levels of risk types compared to low levels of risk types. Nor did we find strong support for H6b, trust was not significantly higher in later stage automation conditions. However, there was partial support when considering that

automation B (late stage) remained higher than automation A (early) stage when controlling for time.

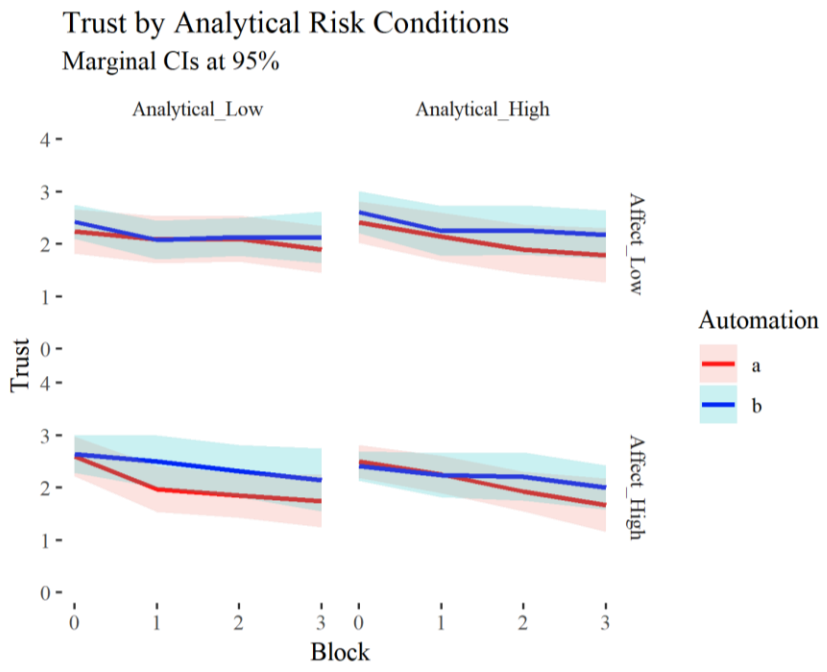


Figure 19. Experiment two trust across conditions and time in experiment two.

In addition to the planned analyses, an exploratory model (Model K, appendix item 21) to compare with experiment one (Model C) was created by adding an additional predictor of trust propensity scores. This variable was centered by overall score mean across participants. This model was higher in its fixed-effects effect size with marginal R^2 = .07. No changes to significance results or beta directions were present in previously existing variables. Propensity on its own was not significant (γ_{04} = -.12, t =-.73, p = .47), nor when interacting with time (γ_{12} = -.07, t =-1.22, p = .22), nor when interacting with time and automation type (γ_{13} = .13, t = 1.59, p = .112).

While trust decreased overall across all conditions, this was partially dependent on two factors. Those who interacted with later stage automation (automation B) reported no change in trust compared to a significant decrease in trust with those who interacted with early stage

automation (automation A). Further, those who had a high negative affect reported a more precipitous decrease in trust compared to those who had less negative affect.

Reliance

A logit mixed-effects (or multi-level) model was created in R using the *glmer* function to model dichotomous outcomes (Model L), involving the same R packages as Model D in experiment one. The dependent variable was reliance behavior of choosing to not rely (0) or to (1) rely on the automated tool at each x-ray trial. The continuous predictor was trust ratings across blocks. The categorical predictors, all of which had two levels, were affective risk, analytical risk, and automation type. In order to answer our related hypotheses, these predictors were all given combined interaction terms. This resulted in a three-level model (figure 18) where reliance (level 1) was nested in trust measurements (level 2) and all blocks were crossed between experimental manipulation groupings (level 3).

$$
\begin{aligned}
\text{Level 1: Reliance}_{ibt} &= \square_{0ibt} + e_{ibt} \\
\text{Level 2: } \square_{0bt} &= \beta_{00bt} + \beta_{01bt}(\text{Trust}) \\
\text{Level 3: } \beta_{00t} &= \gamma_{000t} + \gamma_{001}(\text{Analytical\_high}) + \gamma_{002}(\text{Affective\_high}) + \\
&\quad \gamma_{003}(\text{Automation}) + \gamma_{004}(\text{Analytical\_high*Affective\_high}) + \\
&\quad \gamma_{005}(\text{Analytical\_high*Automation}) + \\
&\quad \gamma_{006}(\text{Affective\_high*Automation}) + \\
&\quad \gamma_{007}(\text{Analytical\_high*Affective\_high*Automation}) \\
\beta_{01t} &= \gamma_{010t} + \gamma_{011}(\text{Analytical\_high}) + \gamma_{012}(\text{Affective\_high}) + \\
&\quad \gamma_{013}(\text{Automation}) + \gamma_{014}(\text{Analytical\_high*Affective\_high}) + \\
&\quad \gamma_{015}(\text{Analytical\_high*Automation}) + \\
&\quad \gamma_{016}(\text{Affective\_high*Automation}) + \\
&\quad \gamma_{017}(\text{Analytical\_high*Affective\_high*Automation})
\end{aligned}
$$

Figure 20. Model M formula for reliance generalized mixed effects model in experiment two.

Initially, our reliance model was fit with an explicit random effects structure of each block at level 2 (Model L, appendix items 22 & 23). The results of this model indicated low amounts of level two variance with an intraclass correlation coefficient (ICC) .09, compared to .40 of level one and .51 of level three. Therefore, the fit of this model was further investigated

using a Hosmer-Lemeshow goodness-of-fit test for binary outcomes (Hosmer & Lemeshow, 2000; Lüdecke, 2019). This indicated a poor overall model fit: with the null hypothesis being that the model was well-fit, we rejected that null hypothesis ($p < .001$, $\square^2 = 52.85$). Because of the poor fit, we considered removing the level two nesting structure based on the ICC being lower than .10 (Chen et al., 2012). The model without level two variance (Model M) did not perform well on goodness-of-fit test either, albeit with a moderately less extreme statistical value ($p < .01$, $\square^2 = 25.83$). This was not especially conclusive on its own to indicate which model to use. Looking at the model outcomes side by side, Model M (figure 19) appeared to have more realistic and interpretable odds-ratios ranges, where Model L had somewhat extreme odds-ratios (for example, 1.29 - 264.35). Because the level two variance in Model M is below the cutoff based on past research and it contains more feasible odds-ratios, this was the model used for interpretation around our hypotheses (table 13, diagnostics in appendix item 24). While there is some concern for the model based on a positive Hosmer-Lemeshow test, the test has been shown to be overly-sensitive to large sample sizes, especially those exceeding 5,000 data points (not the case in experiment one but is the case in experiment two), and does not always mean the model is suspect (Marcin & Romano, 2007).

Table 13.

Reliance Model M results with values in Odds-Ratios (Confidence intervals).

Fixed Effects	Odds-Ratio	Confidence Intervals
Reference Intercept, γ_{000}	0.13**	0.04-0.42
Analytical_high, γ_{001}	0.44	0.09-2.12
Affective_high, γ_{002}	0.8	0.18-3.55
Automation, γ_{003}	0.49	0.09-2.53
Analytical_high*Affective_high, γ_{004}	0.33	0.04-2.87
Analytical_high*Automation, γ_{005}	1.76	0.18-17.10
Affective_high*Automation, γ_{006}	7.76	0.91-66.04
Analytical_high*Affective_high*Automation, γ_{007}	0.24	0.01-5.51
Trust, γ_{010}	0.77	0.51-1.16
Analytical_high*Trust, γ_{011}	**1.86***	**1.10-3.15**
Affective_high*Trust, γ_{012}	**2.56*****	**1.59-4.14**
Automation*Trust, γ_{013}	**2.26****	**1.31-3.89**
Analytical_high*Affective_high*Trust, γ_{014}	0.57	0.29-1.12
Analytical_high*Automation*Trust, γ_{015}	**0.39***	**0.19-0.81**
Affective_high*Automation*Trust, γ_{016}	**0.15*****	**0.08-0.29**
Analytical_high*Affective_high*Automation*Trust, γ_{017}	**6.72*****	**2.59-17.41**
Random Effects	Value	
Between-person variance (τ_{00})	3.29	
Within-person variance (σ^2)	2.71	

*p<.05, **p<.01, ***p<.001

Results did not indicate a significant main effect of trust predicting reliance alone (γ_{010} = .77, z = -1.24, p = .22). There were no main effects of experimental manipulations alone in analytical risk (γ_{001} = .44, z = -1.02, p = .306), affective risk (γ_{002} = .80, z = -.29, p = .77), or automation type (γ_{003} = .49, z = -.86, p = .39); none of the manipulations resulted in significant changes of reliance between conditions. However, many interactions were significant and decomposed visually via marginal plots from '*sjstats*' based on the generalized mixed model output (Lüdecke, 2019).

The amount that trust predicted reliance across each manipulation did have paired interaction effects, such that non-referent risk groups in the model (high analytical risk, high affective risk had a stronger positive relationship between trust and reliance than referent groups (low analytical risk, low affective risk). The high analytical risk group showed high trust scores predicting reliance more strongly compared to the low analytical risk group, (γ_{011} = 1.86, z = 2.30, $p < .05$; see appendix item 25). The high affective risk group showed high trust scores predicting reliance more strongly compared to the low affective risk group, (γ_{012} = 2.56, z = 3.85, $p < .001$; see appendix item 26). Automation stage manipulation showed the opposite such that the referent group (early stage automation) caused a stronger positive trust/reliance relationship than the non-referent group (late stage automation) (γ_{013} = 2.26, z = 2.94, $p < .001$; see appendix item 27).

There was a significant three-way interaction considering how predictive trust is across analytical risk and automation stage (γ_{015} = .39, z = -2.55, $p < .05$). Interpreting this visually (appendix item 28), both automation stages showed a strong prediction of trust to reliance at higher analytical risk, but early automation was higher in both cases (likely related to the obvious effect of the two-way interaction in γ_{012} where automation stage alone indicated early stage automation showed a stronger trust reliance relationship than late stage automation).

There was also a significant three-way interaction considering how predictive trust is across affective risk and automation stage (γ_{016} = .15, z = -5.72, $p < .001$). Interpreting this visually (appendix item 29), early stage automation showed a stronger trust/reliance relationship at high affective risk than late stage automation, but late stage automation showed a stronger trust/reliance relationship at low affective risk than early stage automation. This was echoed by the subsequent four-way interaction.

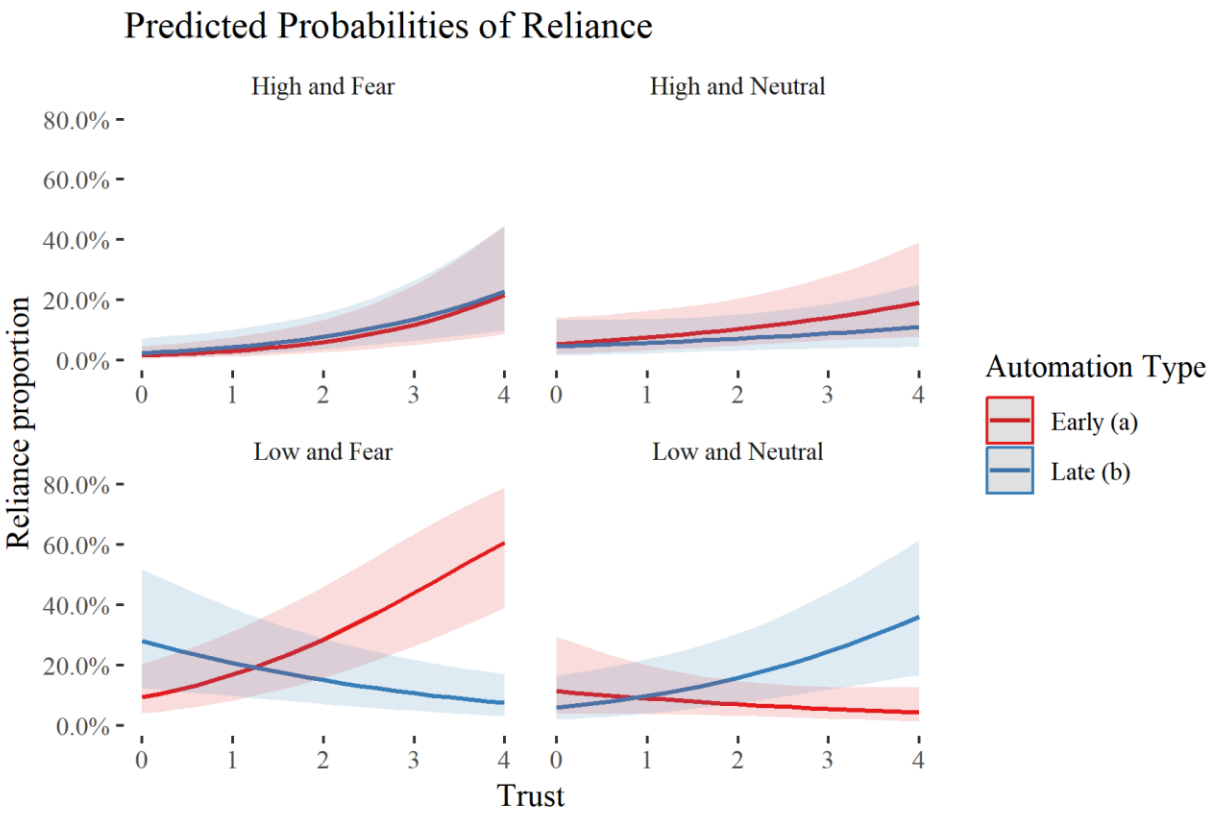


Figure 21. Reliance predicted probabilities in experiment two with conditional CIs

The singular four-way interaction term in the model was significant (γ_{017} = 6.72, z = 3.92, p < .001), as seen in figure 20. This showed that trust similarly predicted reliance across automation stages when both risk types were high. When neither risk type was present, late automation showed a stronger trust/reliance relationship than early stage automation. When only analytical risk was present (upper right quadrant) or only affective risk was present (lower left quadrant), early stage automation showed a stronger trust/reliance relationship than late stage automation.

