

Clemson University

TigerPrints

All Theses

Theses

May 2021

The Impact of Automation Etiquette on User Performance and Trust in Non-Personified Technology

Zachary Joseph Guyton

Clemson University, zjguyt@gmail.com

Follow this and additional works at: https://tigerprints.clemson.edu/all_theses

Recommended Citation

Guyton, Zachary Joseph, "The Impact of Automation Etiquette on User Performance and Trust in Non-Personified Technology" (2021). *All Theses*. 3516.

https://tigerprints.clemson.edu/all_theses/3516

This Thesis is brought to you for free and open access by the Theses at TigerPrints. It has been accepted for inclusion in All Theses by an authorized administrator of TigerPrints. For more information, please contact kokeefe@clemson.edu.

THE IMPACT OF AUTOMATION ETIQUETTE ON USER
PERFORMANCE AND TRUST IN NON-PERSONIFIED TECHNOLOGY

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirement of the Degree
Master of Science
Human Factors Psychology

by
Zachary J. Guyton
May 2021

Accepted by:
Dr. Richard Pak, Committee Chair
Dr. Erika Rovira
Dr. Patrick Rosopa

ABSTRACT

Previous research has shown that good automation etiquette can yield positive effects on user performance, trust, satisfaction, and motivation. Automation etiquette is especially influential in personified technologies – users have increased etiquette expectations from technology that has human characteristics. Designers deliberately integrate etiquette into personified technologies to account for users’ anthropomorphization and meet user needs. The current study examined the impact of etiquette in non-personified technologies. The study aimed to demonstrate that automation etiquette also affects performance, trust, perceived workload, and motivation in technologies that possess little to no human characteristics. The study used a computer-based automation task to examine good and bad etiquette models and different domain-based perceived task-importance, or “criticality” levels (between-subjects) that contained various stages of automation and automation reliability levels (within-subjects). The study found that bad etiquette automation produced better performance in certain conditions. Confirming previous research, we found that users trust good etiquette automation more than bad etiquette automation in some trust categories. This study provides evidence that automation complexity correlates with automation etiquette’s impact – as automation complexity increases, so does automation etiquette’s impact on performance and in some cases trust. We found that bad automation etiquette can increase user’s subjective workload. Last, we confirmed that our domain-based task criticality manipulation was effective. Future research should examine additional domains, tasks, etiquette delivery mechanisms, and

etiquette scales coupled with varied degrees of automation complexity to better understand etiquette's role in human-automation interaction.

TABLE OF CONTENTS

	Page
TITLE PAGE.....	i
ABSTRACT.....	ii
LIST OF TABLES.....	v
LIST OF FIGURES.....	vi
INTRODUCTION.....	1
Why Etiquette Matters.....	2
The Benefits of Etiquette Based Design.....	3
Current Study.....	5
METHOD.....	6
Participants.....	6
Materials.....	7
Measures.....	11
Design.....	12
Procedure.....	15
RESULTS.....	16
Hypothesis 1: Performance, Etiquette, and Task Criticality.....	18
Hypothesis 2: Etiquette and Trust.....	25
Hypothesis 3: Performance, Etiquette, and Stage of Automation.....	28
Hypothesis 4: Trust, Etiquette and Stage of Automation.....	29
Hypothesis 5: Workload, Etiquette, and Task Criticality.....	30

Table of Contents Continued	Page
Additional Findings.....	32
DISCUSSION.....	34
Performance, Etiquette, and Task Criticality.....	34
Performance, Etiquette, and Stage of Automation.....	35
Etiquette, Criticality, and Trust.....	37
Trust, Etiquette and Stage of Automation.....	39
Workload, Etiquette, and Task Criticality.....	40
Limitations and Future Direction.....	41
Conclusion.....	43
APPENDICES.....	45
Appendix A: Automation Induced Complacency Potential.....	46
Appendix B: Trust Measure (Lee and Moray, 1994)	47
Appendix C: NASA-TLX (Hart and Staveland, 1988)	48
Appendix D: Motivation/Affect and Criticality Questionnaire	49
Appendix E: Pilot Test Description and results.....	50
Appendix F: Experimental Counterbalance Sample Size.....	52
REFERENCES.....	53

LIST OF TABLES

Table	Page
Table 1. Descriptive Statistics for Participant Demographics, CPRS, and AICP.....	16
Table 2. Descriptive Statistics for Etiquette and Criticality Conditions.....	17

LIST OF FIGURES

Figure	Page
Figure 1. Four components of low-criticality task interface Taxi Dispatching Task.....	9
Figure 2. Interface of high-criticality Battlefield Simulation Task.....	10
Figure 3. Etiquette feedback matrix – low-criticality taxi task	13
Figure 4. Etiquette feedback matrix – high-criticality targeting task.....	14
Figure 5. Significant interaction between etiquette and criticality on decision accuracy.....	20
Figure 6. Significant differences of performance and stage of automation.....	21
Figure 7. Significant differences of performance between high (.80) and low (.60) reliability on Decision Accuracy.....	22
Figure 8. Stage 3 Automation performance significantly faster in all conditions and sub-conditions.....	23
Figure 9. High reliability quicker response time than low reliability in the Stage 3 Automation taxi condition.....	23
Figure 10. Significant interaction. Bad etiquette participants did better than good etiquette participants on Secondary Task Performance in the low-criticality taxi condition.....	25
Figure 11. Higher trust with good etiquette on trust question 3 – “to what extent are you self-confident that you could successfully perform without the automation aid in this scenario?”	27
Figure 12. Interaction between reliability and stage of automation. Larger difference of stage of automation in the high reliable condition than the low reliable condition.....	28
Figure 13. Impact of etiquette on performance significant in stage 3 automation but not stage 2.....	29
Figure 14. Impact of etiquette on trust significant in stage 3 automation but not stage 2. Trust Question 2 - “to what extent do you rely on (i.e., actually use) the automation aid in this scenario?.....	30
Figure 15. Perceived workload interaction between etiquette and criticality.....	31

List of Figures continued	Page
Figure 16. Perceived workload interaction between reliability and automation stage	32
Figure 17. Perceived Criticality of different criticality and etiquette conditions.....	33
Figure 18. Low-criticality (taxi) task performance mapped onto Yerkes Dodson Curve.....	35
Figure 19. Stimulus driven attention spike from different etiquette conditions.....	37
Figure 20. Direct Relationship with Automation complexity and automation etiquette's impact.....	39

INTRODUCTION

Humans depend on etiquette in virtually every interpersonal interaction they encounter – meeting new people, talking with a loved one, addressing a superior at work, etc. The importance of human-human etiquette is obvious in its absence. What if a person reached out to shake a someone’s hand and they did not extend their arm to reciprocate? This would be instantly categorized as rude and potentially offensive. Etiquette is defined as the socially understood conventions that facilitate smooth and effective interactions between people (Hayes & Miller, 2011). Moreover, good etiquette can be classified in a binary manner of prescriptive norms – things that people should do; and prohibitive norms – things that people should not do (i.e., behaviors, verbal and non-verbal communications, expressions, and actions). Good etiquette depends on doing what is appropriate in context, not necessarily doing what is polite or nice. The role and importance of etiquette will continue to occupy a major role in human-human interaction. But what is etiquette’s role and importance when humans interact with automation and how does it impact performance and trust?

Designers, engineers, and programmers integrate rules of etiquette when developing human-technology interactions. Etiquette is particularly relevant in more personified or human-resembling technology such as voice-based assistants (e.g., Siri, Alexa, Google Assistant). These assistants adhere to social niceties and although this creates inefficiencies in communicative brevity (i.e., the extraneous please or thank you during conversations), users appreciate and expect these colloquial norms (Nass, 2004). Thus, increased etiquette in personified technology is understandable. Another example is

how an interruptive and uncooperative (bad etiquette) voice-automated call menu is likely to have negative impacts on performance, satisfaction, and may even cause users to disengage with the system altogether. But how important is etiquette in simpler, non-personified technology? Do users expect the same level of etiquette and if so, how will etiquette violations impact users? With human-technology and human-automation interaction increasing at an exponential level, these questions require further investigation.

Why Etiquette Matters

Systems that are not intentionally designed to follow rules of good etiquette may be perceived by users to have neutral or even bad etiquette. Negative or poor etiquette is rarely, if ever, deliberately integrated into design. Rude, interruptive, or even threatening interactions are created to be appropriate for the context (e.g., a demanding order to “stay back, danger, incoming train” at a subway station). Thus, “neutral etiquette” results from designers’ lack of implementing good etiquette. Unbeknownst to developers, neutral etiquette can quickly translate into negative or poor etiquette (e.g., persistent and distracting update reminders on computers or the terse, robotic commands signaling bagging errors at self-checkout lines). The aim of this study is to demonstrate that neutral or poor etiquette, even when integrated in simple automation and technology, is not sufficient. Therefore, good etiquette, or least the avoidance of users encountering negative emotions through poor or neutral etiquette, is critical in current design.