Discussion

Trust

The effects of risk types and automation stages on trust attitudes alone were small in that main effects were all non-significant. However, trust did vary over the time of the experimental trials, with a slight decrease overall. This general decrease in trust over time was not seen when controlling for automation type: late stage automation did not result in a trust decrease compared to early stage automation, in line with past research (Calhoun, Draper, & Ruff, 2009; Lorenz, Nocera, Rottger, 2002).

There were significant effects of affective risk over time: when participants were exposed to negative affect conditions, their trust in the automation decreased more sharply than neutral affect conditions. While little work has examined the impact of affect on human-automation trust, Perkins and colleagues (2010) found mixed support for more trust in high risk conditions, and their risk condition was postulated (but not measured) to contain an affective risk component. Our findings go against their results, showing decrease in trust over time with high affective risk, in contrast to H3a. One way to account for this difference is that their risk manipulations did not, in fact, include any truly affective components. Our findings do have some theoretical, but not empirical, support: past human-human trust work has shown negative affect to be related to less cooperation and this could be extended to human-automation trust (Lee, 2006). While one empirical study showed similar findings in that negative mood caused less trust in automation, these findings are not necessarily robust considering their failed manipulation check with the PANAS (Stokes et al., 2010).

With our results of affective risk in contrast to those of Perkins and colleagues, it is likely that their risk manipulation was more plainly analytical. Even if we consider their findings

through a purely analytical risk lens, we found contrasting results to theirs: there were no changes in trust due to analytical risk. Our null findings (no trust change with analytical risk) are also supported by more recent experimental evidence (Lyons & Stokes, 2012; Satterfield et al., 2017). Therefore, empirical evidence is growing that trust itself does not change exclusively due to the analytical risk but may change due to affective risk, perhaps because of the primarily emotional nature of trust formation (Lee & See, 2004). This supports H3b. Results from experiment two indicate trust judgements are susceptible to affective risk components, but not analytical risk components, and therefore highlight the importance of controlling for both subcomponents of risk in human-automation research.

In our exploratory analyses, we did not find support for any changes due to trust propensity on changes in immediate trust over time. This is in contrast to our experiment one findings that indicated trust propensity in higher levels would reverse the decline in trust over time. This could be due to the affective component present in experiment two; affective risk may be more influential than general trust propensity. It could also be due to the flexibility of the automation in experiment one, which could vary in stage by choice of the operator but experiment two did not allow participants to vary automation stages.

Reliance

Our experimental results indicated complex interactions across risk subtypes and automation stages. Generally, reliance behaviors did not differ due to manipulations of risk or automation stage (when controlling for trust), showing no support for H5. Past research has shown significant, but conflicting, differences in reliance related to risk. Lyons and Stokes (2012) found less reliance in a human adviser at high risk situations, possibly indicating more reliance on the other information source that was automated. However, this was not directly

measured, which may account for differences from our results. Satterfield and colleagues found significantly less reliance in high risk situations (2017). Again, this could be related to a difference in measurement, as results from our first experiment showed that their behavioral measurement may be better fit by *behavioral trust* than reliance. Further, their method of analysis (aggregating binomial data into a continuous measure) has been shown to increase the chance of type one errors (Jaeger, 2008).

Both risk subtypes at high levels, on their own while controlling for the other, showed significantly stronger predictive power between trust and reliance when compared to low levels at each subtype. This was in line with findings by Perkins and colleagues (2010), whether we consider their manipulation to be analytical or affective. Other studies investigating risk around automation use have not directly compared the predictiveness of trust measurements to reliance outcomes. Apparent complexity in these results arises when we consider that there was no significant interaction effect of trust predicting reliance due to analytical risk by affective risk (general low risk was not different than general high risk), unless automation stage was also included in the interaction effect (general low risk in early stage automation was different than general high risk in late stage automation). These results provide mixed support for H7a.

Analytical risk (controlling for affective risk) showed that early stage automation resulted in a generally stronger trust/reliance prediction across risk conditions, but both automation stages still showed a stronger trust/reliance relationship when high analytical risk was present (γ_{015}). Affective risk (controlling for analytical risk) showed a stronger trust/reliance relationship with early stage automation at high affective risk levels, but low affective risk showed the opposite with late stage automation resulting in a stronger trust/reliance relationship (γ_{016}). The four-way interaction (γ_{017}: analytical risk by affective risk by automation stage by trust) reconfirmed the

three-way interaction results of γ_{016}: high affective risk with low analytical risk caused a stronger trust/reliance relationship with early stage automation compared to late stage. The four-way interaction (γ_{017}) superseded the results of the three-way interaction of analytical risk, automation stage, and trust (γ_{015}). When analytical risk was high and affective risk was low, early stage automation caused a stronger trust/reliance relationship than late stage (similar to high affective risk but low analytical risk, γ_{016}), instead of showing no differences between automation stages.

There is a dearth of empirical results related to risk *and* automation stage predicting trust and reliance, so there is no past literature that can be used to judge these results. A possible reason that the affective risk alone with late stage automation would lower the trust-reliance relationship is that negative mood causes a more critical, narrow information processing approach (Lee, 2004), and the desire to critically consider automation processes and recommendations would be hindered by the lack of transparency related to its late stage nature. Similarly, other lines of research in decision-making indicate that negative affect could motivate a person to seek out diverse and alternative information (Mosier & Fischer, 2010). Therefore, late stage automation may be trusted as much, but still cannot be used because its style of support to the operator is incongruent with the critical thinking style negative affect brings out. Similarly, a higher analytical risk could also result in more logical thinking based on the motivational nature of the manipulation, leading to the same hindrance with late stage automation (albeit to a slightly smaller effect than affective risk). This would be in line with Perkins and colleagues' findings (2010), where their automation at the same stage as our early automation condition showed a positive trust/reliance relationship.

The question remains then as to why this trust/reliance relationship would be positive again with late both types of risk present in late stage automation compared to neither risk

present alongside early stage automation. Numerous theories have postulated that negative affect can change the impact of an identical piece of information given in neutral/positive affect (Loewenstein et al., 2001; Slovic et al., 2007). While neither of these theories are directly about risk subtypes' effects on the trust and reliance relationship, they at the very least indicate how affective and analytical risk are more than the sum of their parts. It is possible that the way these two subtypes of risk interact do not form a narrow, logical cognitive style that requires close investigation of transparent process in an automated tool, even though that cognitive style is generated with each risk subtype individually. Ultimately, these results give mixed support for H7a: while the model results technically fit the apriori hypothesis (stronger trust/reliance relationship in late stage automation with both risk types high compared to early stage automation with both risk types low), the risk subtypes' individual interactions make important caveats for generalizations.

Visual inspection of data plots

Examining figure 20 through visual inspection, it is obvious that high analytical risk (the top row of the figure's quadrants) produced more consistent trust/reliance regression lines. Clear outliers exist only in the low analytical risk conditions but mixed between automation type and affective risk conditions. It is possible this uneven variance across analytical risk was related to the poor goodness-of-fit test in Model M. None of these outliers were removed because figure 17 indicated that the separation between points in individual reliance proportions is not extreme. It is possible that, despite controlling for within-individual variation and establishing unique logit regressions per participant via the generalized mixed effects model approach, low analytical risk situations require a qualitatively different strategy for use of an automated tool. This experiment has provided a motivation for further investigation into whether the same underlying individual

strategies are used across high and low analytical risk conditions but did not specifically investigate what these differences are.

General Discussion

Across the two experiments conducted, we gained some evidence for specific ways that trust and reliance can vary due to risk types and automation stages. Affective risk often had an impact on trust whereas analytical risk and automation stage did not. Similarly, reliance did not vary due to experimental conditions, but we did find effects when considering how strongly trust predicts reliance across conditions. Hypotheses and results are summarized below in table 14.

Table 14.

Summary of hypotheses and results.

Hypothesis		Support	Findings
H1a	Reliance is expected to lower as analytical risk increases	No	Experiment one and two found null results
H1b	Reliance is expected to increase as affective risk increases	No	Experiment two found null results
H2	Reliance is expected to be higher when both risk types are high than when both are low.	No	Experiment two found null results
H3a	Trust is expected to increase as affective risk increases	No, contrasted	Experiment two showed decrease in trust over time with affective risk
H3b	Trust is expected to remain the same across analytical risk changes	Yes	Experiment one and two found null results
H4	The relationship of trust and reliance is expected to strengthen as affective risk increases.	Yes	Experiment one found that trust was more predictive of reliance at high analytical risk, experiment two found the same main effect
H5	Reliance is expected to be higher in higher risk conditions	No	Experiment two found null results
H6a	Trust is expected to be higher in high risk conditions	No	Experiment two found null results
H6b	Trust is expected to be higher in late stage automation conditions	No	Experiment two found null results
H7a	The relationship of trust and reliance is expected to be weaker in early stage automation conditions	No	Experiment two found main effects of stronger relationship with early stage automation. Interactions showed cases where late automation was stronger.
H7b	The impact of high risk conditions is expected to intensify the impact of automation stage's effect on the trust-reliance relationship	Mixed	Experiment two found stronger relationship when both risk types were high in late stage automation but weaker relationship with early automation in low risk conditions. Interactions support a more nuanced view.

In both experiments, there was relatively little variation by risk or automation on trust alone. Most existing literature on risk and automation use that measured trust did so with an analytical risk manipulation (Lyons & Stokes, 2012; Perkins et al., 2010; Satterfield et al, 2017). These studies did not find variation in trust related to risk alone. Therefore, our results support that risk of an analytical nature does not have an impact on trust attitudes. However, we did find that affective risk caused a steeper decrease in trust over time than without affective risk present. This influence of affect on trust over and above analytical considerations is well-supported by previous theories of trust and reliance: trust in automation is most influentially generated by affective processes (Lee & See, 2004). Our findings are novel with respect to past work by Lee and See that focused mainly on affect towards the automation directly, because affective stimuli were environmental within the broader task and not embedded within a participant's relationship directly with the automation. The environmental affective influence on trust in automation has been investigated with positive affect specifically. Multiple experiments found that participants were more likely to trust an automated tool when in a positive affect condition (Merritt, 2011; Stokes et al., 2010). This gives initial evidence for a simple contrasting nature of positive and negative affect in the environment increasing or decreasing trust in automation, respectively. According to our experiment on negative affect and existing research on positive affect, not only is trust dependent on interpersonal affective judgements with automation, but those judgements are contextualized in environmental affective cues.

Experiment one also investigated a measure of trust postulated to be closer to reliance, while not reliance itself: behavioral trust (Satterfield et al., 2017). This represents a behavioral interaction with an automated tool that is not yet the ultimate moment of reliance engagement with the tool. This could be important practically, as automated tools can involve interactions

that are not yet reliance but may be related (i.e., peeking at a suggestion before moving forward with the suggested action). We investigated their results through the lens of reliance and behavioral trust to ensure the constructs were delineated properly and found support for their delineation. While their experiment did not measure reliance, they used behavioral trust as a proxy for reliance in the discussion. Our results caution this approach in other experiments or applications. While behavioral trust was nearly always preceding reliance, this was dependent on whether individuals demonstrated subjectively low- or high-trust of the automation. Subjectively low-trust participants almost always performed behavioral trust actions before reliance and subjectively high-trust individuals used a different strategy by only engaging in behavioral trust occasionally before reliance. Behavioral trust, as properly delineated from the construct of reliance, has not been a subject of much research, but our results indicate it may have a unique place worthy of further investigation in the attitude-behavior spectrum of trust to reliance.

In addition to trust measurements with the present automated tool, trust propensity was also measured to assess whether participants had stable attitudes toward automation in general. Our experimental results gave mixed support for the importance of trust propensity. Experiment one showed that trust in the specific automated tool decreased over time, but only if trust propensity was generally low. Experiment two did not show any effects of trust propensity on trust in the automated tool. There is past literature that supports similar conclusions to those found in experiment one: that users with high trust propensity will have higher specific trust with moderately reliable automation compared to low trust propensity users (Merritt & Ilgen, 2008). One difference in experiment two that could account for the lack of significant propensity effects is the nature of the automation: peeking was not possible, so the tool was less flexible to participants across automation stages. Another study by Merritt showed positive trust propensity

effects on immediate trust: when automation made obviously correct decisions, trust propensity significantly increased (Merritt et al., 2013). It is possible the greater flexibility of automation in experiment one made its decisions more obvious to those participants. Further, affective manipulations were ineffective in experiment one, so it is also possible that effects of trust propensity were not as strong as those of negative affect that were present only in experiment two.

Overall, trust appears to be resistant to effects related to the manipulations of these experiments. There was an overall slight decrease in trust across experiments, indicating that our reliability level was slightly below acceptable. Automation stage on its own only influenced a user's trust over time. Late stage automation users maintained steady trust levels compared to a decrease in early stage automation users. Previous research on trust and automation stages found similar results (Calhoun, Draper, & Ruff, 2009; Lorenz, Nocera, Rottger, 2002). Because a general decrease in trust was found controlling for automation stage (indicating low reliability perception), it is possible those in the late automation stage condition did not have adequate transparency to diagnose the sub-optimal reliability of the automation but those with early stage automation were able to make more informed judgements.

Risk is generally minimal in its effects on trust as well. Analytical risk had no discernable effect on a user's trust. Affective risk did not broadly have an effect on trust, however, the presence of affective risk did cause a steeper decline in the already present downward slope of trust across experimental blocks. Therefore, it is important to consider affective risk when measuring trust, especially over a series of trust judgements.

In addition to trust attitudes, reliance behaviors were also investigated across both experiments. Similar to trust, reliance behaviors on their own (without considering trust) were

fairly consistent across experimental manipulations. Both experiment one and two showed no significant effects of analytical risk on reliance. This is generally contrasted with related findings of risk on automation reliance, where higher risk generally is related to less automation reliance (Perkins et al., 2010). In experiments with explicit feedback on correct/incorrect performance, both reliabilities were set to 70% for the automation and found less reliance at higher risk (Ezer et al., 2008; Satterfield et al., 2017). Our automation reliability was set to 76% across both experiments. While past work has only experimentally investigated calibration across reliability magnitudes of 20% (Wiegmann, Rich, & Zhang, 2001), it is theoretically possible that this 6% reliability difference could account for our null results compared to significant past results because of our automation being perceived as fundamentally more reliable compared to past work. If this were the case, it is possible that risk is only related to reliance directly at lower reliability levels. In contrast to the general body of research findings and our current experiment, some other past work found more automation reliance in higher risk situations (Lyons & Stokes, 2012). Their differences could be accounted for in that reliance was only measured indirectly by considering disuse of an adjacent human adviser.

Similar to analytical risk, affective risk did not have a significant main effect on reliance alone. Perkins and colleagues considered that an affective component of risk in their experiment may have caused a sudden increase in reliance when their hazards were unexpected and more vividly risky (2010). This claim is less supported than ours based on experimental results, as they did not measure affective manipulations on participants. Therefore, the results are somewhat speculative, and we consider our findings more robustly supported (no singular effects of affective risk on reliance).

In experiment two, automation stage was varied across conditions, but results did not indicate any changes to reliance due to the main effect. While some research has indicated a trend for greater reliance with earlier stage, more transparent automation, this was only seen across varying reliability levels of the automation (Seong & Bisantz, 2008). Again, similarly to risk types, main effects of reliance alone due to automation stage manipulations may only appear when reliance is variable also, especially at 70% reliability or lower. Despite the generally null results for main effects on reliance (and trust), there were notable findings based on the predictive power of trust on reliance across experimental manipulations.

Only experiment one showed a general effect of trust attitudes predicting reliance behaviors, but both experiments showed an effect of trust predicting reliance more strongly when analytical risk was high. Taken together, higher risk situations attenuate, or are necessary for, trust's predictive power of reliance with a specific automated tool. This is mostly in line with the little research that has compared the trust/reliance relationship across risk levels (Perkins et al., 2010). No work has investigated affective risk's influence on trust/reliance; therefore, our results provide initial evidence that high affective risk, when controlling for analytical risk and automation stage, increases the trust/reliance relationship similarly to analytical risk alone. The effects of analytical and affective risk together on trust/reliance were not seen in the interaction. In this way, these two risk types do not have a simple summative or multiplicative effect on trust/reliance.

Early automation stage, when controlling for risk types, showed a stronger trust/reliance relationship than late stage automation. Past research shows that early stage automation is trusted more than late stage automation (Rovira, McGarry & Parasuraman, 2007; Verberne et al., 2012), and our study indicates that this generally leads to stronger trust/reliance relationship as

well. It was hypothesized that late stage automation would show a stronger trust/reliance relationship because there would be less evidence outside of trust judgements for a user to weigh when choosing to rely on late stage automation, compared to early stage automation. However, this was not indicated by our results. It may be the case that more evidence about the automation's decision process allows for a better calibration of a user's trust and reliance.

Measurements of trust predicting reliance also depend on the interaction of risk types and automation stage. Our results found that the trust/reliance relationship of each risk type alone was stronger with earlier stage automation. However, the trust/reliance relationship did not vary by automation stage when both risks types were high, but late stage automation was more predictive than early stage automation with no high risk conditions present. This, as mentioned previously, gives some indication that analytical risk and affective risk do not simply combine to add more risk. Previous theoretical work indicates that a rational calculation of risk with automation is unlikely when affective components are present in the risk (Lee, 2006). A possible reason for this interaction is that risk types alone form a critical, narrow cognitive style. Negative affect has been shown to create such a style in decision-making (Lee, 2006; Mosier & Fischer, 2010). This may lead to disuse of trusted late stage automation, as the critical cognitive style leads to strong motivation to acquire more specific information in a decision-making process. When this extra information is not available in late stage automation, it could lead to disuse of the automation (whether or not the automation is trusted). On the other hand, early stage automation would provide extra information congruent with the cognitive style. Analytical risk on its own, controlling for affect, would theoretically result in more motivation for the participant to carefully consider decision options (Harinck et al., 2007). This motivation could inherently promote a more critical cognitive style as well.

If a critical cognitive style were the underlying cause for the weak-to-negative relationship of trust and reliance with individual risk types and late stage automation, it would then follow that both risk types together ameliorate the critical cognitive style (based on our model results). This would account for why both risk types together with late stage automation showed a strongly positive trust/reliance relationship compared to early stage automation with no high risk conditions. There is some related theoretical evidence of this where a study found that the judged benefit of a risky scenario was low when affect was negative, compared to positive affect (Finucane, Alhakami, Slovic, & Johnson, 2000). In our study, the benefit of the high analytical reward could be seen as low in benefit when affect was negative. This could have resulted in an ambiguous cognitive style. The present study did not measure these mechanisms, as it was previously unknown if affective risk and analytical risk had interactive effects in general or across automation stages at all. Risk types' interaction with automation stage does point to a change in the cognitive style, based on our findings.

Ultimately, results from this study show that Hoff and Bashir's characterization of risk (2013) as a situational variable influencing trust/reliance is insufficient. Risk, analytically, exerts a cognitive change for users of automation. Further, risk can contain affective components that are also impactful in an automation user's cognition. While the objective component of analytical risk is situational, its ability to measure trust and reliance is dependent on the internal state of the user (affective risk). Therefore, it would be more appropriate to consider risk types from a cognitive vantage point, rather than a purely external perspective that focuses on automation use, as situational effects on automation trust/reliance are dependent on the internal cognitive states of the user.

Limitations

Our experiments used a moderate reliability for all forms of automation (about 75%). Reliability is an important feature for trust formation and reliance behaviors over time. Our results can therefore only speak to moderately reliable automation, and the manipulations' effects may change if automation reliability were drastically lower or higher. Another consideration not measured in our experiments was the qualitative distinction of reliance ('all is well' signal) and compliance ('alarm' signal) with the automation (Dixon & Wickens, 2006). It is possible that risk types or automation stages may differentially affect one of these subtypes of behavior with automation. We chose not to include this variable as it would have been challenging or impossible to measure these differences in late stage automation (as the automation did not give information about its signal to the participant).

Future Research

These experiments have provided evidence that risk exists in multiple forms when considering the effect on trust/reliance with automation. While our discussion has speculated on the exact mechanisms that underpin the cognitive differences across conditions, future research could include target manipulations of cognitive style to see if these produce similar results as risk types more generally. Future research could also consider how risk types might change across automation reliability levels. Risk types may play a different role if automation is close to perfect or far below levels of acceptability.

Conclusion

Risk is an important variable to consider when measuring trust in automation and how that trust affects a user's behavioral reliance. Trust judgements are susceptible to risk, particularly when risk involves notable negative affective influences on a user's emotional state.

Trust is predictive of reliance, but this is dependent on how risky the situation is in general and the stage of automation in the tool at hand. Further, risk types can form unique cognitive states that affect users' mental strategies of automation use; how congruent an automation stage is to the present cognitive state can have a strong influence on how much trust judgements can predict subsequent reliance.

Automation has become a common feature of modern life, and accidents are high profile when they involve how humans trust and rely on automation (Ferris, 2016). The results from this set of experiments can help inform our understanding of these accidents. As many situations involving a failure of trust calibration require extensive accident reporting (National Transportation Safety Board, 2017), this can help provide descriptions of the operator's current cognitive state based on situational features and how such cognitive stages are related to their interaction with automation. Results from the current work can also help to inform not only what happens when something goes wrong with automation, but how engineers can design automation that prevents accidents from happening.

If risk of a certain type is present, engineers can use that information to have automation that shifts its decision-making transparency based on context, such as airport security risk levels in an x-ray scanning tool, traffic complexity around a partially self-driving car, or the context of use with a medical imagery scanning device. In the latter example, automated medical scanners support many doctors who report varying levels of trust in their attitudes towards the scanners (Harris, 2019). Our results could help inform a theoretical interpretation of these trust attitudes (and subsequent reliance) by appropriately considering the risk sub-types that are present when the scanners are used. While medical imaging analysis would always be analytically risky, some situations may be more affectively risky. In an active surgical room, the affective risk perceived

by a doctor would be conceivably different from a quiet, calm office. Therefore, the doctor's trust could differentially impact reliance on the automated scanner, especially when also considering the information stage that the scanner supports the doctor.

Trust has not always been predictive of reliance in research around automation. These novel experiments show that the relationship of trust and reliance is systematically dependent on risk types and automation stages. Both risk types (affective and analytical) have unique effects and must be considered in tandem on new research surrounding risk and automation use. Further, risk types must be considered from a cognitive perspective, as it is in the user's cognition that risk's influence on trust/reliance with certain types of automation are considered before action is taken.

REFERENCES

Albert A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71, 1.

Bates, D., Maechler, M., Bolker B., and Walker S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. http://CRAN.R-project.org/package=lme4

Berenson, M., Levine, D., Szabat, K. A., & Krehbiel, T. C. (2012). Basic Business Statistics: Concepts and Applications. Pearson Higher Education AU.

Boettcher, W. (2004). Presidential Risk Behavior in Foreign Policy. Palgrave MacMillan.

Bradley, M. M., Lang, P. J., & Cuthbert, B. N. (1993). Emotion, novelty, and the startle reflex: Habituation in humans. Behavioral Neuroscience, 107, 970–980.

Buhrmester, M., Kwang, T., Gosling, S. D., Buhrmester, M., Kwang, T., & Gosling, S. D. (2017). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality Data?, 6(1), 3–5.

Calhoun, G., Draper, M., & Ruff, H. (2009). Effect of level of automation on unmanned aerial vehicle routing task. Proceedings of the Human Factors International, 197–201. Retrieved from http://pro.sagepub.com/content/53/4/197.short

Cameron, A. C., & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. Journal of Econometrics, 46(3), 347-364.

Chen, Q., Kwok, O., Luo, W., and Willson, V. L. (2012). The impact of ignoring a level of nesting structure in multilevel mixture model: a Monte Carlo study. Structural Equation Modeling. 17, 570–589. doi: 10.1080/10705511.2010.510046

Coan, J. A., & Allen, J. J. (Eds.). (2007). Handbook of Emotion Elicitation and Assessment. Oxford university press.

Dan-Glauser, E. S., & Scherer, K. R. (2011). The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. Behavior Research Methods, 43(2), 468.

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.

Ezer, N., Fisk, A. D., & Rogers, W. A. (2008). Age-Related Differences in Reliance Behavior Attributable to Costs Within a Human-Decision Aid System. Human Factors, 50(6), 853–863. https://doi.org/10.1518/001872008X375018

Ferrari, A., & Comelli, M. (2016). A comparison of methods for the analysis of binomial clustered outcomes in behavioral research. Journal of Neuroscience Methods, 274, 131-140.

Ferris, R. (2016). Tesla Autopilot system's 'limitations' played 'major role' in 2016 crash: NTSB. CNBC. Retrieved from https://www.cnbc.com/2017/09/12/operational-limitations-of-tesla-autopilot-system-played-major-role-in-2016-crash-ntsb.html

Finucane, M. L., Alhakami, A. L. I., Slovic, P., & Johnson, S. M. (2000). The Affect Heuristic in Judgments of Risks and Benefits. Journal of Behavioral Decision Making, 17, 1–17.

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. MIS Quarterly, 27(1), 51-90.