Understanding the nature of human’s anthropomorphization of machines provides important insight on why etiquette matters within technological interactions. Research

has shown that humans tend to subconsciously treat computers politely, even extremely basic ones (Reeves & Nass, 1996; Nass & Moon, 2000). Further research has defined eight specific categories that influence human's anthropomorphization. These categories are language use, voice, face, emotional manifestation, interactivity (especially over time), engagement with and attention to the user, autonomy, and the filling of roles traditionally filled by humans (Reeves and Nass, 1996; Nass and Moon, 2000; Nass, 2004). Although some of these categories seem complex (i.e., filling roles filled by humans, engagement with/attention to the user, emotional manifestation), surprisingly they can be invoked by simple technology. For example, Giga pets from the 1990s consisted of a simple, basic-feature, low technology gadget but could invoke complex human emotions, attachments, and interactions. Appreciating the existence of human's subconscious or conscious technological anthropomorphization provides a foundation to understanding the importance of human expectancy and etiquette. If users perceive humanlike characteristics from technology, they are likely to expect or at least respond to reciprocity, introducing etiquette into the equation.

The Benefits of Etiquette-Based Design

All other variables being constant, users trust and comply more with polite automation. Parasuraman and Miller (2004) conducted a study with both polite and rude automation assisting pilots with simulated flight alerting systems common in modern aircraft. The study used experienced personnel from the aviation industry (pilots and non-pilots) as participants and found increased performance and trust with the polite etiquette automation. The effects of automation etiquette were so profound in the experiment that

the good-automation-etiquette condition with low reliability (60% reliable) nearly matched performance of the poor-etiquette-condition with high reliability (80% reliable; Parasuraman and Miller, 2004). Bad automation etiquette, in the form of rudeness, has revealed detriments to user compliance, trust, and perceived workload (Miller et al., 2006). Yang and Dorneich (2018) found that effective etiquette integration in automated tutors yielded improvements to user motivation, confidence, satisfaction, and performance. In this study, Yang and Dorneich highlighted the importance of preventing negative interactions with automation – avoiding the inadvertent neutral to poor etiquette transition discussed above. These findings indicate that good system etiquette may increase performance and cause users to adopt a more synchronous calibration of either appropriately trusting (using) or distrusting (not using) automation based on etiquette.

As substantial as etiquette appears to be on human perceptions of automation and subsequent performance outcomes (Parasuraman and Miller, 2004), the role of etiquette on human-automation interaction requires further inquiry. First, Parasuraman and Miller only explored high-criticality tasks (i.e., flying a plane in a simulator). This leaves unanswered etiquette questions relating to lower criticality systems and is particularly relevant in determining if etiquette's effects apply to more widespread, everyday systems. Second, Parasuraman and Miller did not analyze etiquette across different stages of automation. Given that different stages of automation have distinct effects on human performance (Rovira, Pak, McLaughlin, 2017), etiquette may exert different effects for different stages. Third, Parasuraman and Miller used a small, sixteen participant sample size and the results of the study could be difficult to generalize to other users. The sixteen

participants included general aviation pilots and non-pilots. A more diverse and expansive participant pool completing an unfamiliar task would strengthen the findings of subsequent etiquette research. The final limitation in Parasuraman and Miller's research is the absence of exploring the relationship between workload and etiquette. It is plausible that the efficacy of etiquette found in their study is moderated by the level of workload imposed by the task.

The Current Study

The current study aimed to fill in the gaps remaining from Parasuraman and Miller's experiment by examining etiquette in a low-criticality task. Additionally, we attempted to replicate Parasuraman and Miller's high-criticality findings. We used a unique task paradigm that allows us to manipulate the perceived criticality of the task (via domain) without altering any other aspect of the task. Our fundamental goal was to corroborate existing conclusions that etiquette matters, even in minimally personified automation. We hoped to enhance the generalizability of this concept across multiple domains. Our study observed the effects of etiquette delivered by different stages of automation (stage 2 – information analysis, and stage 3 – decision support) (Parasuraman, Sheridan & Wickens, 2000). Our study examined etiquette in high and low-criticality tasks. We investigated etiquette's relationship with automation reliability (high, low). Last, we included a secondary task to better characterize how workload moderates the influence of etiquette.

Our initial hypotheses for the experiment

1. Hypothesis 1: Good automation etiquette will produce better performance than bad automation etiquette.
2. Hypothesis 2: Good automation etiquette will produce higher trust than bad automation etiquette.
3. Hypothesis 3: There will be main effects of etiquette on performance in stage 3 automation but not in stage 2 automation.
4. Hypothesis 4: There will be main effects of etiquette on trust in stage 3 automation but not in stage 2 automation.
5. Hypothesis 5: The bad etiquette automation will produce higher subjective workload than the good etiquette automation; and the high-criticality targeting task will produce higher subjective workload than the low-criticality taxi task.

METHOD

Participants

Two hundred and eight undergraduate students ages 18-22 (159 females, $M_{age}=18.4$, $SD=.8$; 49 males, $M_{age}=18.7$, $SD=.8$) from Clemson University were recruited from the SONA extra credit pool and received coursework credit for their participation in the study. Data from four participants was removed from analysis due to overall task performance of lower than fifteen percent. Fifteen percent was used as the cutoff to maximize sample size and include participants with poor performance. Only twelve participants scored lower than forty percent and all data used fell within three standard deviations of the mean of primary task performance.

Materials

Equipment. Data collection occurred through a web-based online study. Participants received access to the study through the Clemson Sona Psychology Research System following enrollment. Participants completed the experimental task from their home computers using desktops, laptops, or tablets. Participants were instructed to maximize their browsers to full screen, not use any background applications, and complete the study at a location with a good internet connection. Data was compiled and stored using an online repository.

Taxi Dispatching Task (Figure 1). This task represents the low-criticality condition and resembles a task used in previous studies (Rovira et al., 2017; Pak et al. 2017; Rovira, McGarry, & Parasuraman, 2007). The task display consists of four parts – a grid overlaid street map (right), an automated assistant interface (left), feedback display bar (bottom), and a communications input panel (top left) (Figure 1). The grid overlaid street map displays the task information through a series of four colored boxes shown simultaneously. Customers are represented by green boxes from one to six (displayed C1-C6), taxis are represented by red boxes from one to six (T1-T6), competing buses (extraneous distractors) are represented by yellow boxes from one to three (B1-B3), and the taxi dispatching headquarters is an orange box (HQ).

The automated assistant provides the participant with helpful task information. The utility of the information varies based on stage of automation. In the stage 2 condition, the automation provided a list of all possible taxi/customer pairings/distances

listed in a random, unsorted order. In the stage 3 condition, the automation provided a sorted list with optimal pairings ordered from best to worst (top to bottom).

The feedback display bar provides feedback after a trial is completed. The after-trial feedback varies based on the participant's performance (correct, incorrect, timeout). The feedback function is used to manipulate etiquette. The communications input panel delivers a secondary task. The panel displays one of fourteen different names rotating every six seconds. When the name "WARREN" appears among the fourteen, the participant is required to select the "Answer" button.

During the task, participants play the role of a taxi dispatcher located at the taxi headquarters. Participants' primary task is to match the closest customer/taxi pair. If two or more pairs of taxi/customers are the same distance, the participant is instructed to match the pair closest to the headquarters. Ten seconds are allotted for the task and

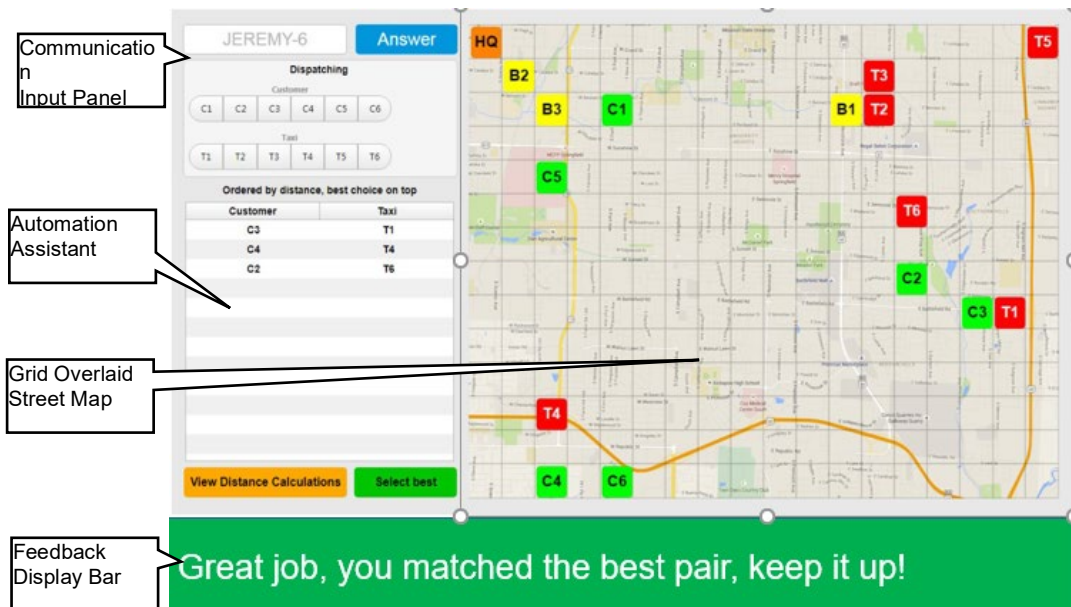


Figure 1. Four components of low-criticality task interface, Taxi Dispatching Task.

participants are instructed to choose quickly. Participants completed four fifty-trial blocks and answer questions on workload and trust between each block.