Harinck, F., Van Dijk, E., Van Beest, I., & Mersmann, P. (2007). When gains loom larger than losses: Reversed loss aversion for small amounts of money. Psychological Science, 18(12), 1099–1105. https://doi.org/10.1111/j.1467-9280.2007.02031.x

Harris, R. (2019). How Can Doctors Be Sure A Self-Taught Computer Is Making The Right Diagnosis? National Public Radio. Retrieved from https://www.npr.org/sections/health-shots/2019/04/01/708085617/how-can-doctors-be-sure-a-self-taught-computer-is-making-the-right-diagnosis. Retrieved on 04/21/2019.

Herwig, U., Baumgartner, T., Kaffenberger, T., Brühl, A., Kottlow, M., Schreiter-Gasser, U., ... & Rufer, M. (2007). Modulation of anticipatory emotion and perception processing by cognitive control. Neuroimage, 37(2), 652-662.

Hoff, A., & Bashir, M. (2013). Trust in automation. IEEE Intelligent Systems, 28(1), 84–88. https://doi.org/10.1109/MIS.2013.24

Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi: 10.1002/0471722146

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language, 59(4), 434-446.

Kleiber, C., and Zeileis, A. (2008). Applied Econometrics with R. New York: Springer-Verlag. ISBN 978-0-387-77316-2. URL https://CRAN.R-project.org/package=AER

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the Open Affective Standardized Image Set (OASIS). Behavior Research Methods, 49(2), 457–470. https://doi.org/10.3758/s13428-016-0715-3

Kuznetsova A., Brockhoff, P.B., Christensen R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software, 82(13), 1–26. doi: 10.18637/jss.v082.i13.

Lee, J. D. (2006). Affect, attention, and automation. Attention: From theory to practice, 73-89.

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

Lenz, G. S., Huber, G. A., & Lenz, G. S. (2014). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk, (July 2012). https://doi.org/10.2307/23260322

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. Psychological Bulletin, 127(2), 267–286. https://doi.org/10.1037/0033-2909.127.2.267

Lorenz, B., Di Nocera, F., Röttger, S., & Parasuraman, R. (2002). Automated fault-management in a simulated spaceflight micro-world. Aviation Space and Environmental Medicine, 73(9), 886–897.

Lüdecke D (2019). sjstats: Statistical Functions for Regression Models (Version 0.17.3). doi: 10.5281/zenodo.1284472, https://CRAN.R-project.org/package=sjstats.

Lyons, J. B., & Stokes, C. K. (2012). Human-Human Reliance in the Context of Automation. Human Factors: The Journal of the Human Factors and Ergonomics Society, 54(1), 112–121. https://doi.org/10.1177/0018720811427034

Marcin, J. P., & Romano, P. S. (2007). Size matters to a model's fit. Critical Care Medicine, 35(9), 2212-2213.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integration model of organizational trust. Academy of Management Review, 20(3), 709–729.

Merritt, S. M. (2011). Affective Processes in Human−Automation Interactions. Human Factors: The Journal of the Human Factors and Ergonomics Society, 53(4), 356–370. https://doi.org/10.1177/0018720811411912

Merritt, S. M., & Ilgen, D. R. (2008). Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. Human Factors: The Journal of the Human Factors and Ergonomics Society, 50(2), 194–210. https://doi.org/10.1518/001872008X288574

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. Human Factors, 55(3), 520-534.

Mongin, P. (1997). Expected utility theory. Handbook of Economic Methodology. Edward Elgar Publishing.

Mosier, K. L., & Fischer, U. (2010). The Role of Affect in Naturalistic Decision Making. Journal of Cognitive Engineering and Decision Making, 4(3), 240–255. https://doi.org/10.1518/155534310x12844000801122

National Transportation Safety Board. (2017). Driver Errors, Overreliance on Automation, Lack of Safeguards, Led to Fatal Tesla Crash. Retrieved from https://www.ntsb.gov/news/press-releases/Pages/PR20170912.aspx

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation. Human Factors: The Journal of the Human Factors and Ergonomics Society, 56(3), 476–488. https://doi.org/10.1177/0018720813501549

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

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and

Cybernetics - Part A: Systems and Humans, 30(3), 286–297. https://doi.org/10.1109/3468.844354

Perkins, L., Miller, J. E., Hashemi, a., & Burns, G. (2010). Designing for Human-Centered Systems: Situational Risk as a Factor of Trust in Automation. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54(25), 2130–2134. https://doi.org/10.1177/154193121005402502

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models: Applications and Data Analysis Methods (Vol. 1). Sage.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. Human Factors, 49(1), 76–87. https://doi.org/10.1518/001872007779598082

Satterfield, K., Baldwin, C., de Visser, E., & Shaw, T. (2017, September). The Influence of Risky Conditions in Trust in Autonomous Systems. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 61, No. 1, pp. 324-328). Sage CA: Los Angeles, CA: SAGE Publications.

Schaefer, A., Nils, F., Philippot, P., & Sanchez, X. (2010). Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. Cognition and Emotion, 24(7), 1153–1172. https://doi.org/10.1080/02699930903274322

Seong, Y., & Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. International Journal of Industrial Ergonomics, 38(7-8), 608-625.

Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation-induced" complacency": Development of the complacency-potential rating scale. The International Journal of Aviation Psychology, 3(2), 111-122.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. European Journal of Operational Research, 177(3), 1333–1352. https://doi.org/10.1016/j.ejor.2005.04.006

Stokes, C. K., Lyons, J. B., Littlejohn, K., Natarian, J., Case, E., & Speranza, N. (2010). Accounting for the human in cyberspace: Effects of mood on trust in automation. 2010 International Symposium on Collaborative Technologies and Systems, CTS 2010, 180–187. https://doi.org/10.1109/CTS.2010.5478512

Verberne, F. M. F., Ham, J., & Midden, C. J. H. (2012). Trust in Smart Systems. Human Factors, 54(5), 799–810. https://doi.org/10.1177/0018720812443825

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. Theoretical Issues in Ergonomics Science, 2, 352–367.

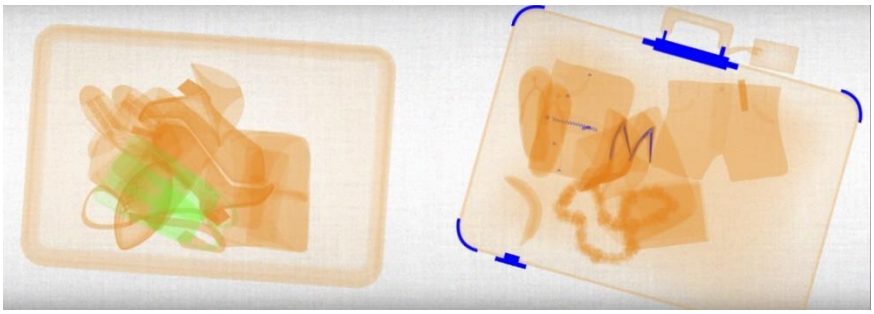
Wilson, K. M., de Joux, N. R., Finkbeiner, K. M., Russell, P. N., & Helton, W. S. (2016). The effect of task-relevant and irrelevant anxiety-provoking stimuli on response inhibition. Consciousness and Cognition, 42, 358-365.

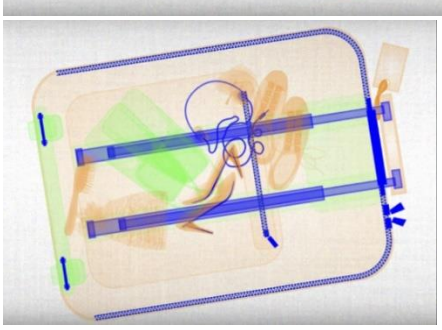
Wright, C. I., Fischer, H., Whalen, P. J., McInerney, S. C., Shin, L. M., & Rauch, S. L. (2001). Differential prefrontal cortex and amygdala habituation to repeatedly presented emotional stimuli. Neuroreport, 12, 379–383.

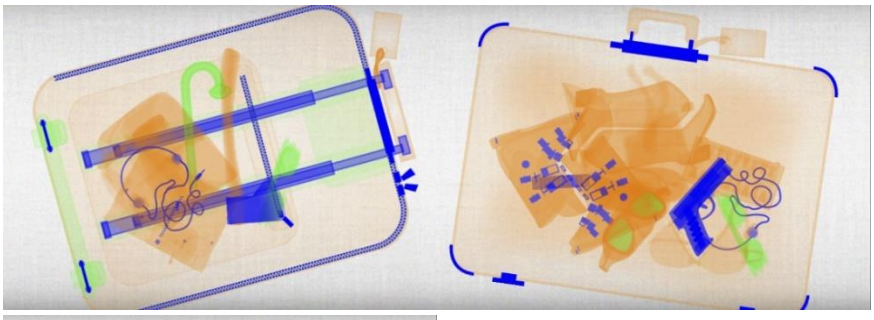
APPENDICES

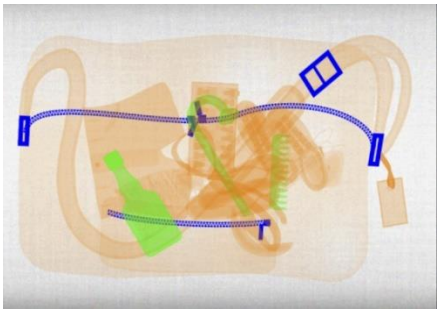
Item 1.

X-ray target images (1-3 are clear, 4-6 contain contraband)









Item 2.

Affective Video Stimuli

Movie	Length	Emotion	Description
The Blair Witch Project	3:57	Fear	"Anxious-provoking scene by the end of the movie: Heather (Heather Donahue), and Mike (Michael Williams) – who is filming Heather - are looking for Joshua (Joshua Leonard) in the woods, at night. They hear screams, apparently form Joshua. They find a house, from where the screams are coming. Mike and Heather go to the second floor. Next, Mike comes back without Heather. Heather's screams can be heard. Eventually, Mike's camera falls on the ground, and keeps filming."
The Shining	4:15	Fear	"Jack (Jack Nicholson) pursues his wife with an axe. "
Blue	0:40	Neutral	"A man clears out the drawers of his desk; a woman arrives walking in an alley. She greets another woman and continues walking."
Blue	0:16	Neutral	"A person passes a piece of aluminum foil through the window of a car."

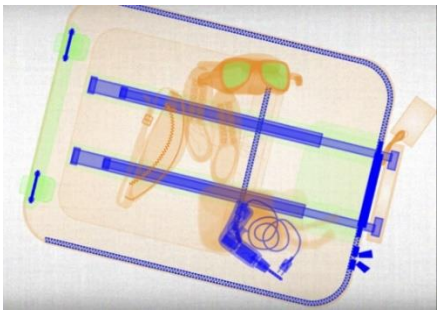
Item 3.

Experiment one images (first four neutral and second four fearful).

Item 4.

Screen shots of experiment one protocol

1. Target image



2. Decision point

What is your decision for the luggage you just viewed?

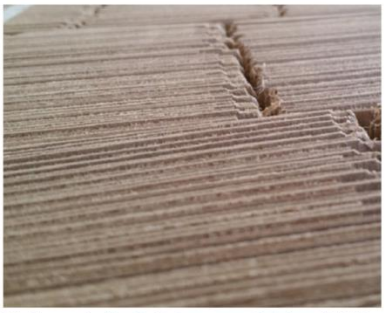
Check (contraband present)

Clear (NO contraband present)

Use AWD decision

Peek at AWD recommendation (10% point cost)

3. If 'peek' was chosen



What is your decision for the luggage you just viewed? [AWD recommendation]

Check (contraband present) **[RECOMMENDED BY AWD]**

Clear (NO contraband present)

>>

4. Final point/feedback screen

Your response was **correct**. There was a contraband item present in the luggage.

Your total score is 9

Click 'next' when you're ready to view the next slide, as it will start immediately.

>>

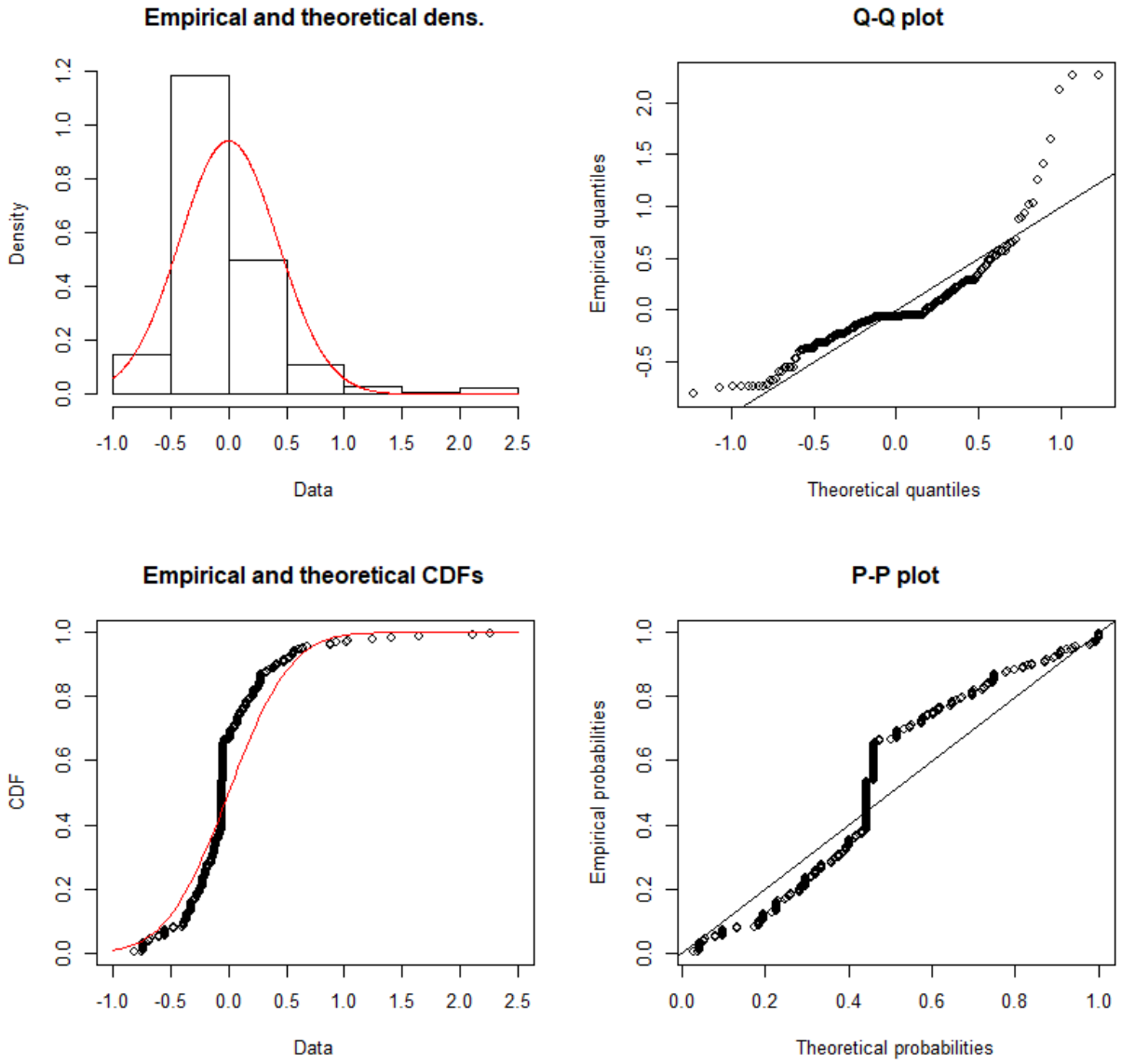
Item 5.

Experiment one descriptive statistics of manipulation check outcomes.

	High affect	Low affect
N	34	31
Affective means (sd)		
Block 1	1.01 (1.09)	0.85 (1.15)
Block 2	0.42 (.63)	0.68 (.98)
Block 3	0.39 (.65)	0.74 (1.09)
Block 4	0.38 (.72)	0.75 (1.07)

Item 6.

Experiment 1 manipulation check diagnostic plots



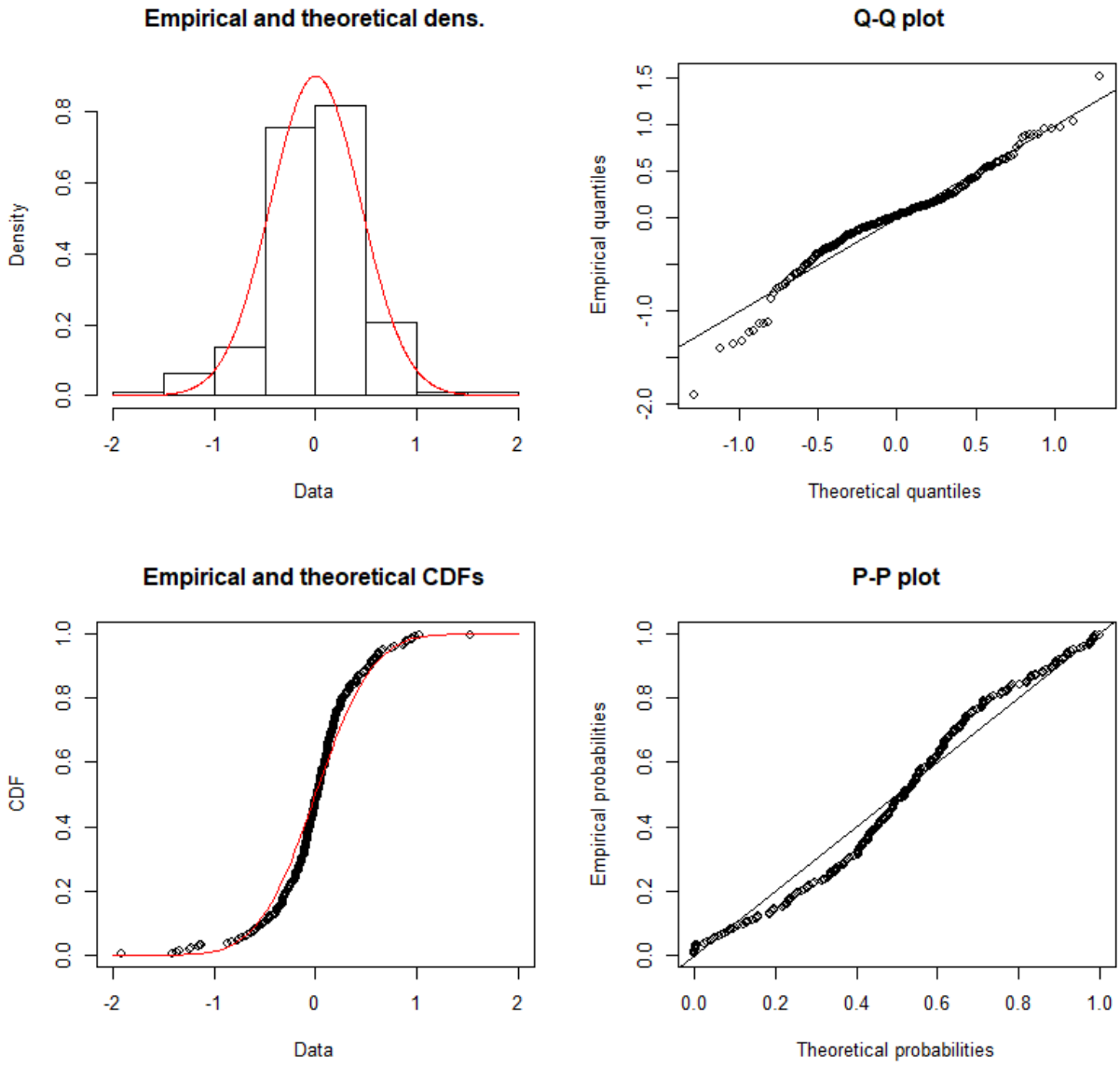
Item 7.

Experiment one trust descriptives.

Condition	Propensity (SD)	Trust (SD)			
		Block 1	Block 2	Block 3	Block 4
High Analytical / High Affective	2.29 (.57)	2.41 (1.06)	2.40 (1.14)	2.10 (1.14)	2.07 (1.18)
High Analytical / Low Affective	2.76 (.48)	2.72 (.38)	2.77 (.52)	2.83 (.58)	2.82 (.61)
Low Analytical / High Affective	2.41 (.36)	2.31 (.87)	2.39 (.78)	2.33 (1.05)	1.92 (1.00)
Low Analytical / Low Affective	2.72 (.63)	2.51 (.96)	2.52 (1.20)	2.43 (1.24)	2.36 (1.20)

Item 8.

Experiment 1 trust model diagnostic plots.



Item 9

Experiment one exploratory trust model C formulation.

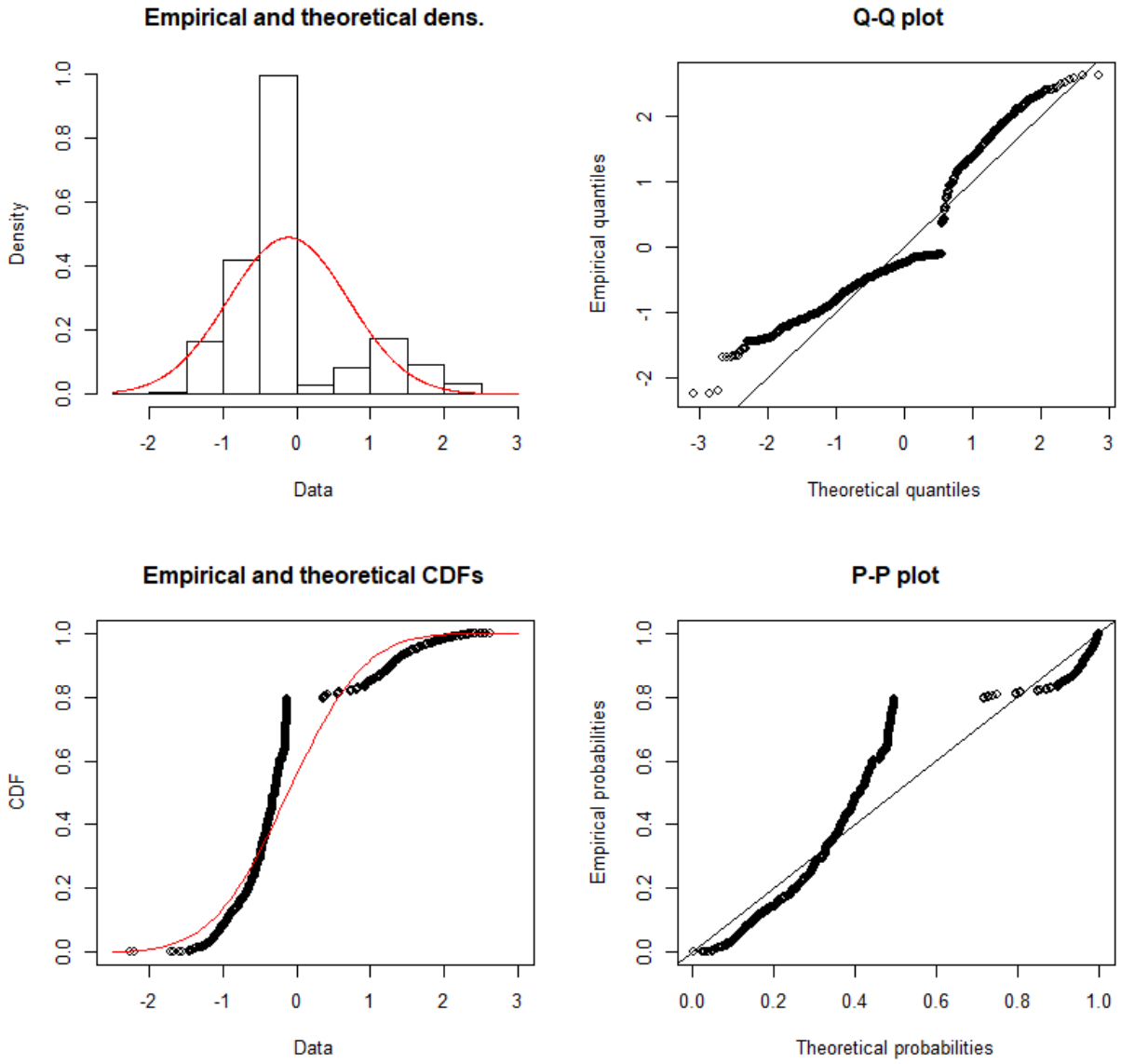
Level 1: $Trust_{it} = \beta_{0it} + \beta_{1it}(Block) + r_{it}$

Level 2: $\beta_{0i} = \gamma_{00} + \gamma_{01}(Analytical\_high) + \gamma_{02}(Affective\_high) + \gamma_{02}(Trust\_propensity) + u_{0i}$

$\beta_{1i} = \gamma_{10} + \gamma_{11}(Trust\_propensity)$

Item 10.

Experiment one reliance diagnostics



Item 11.

Experiment one reliance Model D formulation.

$$\text{Level 1: Reliance}_{ibt} = \square_{0ibt} + e_{ibt}$$
$$\text{Level 2: } \square_{0bt} = \beta_{00bt} + \beta_{10bt}(\text{Trust}) + r_{bt}$$
$$\text{Level 3: } \beta_{00t} = \gamma_{000t} + \gamma_{001}(\text{Analytical\_high}) + \gamma_{02}(\text{Affective\_high}) + u_{00t}$$
$$\beta_{01t} = \gamma_{010t}(\text{Analytical\_high})$$

Item 12.

Experiment one reliance Model D results with values in Odds-Ratios (Confidence intervals)

Fixed Effects	Model E
Reference Intercept, γ_{00}	.04*** (.01-.13)
Trust, γ_{10}	1.37* (.96-1.95)
High analytical, γ_{11}	.44 (.09-2.29)
High affective, γ_{12}	1.167 (.44-3.07)
High analytical*Trust, γ_{13}	1.64 (.96-2.80)
Random Effects	
Between-person variance (τ_{00})	3.29
Within-person block variance (e_{00})	.05
Within-person trial variance (σ^2)	3.49

*p<.05, **p<.01, ***p<.001

Item 13.