Battlefield Simulation Task (Figure 2). This task represents the high-criticality condition in the experiment and was also used in previous studies (Rovira et al., 2017; Pak et al. 2017; Rovira, McGarry, and Parasuraman, 2007). The task resembles the taxi dispatching task and contains the same four components (grid overlaid street map, automated assistant interface, feedback display bar, and communications input panel). The grid overlaid street map also displays four colored boxes shown simultaneously (Figure 2).

The battlefield task replaces the customers, taxis, and buses from the taxi dispatching task with enemy units (displayed red, E1-E6), friendly units (displayed green, A1-A6), friendly battalion units (displayed yellow, B1-B3), and the headquarters unit

(displayed orange, HQ) respectively. Additionally, the battlefield grid overlaid map display uses a satellite terrain background oppose to the street map background from the

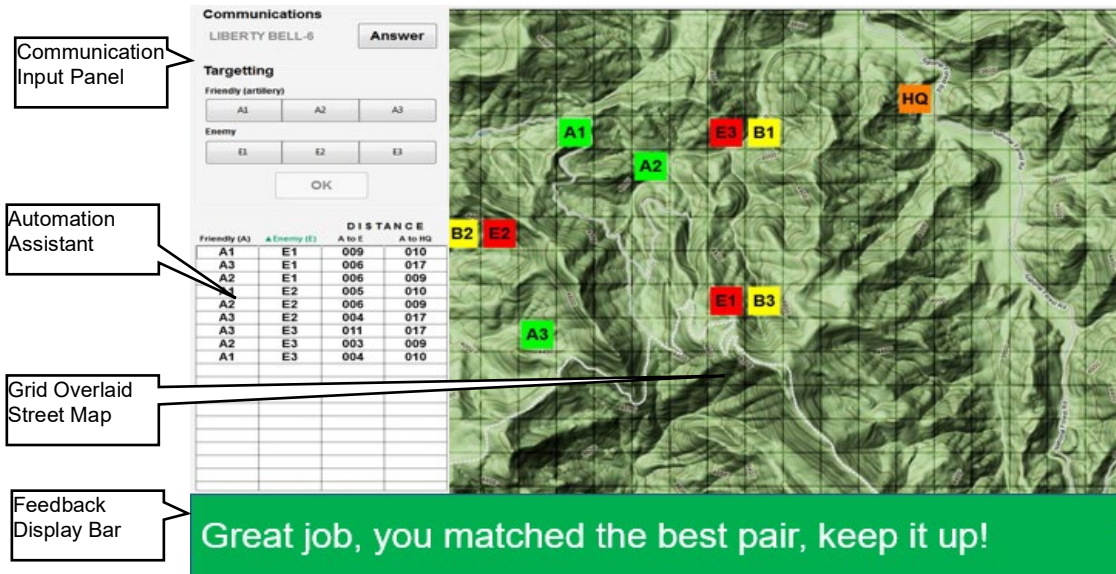


Figure 2. Interface of high-criticality Battlefield Simulation Task.

taxi task. Workload is manipulated through the number of enemy, friendly, and battalion units. The automated assistant, feedback display bar, and communications input panel serve the same purpose and possess the same capabilities and limitations as the taxi task.

During the task, participants play the role of a military battlefield commander located at the headquarters. Participants' primary task is to match the closest friendly/enemy unit pair. If two or more pairs of friendly/enemy units are the same distance, the participant is instructed to match the pair closest to the headquarters. Ten seconds are allotted for the task and participants are instructed to choose quickly. Participants completed four fifty-trial blocks and answered questions on workload and trust between each block.

Measures

Trust and individual differences. To control for participants' attitudes toward automation, we integrated the Automation-Induced Complacency Potential Scale (AICP; Merritt et al., 2019). This measure contains ten questions on a five-point scale ranging from 1 (strongly disagree) to 5 (strongly agree) and was taken after the participants completed the trials (Appendix A). Additionally, between each of the four fifty-trial blocks, we measured participants trust through a four-question evaluation validated in previous research (Lee & Moray, 1994; Appendix B).

Perceived workload. Participants completed the subjective NASA Task Load Index (NASA-TLX) after finishing each trial block (Appendix C). This assessment scale measured six dimensions: mental, physical, temporal, performance, effort, and frustration (Hart & Staveland, 1988). We aimed to identify differences in perceived workload between the etiquette conditions with this test.

Motivation/Affect. Participants completed a modified version of the Intrinsic Motivation Inventory (IMI) questionnaire after completing all trial blocks. The IMI measures intrinsic motivation defined as “doing an activity for its inherent satisfactions rather than for some separable consequence,” and has been used and validated in previous research. (Ryan and Deci, 2000; McAuley, Duncan, and Tammen (1987). We used this questionnaire to capture differences in participants' affective attitudes and motivation for future use toward the different etiquette conditions. The IMI consist of five questions scored on a scale of 1-7 (1=strongly disagree and 7=strongly agree (Appendix D).

Design

The experiment was a 2 (etiquette: good, bad) x 2 (criticality: high, low) x 2 (stage of automation: information analysis, decision automation) x 2 (automation reliability: low, high) mixed-factorial design. Dependent variables include (1) performance on primary/secondary task (2) trust in automation (3) perceived workload (4) motivation/affect toward automation use. Etiquette and criticality were between-subject variables while automation stage and automation reliability were within-subject variables.

Etiquette. Etiquette was a between-subjects factor with two levels, good and bad. Participants were either in the good or bad etiquette condition. To manipulate etiquette, our experiment focuses on communication style. Redressive language, as defined by Brown and Levinson, dictates the perceived tone and attitude of the automation (Brown and Levinson, 1987; Miller et al, 2006; Miller et al, 2008). The task contains redressive language through after-trial feedback. The good etiquette condition positive messages (correct responses) have a Praise, Inform, Encourage (PIE) construct (e.g., “great job, you matched the best pair, keep it up”) while the negative messages (timeout or incorrect response) have a Apologize/minimize, Inform, Encourage (AIE) construct (“I’m sorry, that’s incorrect, you will get the next one”). The bad etiquette condition positive messages strictly inform (e.g., “correct answer”) and the negative messages Call out, Blame, Quantify (CBQ) (e.g., “you are wrong, you cost the company the fare). Each etiquette condition contained twelve potential responses. Additionally, color and punctuation were strategically coupled with messages within the feedback display bar

(Figure 3, Figure 4). The good etiquette condition contained green with correct responses and no color with incorrect responses/timeouts. It punctuated with exclamation marks for correct responses and periods for incorrect responses. These manipulations aimed to accentuate good performance. Conversely, the bad etiquette condition contained red and punctuated with exclamation marks for incorrect responses/timeouts and periods for correct responses. These manipulations aimed to emphasize bad performance. A pilot test with 13 participants revealed the good etiquette messages were perceived as significantly more polite than the bad etiquette messages (Appendix G).

Low-Criticality Taxi Task		
Etiquette	Condition	Message after participant response
Good	Correct response	Great job! You matched the best pair. Keep it up!
	Correct response	Outstanding effort! That's correct! Continue the good work!
	Correct response	Excellent work! You matched the best pair! Keep up the good job!
	Correct response	Terrific job! That's correct! Maintain your performance!
	Time ran out	I'm sorry. Time ran out. You will get the next one (Lost taxi fare).
	Time ran out	Apologies. No time is left. There are plenty more opportunities (Lost taxi fare).
	Time ran out	Whoops. Time ran out. Plenty more customers (Lost taxi fare).
	Time ran out	I'm sorry. Time is up. You will get the next one (Lost taxi fare).
	Incorrect response	I'm sorry. That's incorrect. You will get the Next one (Lost taxi fare).
	Incorrect response	Apologies. That was not the best answer. Next time (Lost taxi fare).
	Incorrect response	Whoops. That is not correct. Plenty more customers (Lost taxi fare).
	Incorrect response	Sorry. That is not the best match. Still lots of business out there (Lost taxi fare).
Bad	Correct response	Correct answer.
	Correct response	Correct answer.
	Correct response	Correct answer.
	Correct response	Correct answer.
	Time ran out	You ran out of time and cost the company the fare!
	Time ran out	You were not fast enough and cost the company the fare!
	Time ran out	You were too slow and cost the company the fare!
	Time ran out	Not quick enough! You cost the company that fare!
	Incorrect response	You are wrong! you cost the company the fare!
	Incorrect response	Wrong answer! you cost the company that fare!
	Incorrect response	You made the wrong selection! You cost the company the fare!
	Incorrect response	You chose the incorrect answer! You lost the company that fare!

Figure 3. Etiquette feedback matrix reveals the message, color, and punctuation combinations participants see in the feedback display bar of the taxi task.