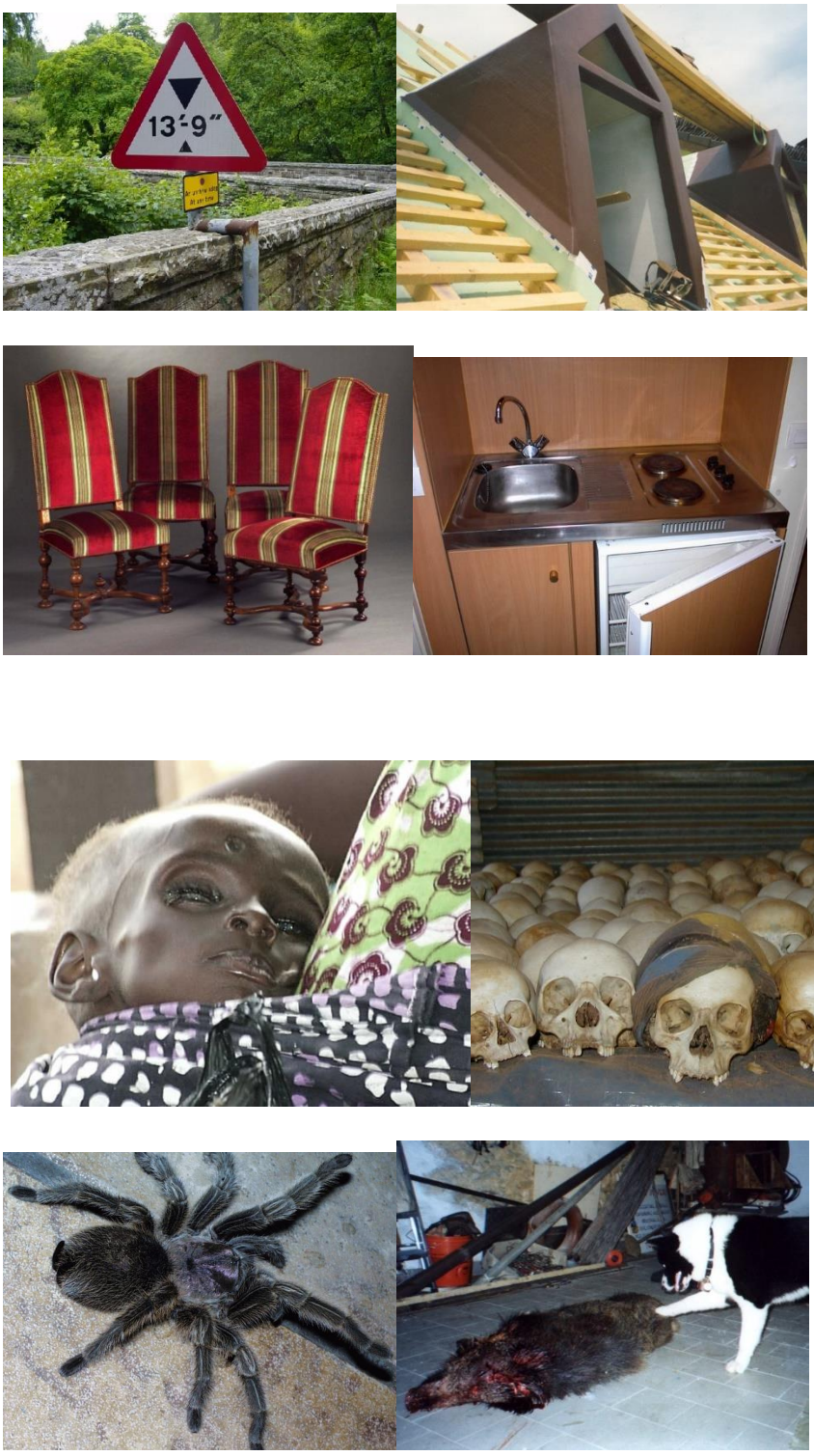
Experiment one reliance Model G results with values in Odds-Ratios (Confidence intervals)

Fixed Effects	Model H
Reference Intercept, γ_{000}	.01*** (.00-.03)
BT, γ_{010}	23141.51*** (836.69-640053.35)
ST, γ_{100}	1.92*** (1.35-2.74)
BT*ST, γ_{110}	.39 (.13-.1.16)
Random Effects	
Between-person variance (τ_{00})	4.98
Within-person block variance (e_{00})	.41
Within-person variance (σ^2)	3.29

*p<.05, **p<.01, ***p<.001

Item 14.

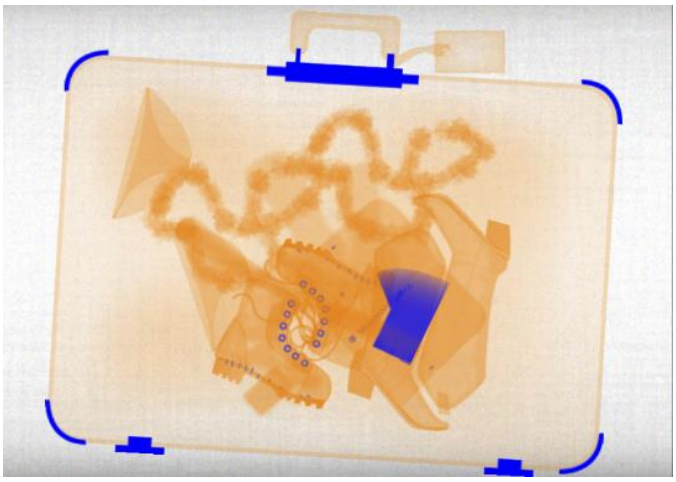
Experiment two affective image stimuli (first four neutral and second four fearful)



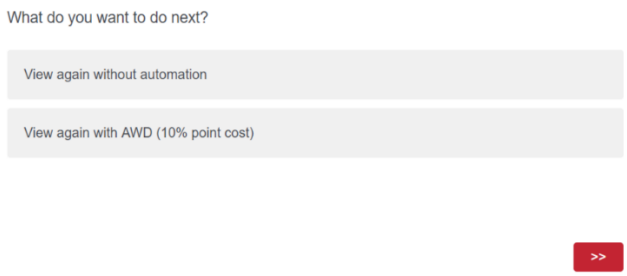
Item 15.

Automation A protocol with automation engagement

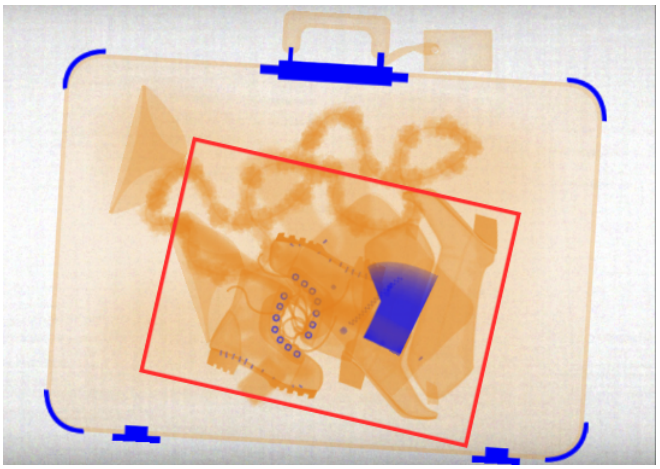
1. Target image



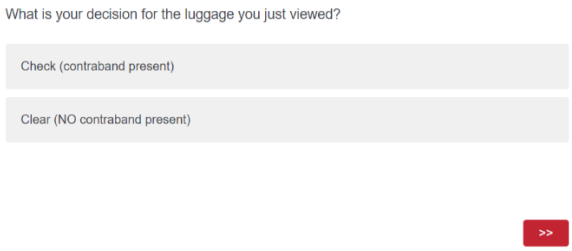
2. Reliance



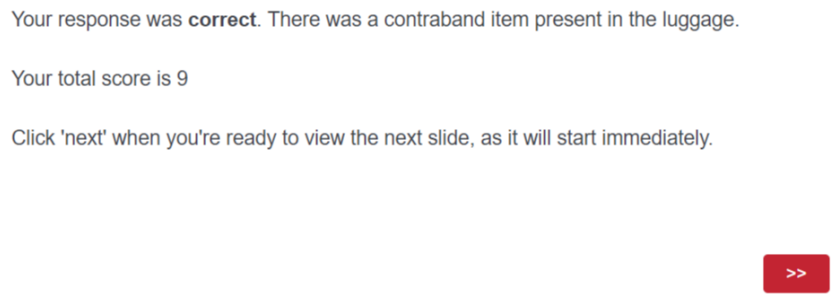
3. Automation engaged



4. Decision



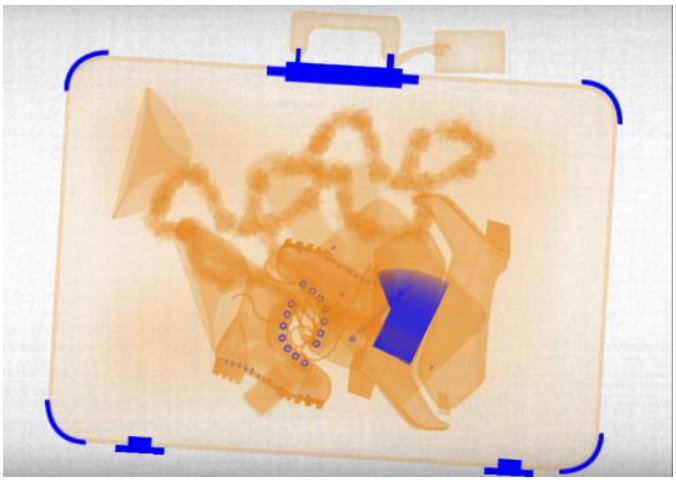
5. Outcome



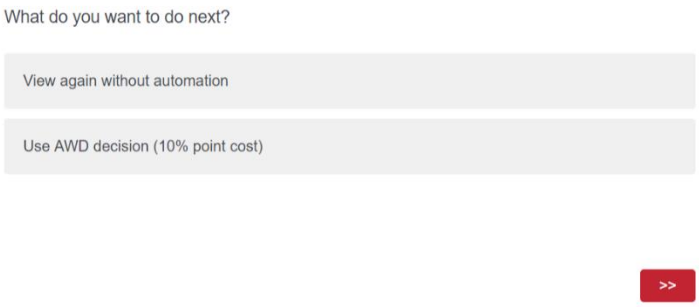
Item 16.

Automation B protocol with automation engagement

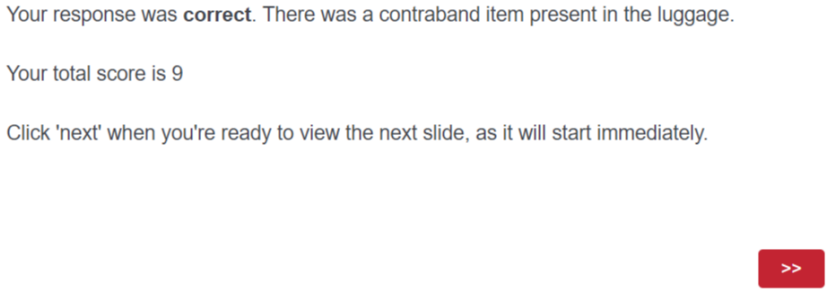
1. Target image



2. Reliance



3. Outcome (no second viewing with reliance)



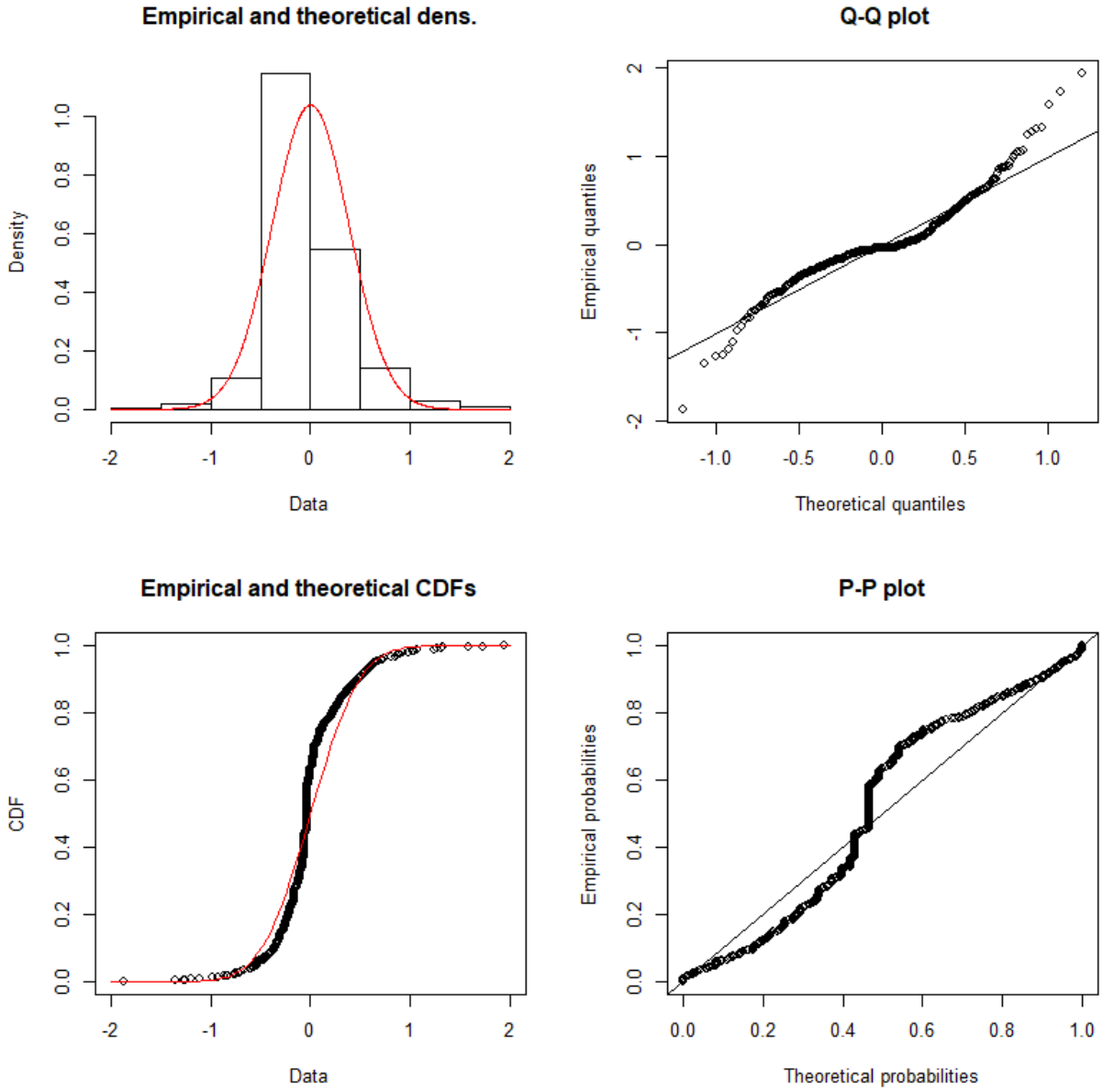
Item 17.