High-Criticality Battlefield Simulation Task		
Etiquette	Condition	Message after participant response
Good	Correct response	Great job! You destroyed the closest enemy! keep it up!
	Correct response	Outstanding effort! That's correct! Continue the good work!
	Correct response	Excellent work! You matched the best pair! Keep up the good job!
	Correct response	Terrific job! That's correct! Maintain your performance!
	Time ran out	I'm sorry. Time ran out. You will get the next one (Lost 1x platoon).
	Time ran out	Apologies. No time is left. There are plenty more opportunities (Lost 1x platoon).
	Time ran out	Whoops. Time ran out. Still plenty of enemy to destroy (Lost 1x platoon).
	Time ran out	I'm sorry. Time is up. You will get the next one (Lost 1x platoon).
	Incorrect response	I'm sorry. That's incorrect. You will get the next one (Lost 1x platoon).
	Incorrect response	Apologies. That was not the best answer. Next time (Lost 1x platoon).
	Incorrect response	Whoops. That is not correct. Plenty more enemy on the battlefield (Lost 1x platoon).
	Incorrect response	Sorry. That is not the best match. Still lots of enemy units out there (Lost 1x platoon).
Bad	Correct response	Correct answer.
	Correct response	Correct answer.
	Correct response	Correct answer.
	Correct response	Correct answer.
	Time ran out	You ran out of time and caused the platoon's demise!
	Time ran out	You were not fast enough and cost the battalion a platoon!
	Time ran out	You were too slow and cost the battalion that platoon!
	Time ran out	Not quick enough! You cost the battalion that platoon!
	Incorrect response	You are wrong! Your response resulted in a friendly platoon destruction!
	Incorrect response	Your response is incorrect! you cost the battalion a platoon!
	Incorrect response	You made the wrong selection! Your actions cost the battalion a platoon!
	Incorrect response	You chose the incorrect answer! You lost a platoon from the battalion!

Figure 4. Etiquette feedback matrix reveals the message, color, and punctuation combinations participants w see in the feedback display bar of the Battlefield simulation

Criticality. Criticality was a between-subjects factor achieved through the two separate tasks. Half of our participants only completed the high-criticality battlefield simulation task while half only completed the low-criticality taxi task. Previous research has shown that users perceive technology domains differently and operational/task features play a role in how users classify the importance of a task (Pak et al., 2016; Mosier and Fischer, 2012). The battlefield simulation task aimed to generate a highly critical environment in which participants perceive increased importance of performance and greater consequences for failures (i.e., the loss of friendly troops vs. the loss of a taxi customer).

Stage of automation. Stage of automation was a within-subjects factor where all participants were exposed to stage 2 automation (information analysis) and stage 3

automation (decision aid) through the automation assistant. The stage 2 automation assistant gave the user raw, unsorted information. In this condition, if the user decided to use the automation, they had to sift through numerous options to find the best one. The stage 3 automation assistant ordered the options from best to worst. If the user decided to use the automation, they just need to select from the top option on the list for the correct answer.

Automation reliability. Automation reliability was a within-subjects factor where all participants were exposed to high (80%) reliability and low (60%) reliability. The study used 80% reliable for high and 60% reliable for low to replicate the levels from the Parasuraman and Miller (2004) experiment. The automation reliability relates to the accuracy of the automation assistant as it provides solutions and calculates options for participants. Calculated solutions were correct 80% of the time in the high reliability condition and correct 60% of the time in the low reliability condition.

PROCEDURE

Participants were randomly assigned into the etiquette and criticality conditions, the between-subject factors, as they began the study. They provided initial written consent prior to beginning. The participants completed a ten-trial practice block with 100% reliable automation to ensure they were familiar with the task and understood the system. The experimental trials consisted of four fifty-trial blocks. The experimental blocks consisted of two stage 2 automation blocks, one high reliability (80%) and one low reliability (60%) and two stage 3 automation blocks, one high reliability (80%) and one low reliability (60%). Blocks were counterbalanced evenly to prevent ordering

effects (see Appendix F for counterbalance sample size). Following each block, users answered four questions relating to trust in the automation (Lee and Moray, 1994) and six NASA-TLX questions (Hart and Staveland, 1988). After participants finished the four fifty-trial blocks, they completed a ten-question AICP questionnaire, and a five-question motivation/affect and criticality exit survey prior to completing the study.

RESULTS

An a priori power analysis determined a sample of 80 participants was required to detect a large effect size of $R^2 > .25$. A total of 208 participants completed the study. Four participants were dropped from the study due to extremely low performance. Data from 204 participants was used for analysis with the exception of Secondary Task Performance. Data from 48 participants was used for analysis on Secondary Task Performance due to extremely poor overall performance ($M=10.1\%$, $SD=18.8\%$) on this metric. All outliers that fell three or more standard deviations from the mean were removed. All data was checked for normality, homogeneity of variance, and independence where applicable. Numerous models violated the normality assumption and homoscedasticity. Thus, we used conservative estimates and Welch’s analysis when applicable.

Table 1
Descriptive statistics for participant demographics, CPRS scores, and AICP scores.

Measure	Good Etiquette Condition				Bad Etiquette Condition			
	Male (n=34)		Female (n=77)		Male (n=14)		Female (n=79)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age	18.8	1.1	18.4	.69	18.5	.75	18.3	.62
CPRS	64.5	5.1	65.0	5.7	66.2	4.5	65.5	5.3
AICP	46.3	6.4	45.7	5.8	48.3	6.4	46.1	5.8

Note. Higher CPRS and AICP scores reflect higher automation complacency potential.

Table 2
Descriptive Statistics for Etiquette and Criticality Conditions on Dependent Variables

Measure	Good Etiquette Condition				Bad Etiquette Condition			
	High-Criticality (Targeting Task) n=58		Low-Criticality (Taxi Task) n=49		High-Criticality (Targeting Task) n=49		Low-Criticality (Taxi Task) n=48	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Decision Accuracy (seconds)	65.6	.16	64.3	.18	65.3	.15	68.9	.14
Response Time	5654	1250	5699	1331	5757	1257	5562	1333
Total Trust	222	75	225	74	222	80	224	74
*Secondary Task Performance	28.1 n=13	22.1	29.9 n=14	23.1	32.1 n=10	22.1	38.6 n=11	29.4
Perceived Workload (Nasa-TLX Sum)	338	75.6	364	73.7	362	70	376	76

*Sample size reduced after filtering participants scoring below 10%

Full Multivariate Model

A 2 (etiquette: good and bad) x 2 (criticality: high and low) x 2 (reliability: high and low) x 2 (automation stage: stage 2 and stage 3) mixed MANOVA revealed significant main effects for etiquette ($F(4,817)=5.41, p<.001, \eta_p^2=.026$), automation level ($F(4,817)=53.1, p<.001, \eta_p^2=.206$), and automation reliability ($F(4,817)=4.922, p<.001,$

$\eta_p^2=.024$). Additionally, the full model revealed significant two-way interactions between etiquette and criticality ($F(4,817)=7.71, p<.001, \eta_p^2=.037$) and automation level and reliability ($F(4,817)=4.52, p<.001, \eta_p^2=.022$). We conducted follow-up analysis for each dependent variable to test our hypotheses.

Hypothesis 1: Good automation etiquette will produce better performance than bad automation etiquette.

Decision Accuracy, Response Time, and Secondary Task Performance were the three metrics used to measure experimental performance. Decision Accuracy represents primary task performance and was the main measure of interest. Decision Accuracy data was averaged for each fifty-trial block and is represented as a percentage of correctly answered trials (e.g., Decision Accuracy of .70 equates to 70% correct of 50 trials or 35/50 correct trials). Data from four participants was removed due to overall Decision Accuracy of less than 15%, indicating they did not understand the task or did not attempt to perform at the task. Response time is depicted in milliseconds and represents the average response time across one fifty-trial block. Secondary task Performance is represented as percent correct. Secondary Task Performance was aggregated as the average of correct responses across one fifty-trial block. Participants' overall Secondary Task Performance was poor. After filtering all participants with less than 10% overall performance, 156 were eliminated, giving us a sample size of 48 when we conducted our analysis of Secondary Task Performance.

Decision Accuracy.

A 2 (etiquette: good and bad) x 2 (criticality: high and low) x 2 (reliability: high and low) x 2 (automation stage: stage 2 and stage 3) mixed ANOVA on Decision Accuracy was significant ($F(15,820)=10.6, p<.001, \eta_p^2=.162$). The model indicated significant main effects for etiquette ($F(1,820)=5.56, p<.00, \eta_p^2=.007$) and automation level ($F(1,820)=135.6, p<.001, \eta_p^2=.142$). Additionally, the model revealed a significant two-way interaction between etiquette and criticality ($F(1,820)=6.68, p<.001, \eta_p^2=.008$). We conducted post hoc analysis on our constructs of interest to test specific effects.

Etiquette and Criticality. A 2 (etiquette: good and bad) x 2 (criticality: high and low) factorial ANOVA on Decision Accuracy was statistically significant ($F(3,820)=2.92, p<.05, R^2=.01$). The model revealed a significant interaction effect between etiquette and criticality ($F(1,820)=4.52, p<.05, \eta_p^2=.005$). The interaction indicated etiquette had an impact on Decision Accuracy in the low-criticality taxi task but not in the high-criticality targeting task. Significant simple effects of etiquette ($F(1,377)=7.42, p<.01, R^2=.02$) indicate participants in the bad etiquette taxi condition scored higher ($M=.689, SD=.14$) than participants in the good etiquette taxi condition ($M=.643, SD=.18$; Figure 5). There were no differences between the etiquette conditions in the high-criticality targeting task.