Experiment two manipulation check descriptives

	High affect	Low affect
N	34	31
Affective means (sd)		
Block 1	1.16 (1.22)	0.43 (.70)
Block 2	1.00 (1.13)	0.57 (.76)
Block 3	1.17 (1.15)	0.49 (.69)
Block 4	.93 (1.06)	0.59 (.73)

Item 18.

Experiment two manipulation check model diagnostics



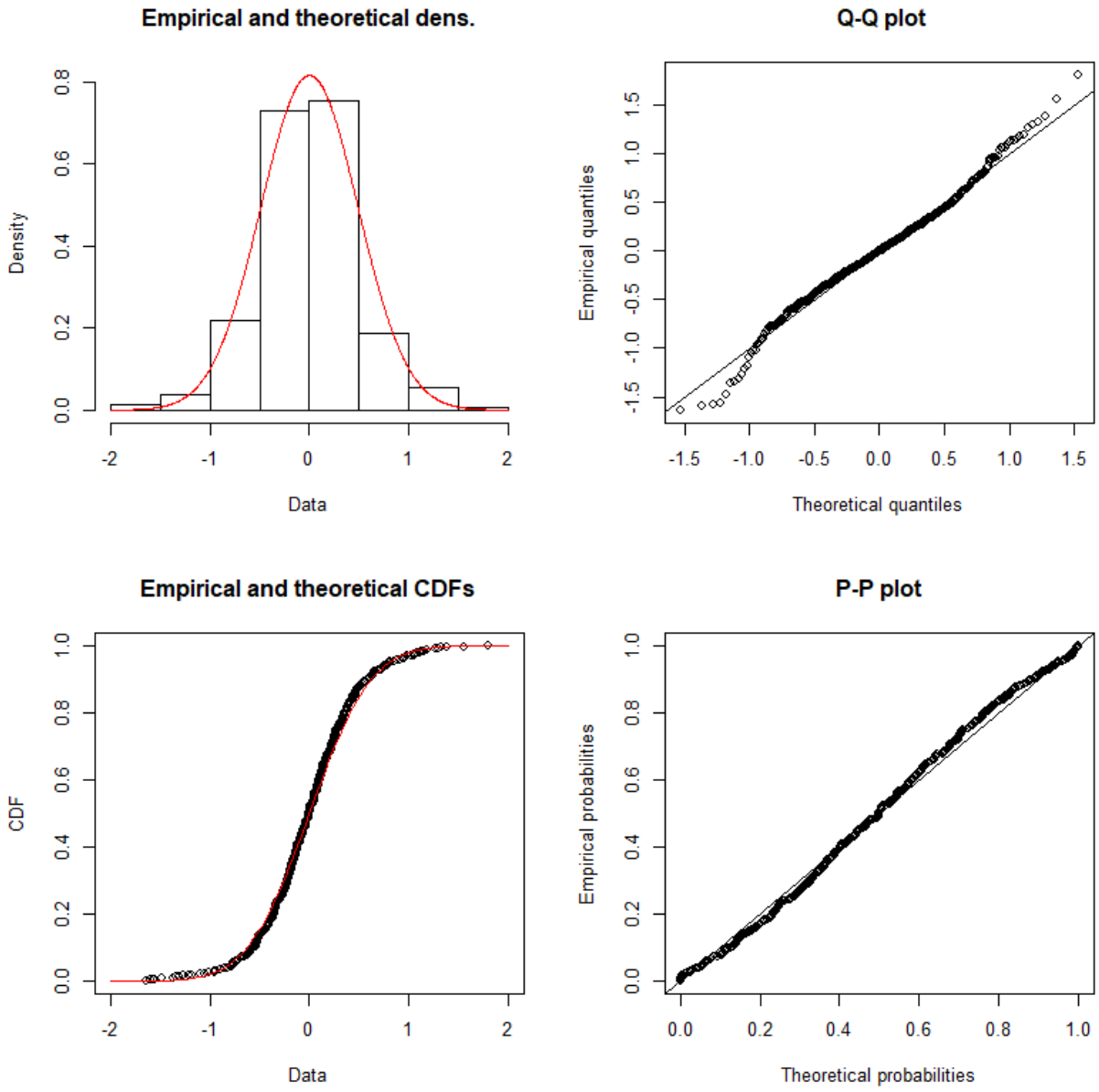
Item 19.

Trust descriptives for experiment two

Condition	Propensity (SD)	Trust (SD)			
		Block 1	Block 2	Block 3	Block 4
A / High Analytical / High Affective	2.68 (.59)	2.50 (.64)	2.25 (.74)	1.93 (.77)	1.67 (1.05)
A / High Analytical / Low Affective	2.67 (.58)	2.42 (.86)	2.14 (1.00)	1.90 (1.01)	1.79 (1.13)
A / Low Analytical / High Affective	2.63 (.46)	2.60 (.84)	1.97 (.97)	1.85 (.94)	1.75 (1.11)
A / Low Analytical / Low Affective	2.52 (.54)	2.24 (.91)	2.08 (.98)	2.10 (.95)	1.90 (.97)
B / High Analytical / High Affective	2.58 (.46)	2.41 (.61)	2.24 (.94)	2.21 (1.03)	2.00 (.93)
B / High Analytical / Low Affective	2.52 (.42)	2.61 (.84)	2.26 (1.00)	2.26 (.99)	2.18 (.96)
B / Low Analytical / High Affective	2.64 (.49)	2.64 (.74)	2.50 (1.04)	2.31 (1.05)	2.15 (1.27)
B / Low Analytical / Low Affective	2.60 (.55)	2.43 (.66)	2.07 (.75)	2.14 (.73)	2.12 (1.00)

Item 20.

Experiment two trust model diagnostics



Item 21.

Experiment two trust propensity Model K formulation

Level 1: $Trust_{it} = \beta_{0it} + \beta_{1it}(Block) + r_{it}$

Level 2: $\beta_{0i} = \gamma_{00} + \gamma_{01}(Analytical\_high) + \gamma_{02}(Affective\_high) + \gamma_{03}(Automation\ type) + \gamma_{04}(Propensity) + u_i$

$\beta_{1i} = \gamma_{10}(Block) + \gamma_{11}(Automation\ type) + \gamma_{12}(Propensity) + \gamma_{13}(Automation\ by\ Propensity)$

$\beta_{2i} = \gamma_{20}(Analytical*Affective)$

Item 22.

Experiment two reliance Model L formulation

$$
\begin{aligned}
\text{Level 1: Reliance}_{ibt} &= \square_{0ibt} + e_{ibt} \\
\text{Level 2: } \square_{0bt} &= \beta_{00bt} + \beta_{01bt}(\text{Trust}) + r_{ib} \\
\text{Level 3: } \beta_{00t} &= \gamma_{000t} + \gamma_{001}(\text{Analytical\_high}) + \gamma_{002}(\text{Affective\_high}) + \\
&\quad \gamma_{003}(\text{Automation}) + \gamma_{004}(\text{Analytical\_high*Affective\_high}) + \\
&\quad \gamma_{005}(\text{Analytical\_high*Automation}) + \\
&\quad \gamma_{006}(\text{Affective\_high*Automation}) + \\
&\quad \gamma_{007}(\text{Analytical\_high*Affective\_high*Automation}) + u_{00t} \\
\beta_{01t} &= \gamma_{010t} + \gamma_{011}(\text{Analytical\_high}) + \gamma_{012}(\text{Affective\_high}) + \\
&\quad \gamma_{013}(\text{Automation}) + \gamma_{014}(\text{Analytical\_high*Affective\_high}) + \\
&\quad \gamma_{015}(\text{Analytical\_high*Automation}) + \\
&\quad \gamma_{016}(\text{Affective\_high*Automation}) + \\
&\quad \gamma_{017}(\text{Analytical\_high*Affective\_high*Automation})
\end{aligned}
$$

Item 23.

Experiment two reliance Model L results

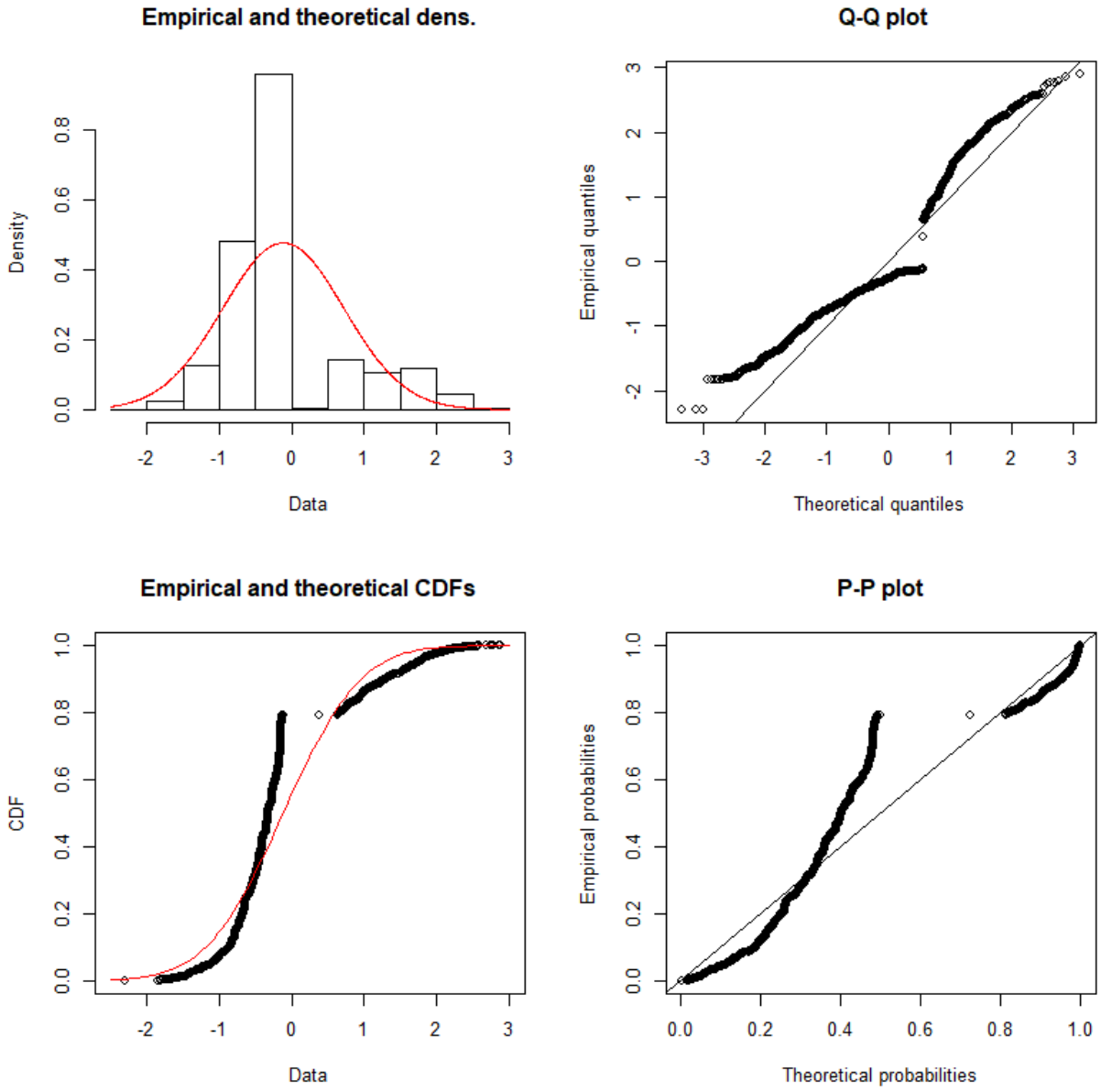
Reliance Model L results with values in Odds-Ratios (Confidence intervals).