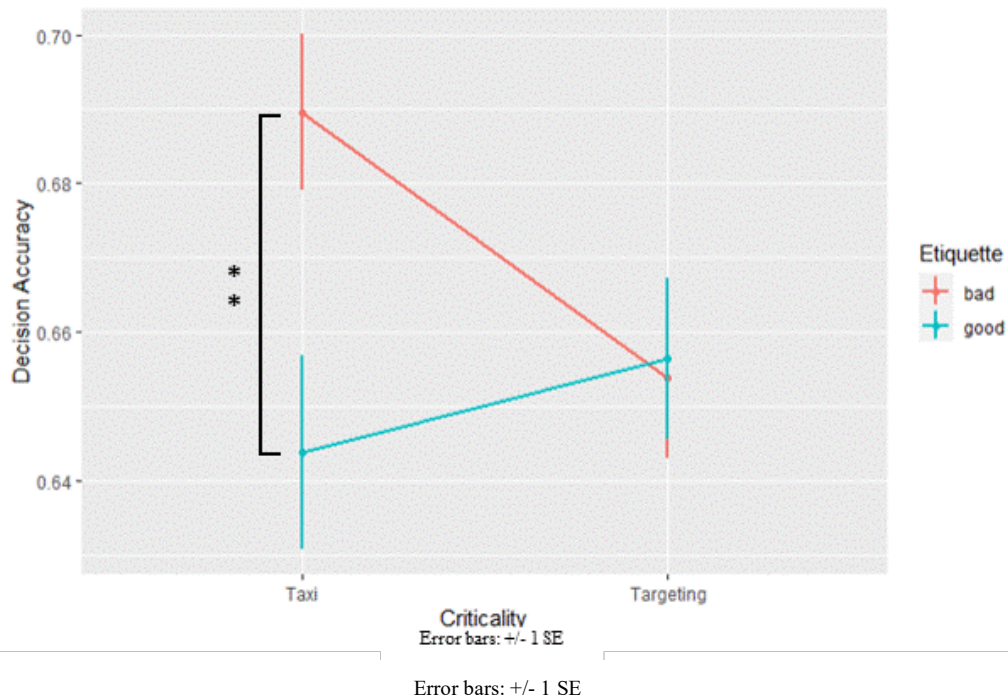


Figure 5. Significant interaction between etiquette and criticality. Bad etiquette participants in in the taxi task scored higher than good etiquette taxi participants.

Etiquette, Automation Stage, and Automation Reliability. A 2 (etiquette: good and bad) x 2 (automation stage: stage 2 and stage 3) x 2 (reliability: high and low) factorial ANOVA was statistically significant ($F(7,816)=22.26, p<.001, R^2=.163$). The analysis revealed a main effect for automation stage ($F(1,816)=143.2, p<.001, \eta_p^2=.147$), a main effect for reliability ($F(1,816)=5.57, p<.02, \eta_p^2=.005$), and a main effect for etiquette ($F(1,816)=3.98, p<.05, \eta_p^2=.004$). The automation stage main effect indicated the stage 3 decision automation performance was significantly higher ($M=.724, SD=.11$) than the stage 2 information analysis automation ($M=.600, SD=.17$; Figure 6). A post hoc test revealed that in the stage 3 automation condition, the effects of etiquette were significant ($F(1,392)=4, p<.03, R^2=.01$) with bad etiquette participants scoring higher ($M=.736,$

$SD=.09$) than good etiquette participants ($M=.711$, $SD=.13$) (Figure 6, stage 3 graphs on right side). The main effect for reliability revealed the high reliability performance was significantly higher ($M=.672$, $SD=.16$) than low reliability performance ($M=.648$, $SD=.15$) (Figure 7).

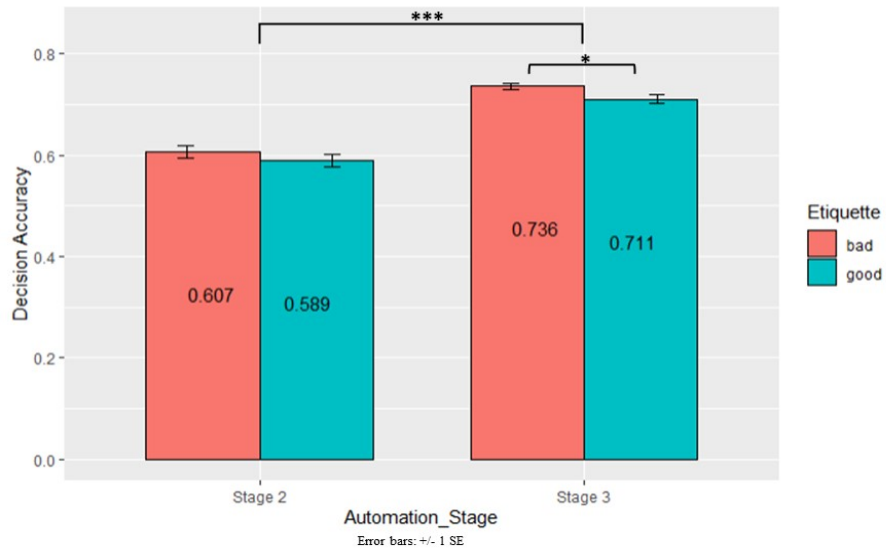


Figure 6. Significant differences of performance between stage 2 and stage 3 automation and significant differences in stage 3 performance between etiquette conditions.

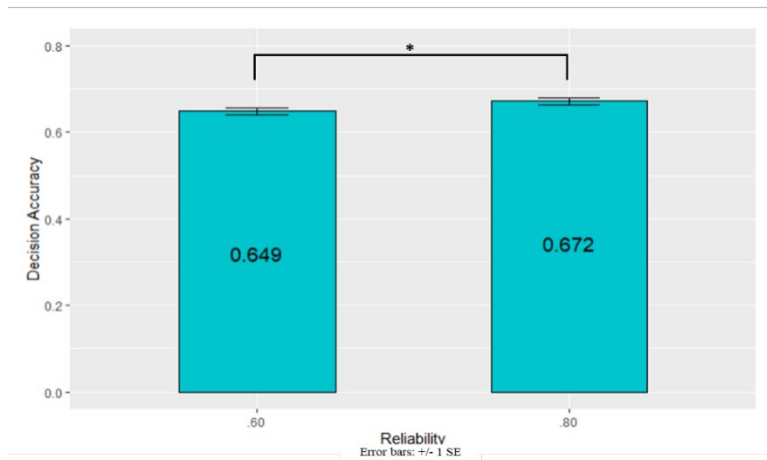


Figure 7. Significant differences of performance between stage 2 and stage 3 automation and significant differences in stage 3 performance between etiquette conditions.

Response Time.

A 2 (etiquette: good and bad) x 2 (criticality: high and low) x 2 (stage of automation: stage 2 and stage 3) x 2 (automation reliability: low and high) factorial ANOVA on Response Time was statistically significant ($F(5,818)=12.7, p<.001, R^2=.07$). The model revealed a main effect for automation stage ($F(1,822)=58.3, p<.001, \eta_p^2=.07$). The stage 3 decision automation performance was significantly faster ($M=5338, SD=1328$) than the stage 2 information automation ($M=6001, SD=1160$). In all experimental conditions and sub-conditions, stage 3 automation performance was better (i.e., quicker response time) than stage 2 automation performance (Figure 8). Additionally, within the stage 3 automation taxi condition, there was a significant ($F(1,412)=, p<.05, R^2=.009$) difference between reliability conditions – high reliability performed better ($M=5207, SD=1287$) than low reliability ($M=5466, SD=1358$) (Figure

9). There were no differences in response time between any conditions or sub-conditions of etiquette or criticality. ANCOVAs using AICP and CPRS as covariates revealed no significant findings on Response Time.

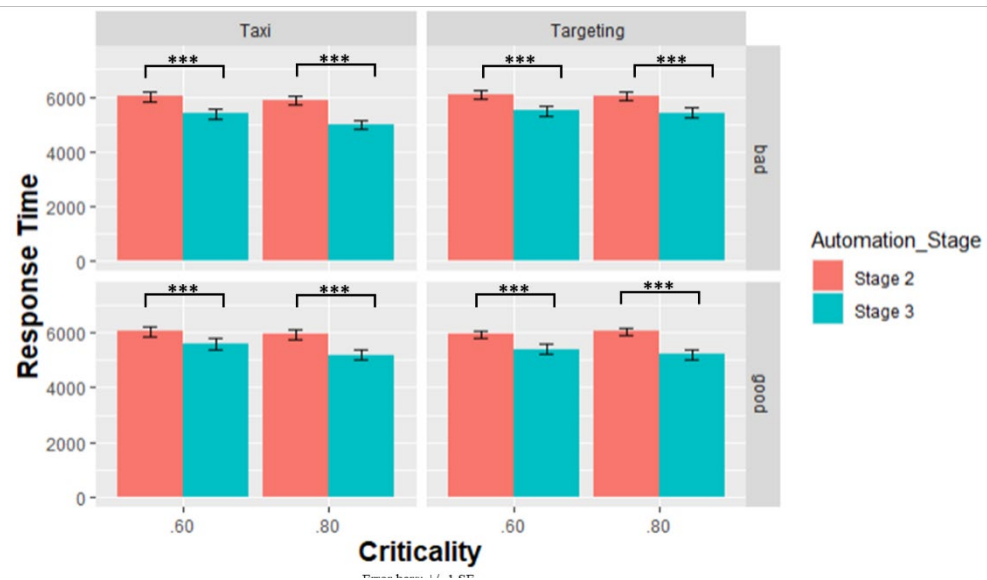


Figure 8. Stage 3 Automation performance significantly faster in all conditions and sub-conditions.

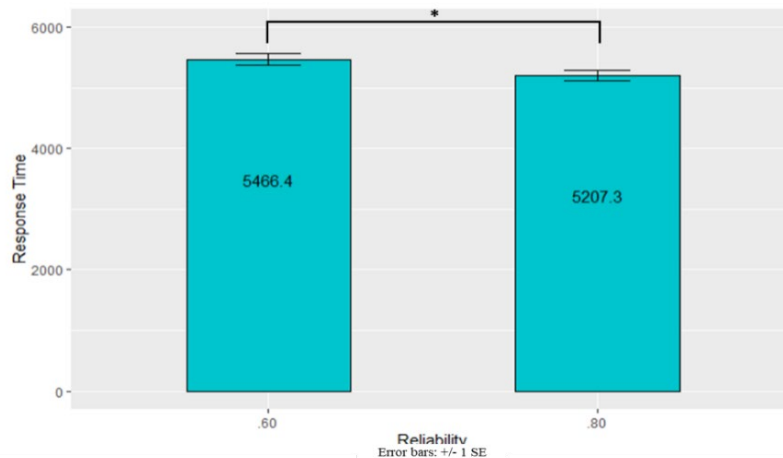


Figure 9. High reliability faster response time than low reliability in the Stage 3 Automation taxi condition.

Secondary Task Performance

A 2 (etiquette: good and bad) x 2 (criticality: high and low) x 2 (reliability: high and low) x 2 (automation stage: stage 2 and stage 3) mixed ANOVA revealed no main or interaction effects. We conducted additional analysis to test our specific hypothesis.

Etiquette and Criticality. A 2 (etiquette: good and bad) x 2 (criticality: high and low) factorial ANOVA on Secondary Task Performance was statistically significant ($F(3,189)=3.09, p<.03, R^2=.05$). There was a significant main effect for criticality ($F(1,189)=6.76, p<.01, \eta_p^2=.008$). The low-criticality taxi participants performed better ($M=.416, SD=.25$) than the high-criticality targeting participants ($M=.372, SD=.20$). This main effect was qualified by a significant interaction effect between etiquette and criticality ($F(1,189)=5.68, p<.03, \eta_p^2=.03$). The interaction shows there were no differences of Secondary Task Performance between the etiquette conditions in the high-criticality targeting task but there were significant differences between the etiquette conditions in the low-criticality taxi task. Simple effects of etiquette ($F(1,99)=5.823, p<.03, R^2=.05$) in the taxi task confirm that the bad etiquette taxi condition scored higher ($M=.482, SD=.27$) than the good etiquette taxi condition ($M=.363, SD=.22$) (Figure 10). There were no significant effects of automation stage or reliability on Secondary Task Performance in any conditions or sub-conditions.

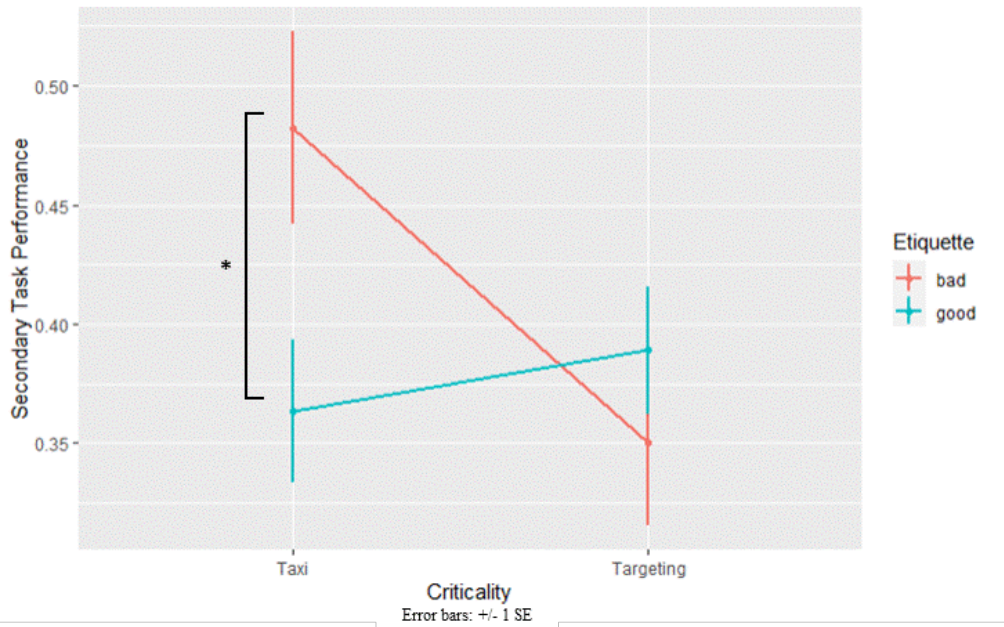


Figure 10. Significant interaction. Bad etiquette participants did better than good etiquette participants on Secondary Task Performance in the low-criticality taxi condition.

Hypothesis 2: Good automation etiquette will produce higher trust than bad automation etiquette.

Trust was measured using the Lee and Moray four-question trust questionnaire. Trust measures were collected a total of four times at the end of each fifty-trial experimental block. Each of the four trust questions was answered on a scale of 10-100 with higher numbers indicating higher trust. We aggregated the responses of each trial block to determine an overall trust measure and followed up all analysis with post hoc test on specific trust questions which measure different constructs (i.e., trust, automation reliance, etc.).

A 2 (etiquette: good and bad) x 2 (criticality: high and low) x 2 (reliability: high and low) x 2(automation stage: stage 2 and stage 3) mixed MANOVA revealed main effects for automation stage ($F(4,805)=9.93, p<.001, R^2=.05$) and automation reliability ($F(4,805)=2.79, p<.03, R^2=.014$). Therefore, we conducted follow up analysis on each of our variables.

Trust, Etiquette, and Criticality. The only significant effect of etiquette on trust was for one of the four questions on the trust questionnaire. Question 3 – “to what extent are you self-confident that you could successfully perform without the automation aid in this scenario?” Participants in the good etiquette condition trusted the automation significantly more ($M=66.7, SD=22$) than participants in the bad etiquette condition ($M=62.8, SD=21; F(1,822)=6.31, p<.02, R^2=.008$; Figure 11). There were no effects of criticality on trust.

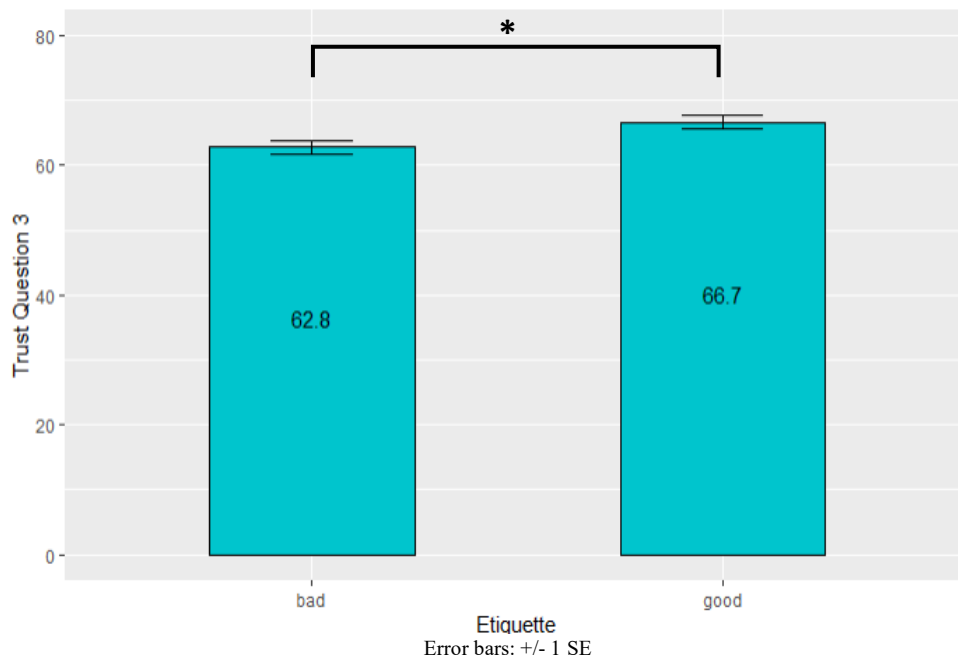


Figure 11. Higher trust with good etiquette on trust question 3 – “to what extent are you self-confident that you could successfully perform without the automation aid in this scenario?”