Fixed Effects	Odds-Ratio	Confidence Intervals
Reference Intercept, γ_{000}	0.16*	0.04-0.70
Analytical_high,γ_{001}	0.3	0.35-1.13
Affective_high, γ_{002}	0.58	0.04-2.08
Automation, γ_{003}	0.2	0.09-3.59
Analytical_high*Affective_high, γ_{004}	0.38	0.03-1.59
Analytical_high*Automation, γ_{005}	4.18	1.08-4.82
Affective_high*Automation, γ_{006}	18.46	1.60-6.42
Analytical_high*Affective_high*Automation, γ_{007}	0.12	0.03-5.31
Trust, γ_{010}	0.63	1.57-7.59
Analytical_high*Trust,γ_{011}	2.28*	0.25-70.61
Affective_high*Trust, γ_{012}	3.2**	1.29-264.35
Automation*Trust, γ_{013}	3.45**	0.18-1.30
Analytical_high*Affective_high*Trust, γ_{014}	0.49	0.09-0.72
Analytical_high*Automation*Trust, γ_{015}	0.25*	0.04-0.25
Affective_high*Automation*Trust, γ_{016}	0.1***	0.00-5.78
Analytical_high*Affective_high*Automation*Trust, γ_{017}	9.81*	2.48-38.77
Random Effects		
Between-person variance (τ_{00})	3.29	
Between-block variance (e_{00})	.59	
Within-person variance (σ^2)	2.71	

*p<.05, **p<.01, ***p<.001

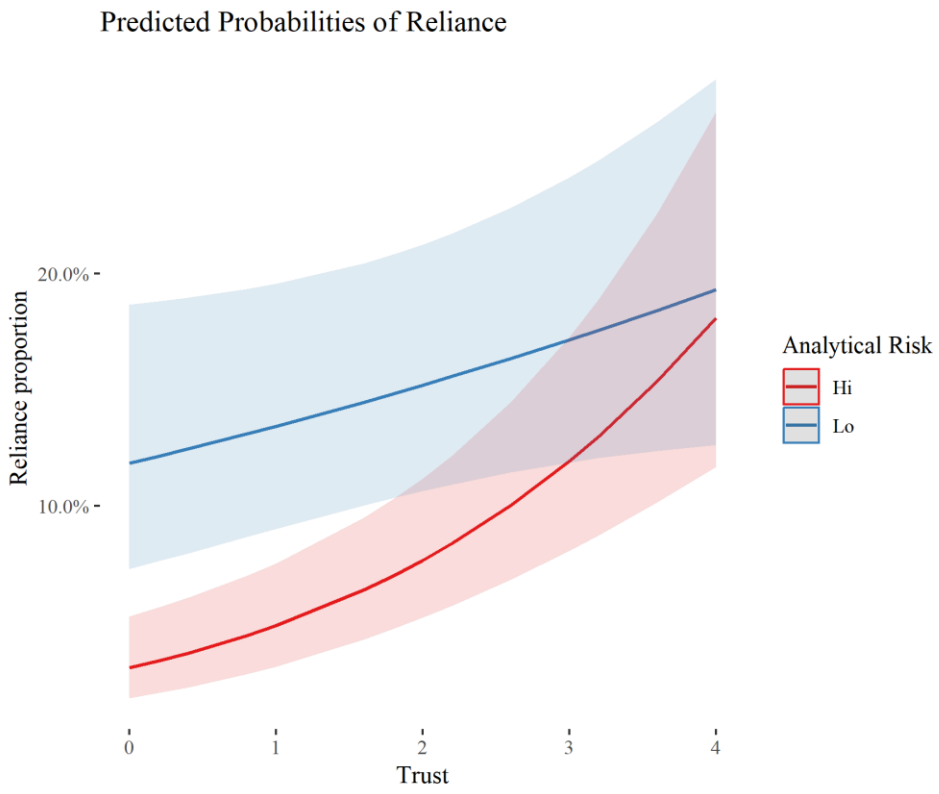
Item 24.

Experiment two reliance diagnostics



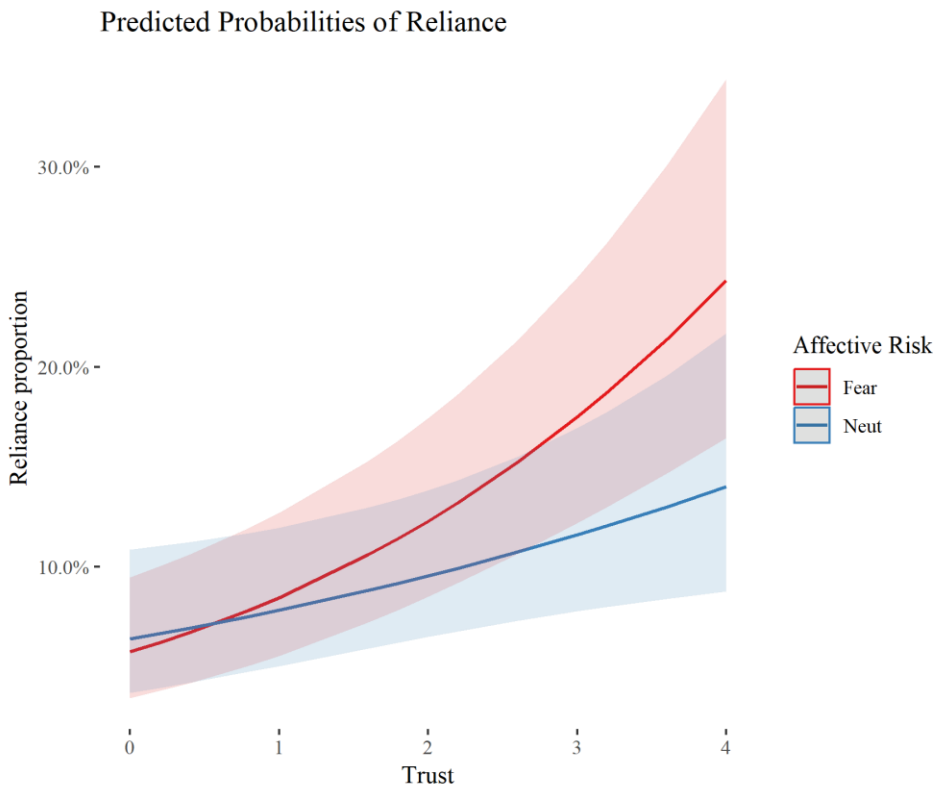
Item 25.

Trust by analytical risk interaction in Model M



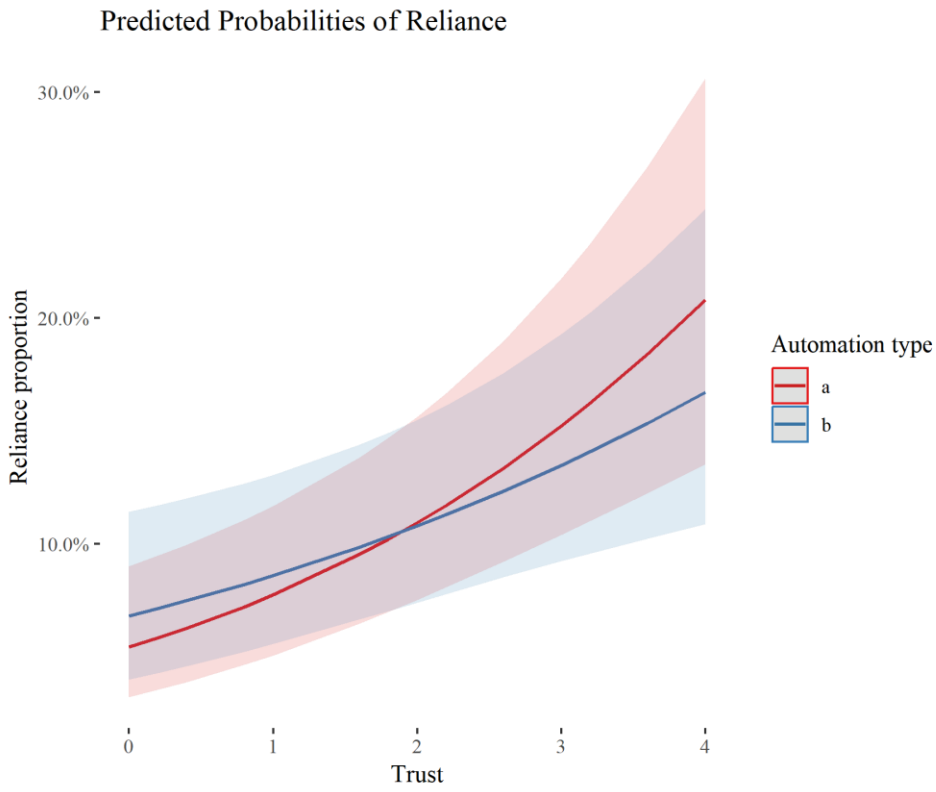
Item 26.

Trust by affective risk interaction in Model M



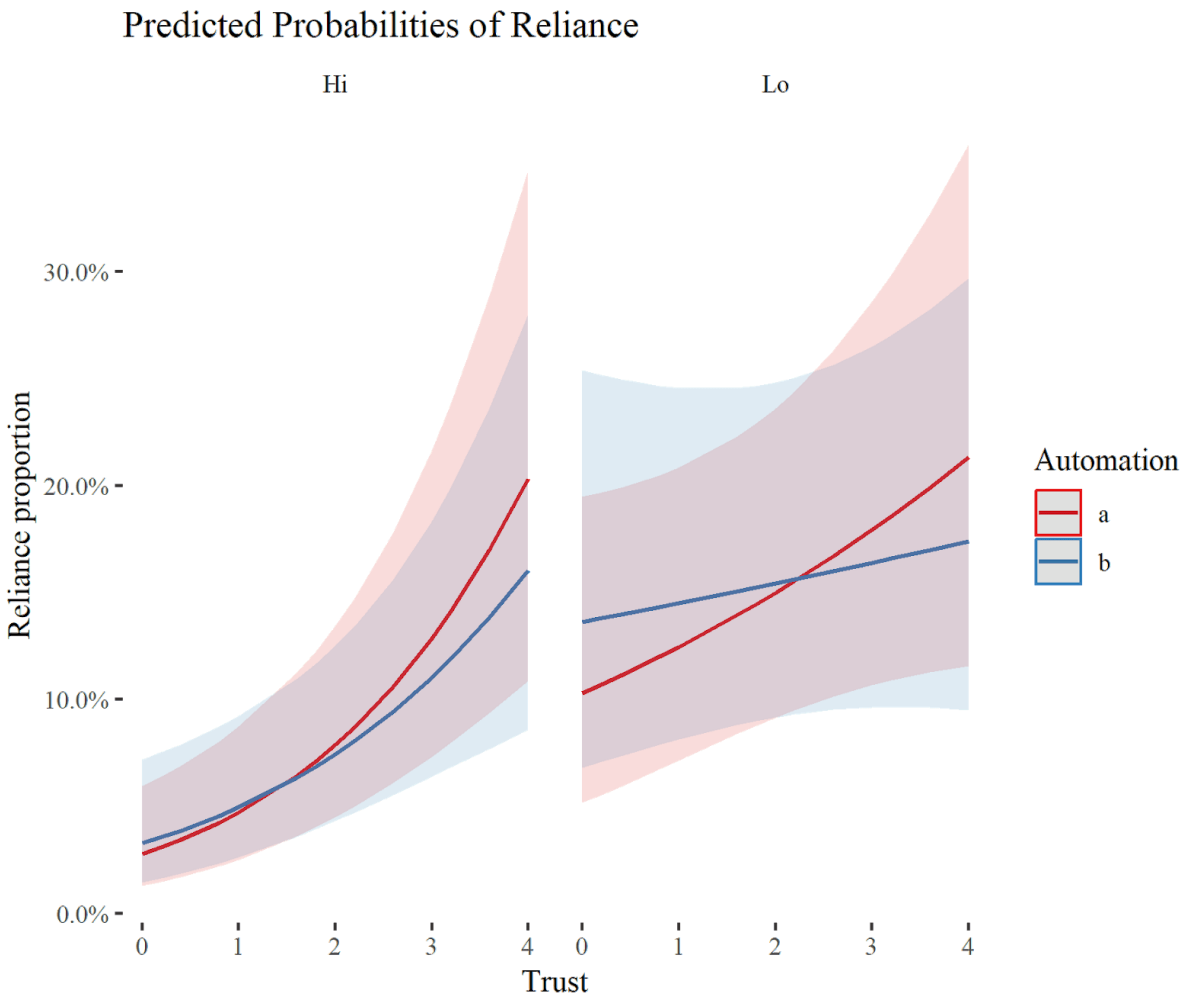
Item 27.

Trust by automation stage interaction in Model M



Item 28.

Trust by analytical risk by automation stage interaction in Model M



Item 29.

Trust by affective risk by automation stage interaction in Model M