Trust, Reliability, and Automation Level. A 2 (automation stage: stage 2 and stage 3) x 2 (reliability: low and high) ANOVA on Trust was statistically significant ($F(3,820)=17.23, p<.001, R^2=.06$). The model reveals a main effect for automation stage ($F(1,820)= 32.5, p<.001, \eta_p^2=.04$). The mean trust for stage 3 automation was higher ($M=238, SD=75$) than the mean trust for stage 2 automation ($M=209, SD=74$). There was a main effect for reliability ($F(1,820)= 11.8, p<.001, \eta_p^2=.01$). The high reliability condition was trusted significantly more ($M=232, SD=79$) than the low reliability condition ($M=214, SD=71$). There was significant interaction effect between reliability and automation stage ($F(1,820)= 7.6, p<.01, \eta_p^2=.009$). The interaction indicates the effect of stage of automation on trust depends on the reliability level. There were

significant differences of trust between stage 3 and stage 2 in both reliability conditions but the differences were larger in the high reliability condition ($F(1,409)=33.33, p<.001, \eta_p^2=.075$) than the low reliability condition ($F(1,411)= 4.67, p<.05, \eta_p^2=.01$; Figure 12).

Hypothesis 3: There will be main effects of etiquette on performance in stage 3 automation but not in stage 2 automation.

Performance and Stage of Automation. A one-way (etiquette: good and bad) ANOVA on Decision Accuracy within the stage 3 automation condition was significant ($F(1,392)=4.99, p<.03, R^2=.01$) while the same analysis across stage 2 automation was

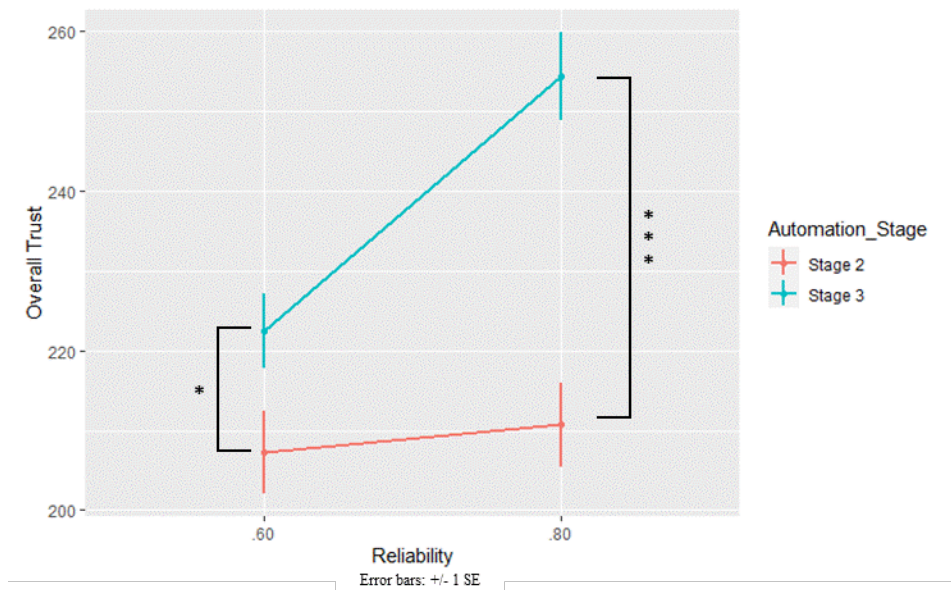


Figure 12. Interaction between reliability and stage of automation. Larger difference of stage of automation on trust in the high reliability condition than the low reliability condition.

not significant. The stage 3 bad etiquette automation performance was significantly higher ($M=.735, SD=.09$) than the stage 3 good etiquette performance ($M=.710, SD=.13$)

(Figure 13). This indicates there was a main effect of etiquette in the stage 3 automation but not in the stage 2 automation.

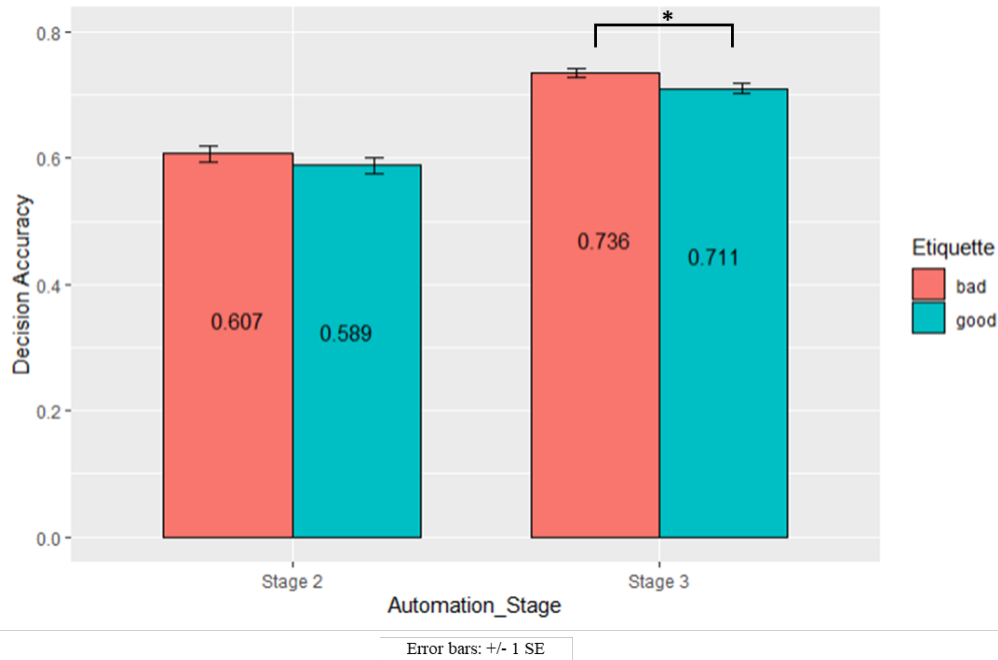


Figure 13. Impact of etiquette on performance significant in stage 3 automation but not stage 2.

Hypothesis 4: There will be main effects of etiquette on trust in stage 3 automation but not in stage 2 automation.

The only significant relationship that supported this hypothesis was on Trust Question 2 – “to what extent do you rely on (i.e., actually use) the automation aid in this scenario?” A 2 (etiquette: good and bad) x 2 (Automation Stage: stage 2 and stage 3) factorial ANOVA on Trust Question 2 revealed in the stage 3 automation, the bad etiquette participants relied on the automation significantly more ($M=59.8$, $SD=.09$) than the good etiquette participants ($M=53$, $SD=.13$; $F(1,412)=5.184$, $p<.03$, $R^2=.01$). In the stage 2 automation there were no differences on automation reliance (Figure 14).

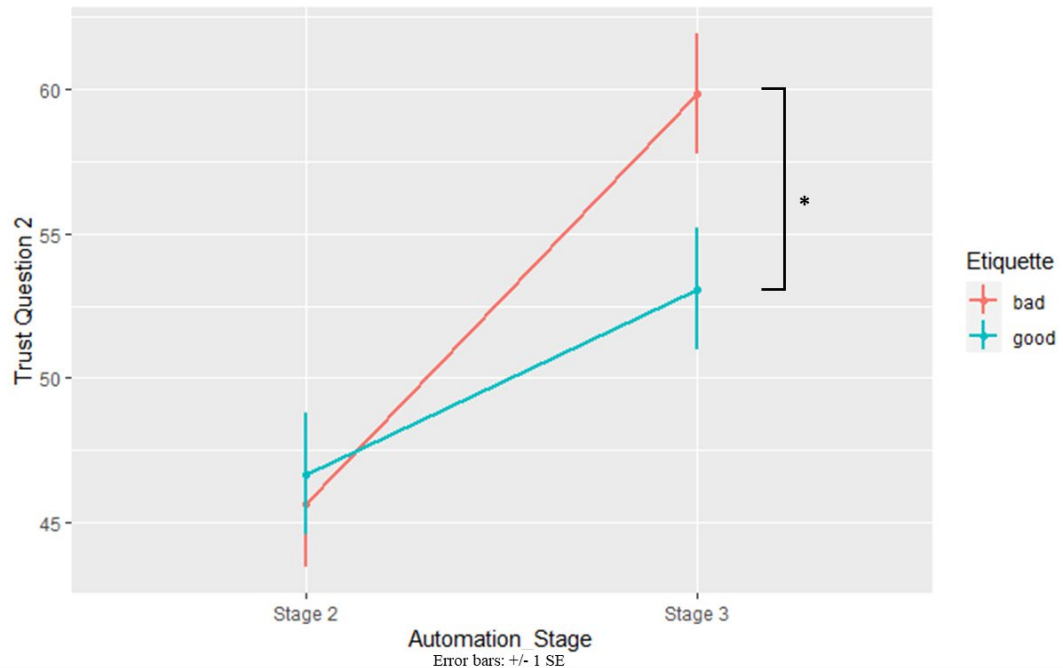


Figure 14. Impact of etiquette on trust significant in stage 3 automation but not stage 2. Trust Question 2 - “to what extent do you rely on (i.e., actually use) the automation aid in this scenario?”

Hypothesis 5: The bad etiquette automation will produce higher subjective workload than the good etiquette automation; and the high-criticality targeting task will produce higher subjective workload than the low-criticality taxi task.

Perceived workload was measured using the NASA-TLX following each fifty-trial block. Unless otherwise noted, data was analyzed using the TLX Raw calculation which takes the sum of all six TLX measures (i.e., the sum of the mental, physical, temporal, effort, performance, frustration).

A 2 (etiquette: good and bad) x 2 (criticality: high and low) x 2 (reliability: high and low) x 2 (automation stage: stage 2 and stage 3) mixed ANOVA on TLX was statistically significant.

($F(15,820)=3.53, p<.001, R^2=.06$). The main effect for etiquette was statistically significant ($F(1,820)=13.2, p<.001, \eta_p^2=.015$) and indicated the good etiquette participants perceived a lower workload ($M=352, SD=75.7$) than the bad etiquette participants ($M=369, SD=73.9$). This main effect was qualified by a significant interaction between etiquette and criticality ($F(1,820)=19.33, p<.001, \eta_p^2=.02$). This interaction reveals the impact etiquette has on perceived workload is qualified by task-criticality. In the targeting task, bad etiquette perceived workload was higher ($M=376, SD=70$) than good etiquette perceived workload ($M=338, SD=84$) while there were no effects of etiquette on perceived workload in the low-criticality taxi task (Figure 15).

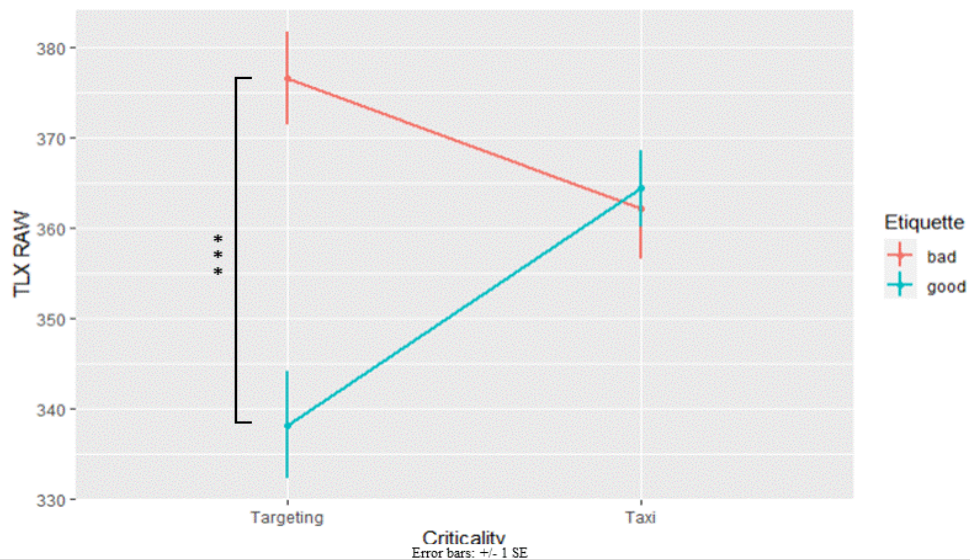


Figure 15. Perceived workload significant interaction between etiquette and criticality. Etiquette's effect on perceived workload depends on criticality with significant differences in the targeting condition but not the taxi condition.

There was a significant interaction effect between automation stage and automation reliability on perceived workload ($F(1,820)=8.94, p<.01, \eta_p^2=.01$). Simple effects of reliability indicate that in the low reliability condition, the stage 2 automation created a significantly ($F(1,413)=2.33, p<.03, R^2=.02$) higher perceived workload ($M=364, SD=77$) than the stage 3 low reliability automation ($M=345, SD=78$). There were no differences on perceived workload between the two stages of automation in the high reliability condition (Figure 16).

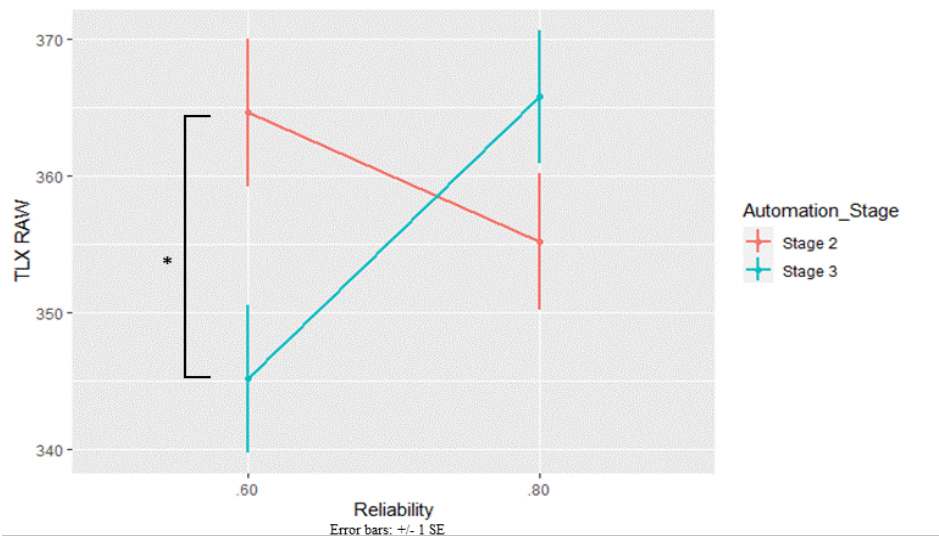


Figure 16. Perceived workload significant interaction between reliability and automation stage. Automation stage’s effect on perceived workload depends on reliability with significant differences in the low reliability but not the high reliability.

Additional Findings

The exit questionnaire measuring perceived task criticality (Appendix F) revealed significant findings. Cronbach’s alpha for the three-item scale was .51. A 2 (etiquette:

good and bad) x 2 (criticality: high and low) factorial ANOVA on perceived criticality (i.e., sum of three criticality questions) was statistically significant ($F(3,809)=12.7$, $p<.001$, $R^2=.05$). The high-criticality targeting task was perceived as more critical ($M=14.8$, $SD=2.86$) than the low-criticality taxi task ($M=13.9$, $SD=3.62$). Additionally, the bad etiquette condition was perceived as more critical ($M=14.9$, $SD=2.96$) than the good etiquette condition ($M=13.1$, $SD=3.44$). Simple effects of etiquette were significant in both criticality conditions but were more profound in the low-criticality taxi task ($F(1,388)=19.78$, $p<.001$, $R^2=.05$) than the high-criticality targeting task ($F(1,421)=4.9$, $p<.03$, $R^2=.011$) (Figure 17).

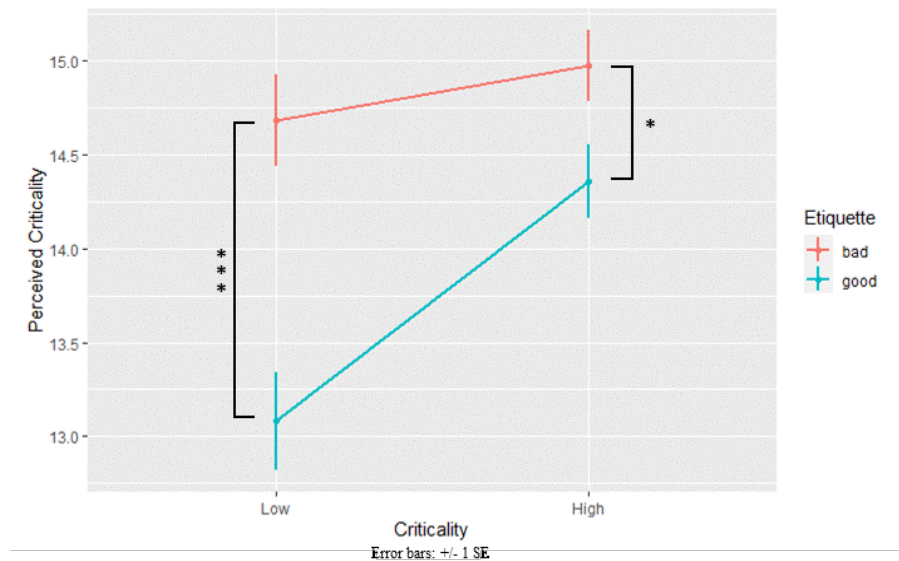


Figure 17. Perceived criticality within etiquette and criticality conditions. Perceived criticality calculated using sum of criticality/ importance subscale asking participants about task consequences, task importance, and task pressure.

DISCUSSION

The current study examined the impact of automation etiquette on performance and trust in non-personified technology. The goal of the study was to determine what role, if any, does automation etiquette play when users engage with technology that contains little to no anthropomorphization. We examined the interactions of etiquette, task criticality, stage of automation, and reliability. In contrast to some previous etiquette research, our task was unfamiliar to participants. This allowed us to control for previous experience, knowledge, and developed habits. Additionally, we introduced a secondary task and measured workload to expand previous research. We also implemented measures of automation complacency to account for individual differences within our study.

Hypothesis 1, which predicted better performance in the good automation etiquette condition, was not supported. Results indicate the bad etiquette automation produced better performance than the good etiquette automation in some conditions of the experiment. Participants in the bad etiquette condition outperformed participants in the good etiquette condition on both the primary and secondary experimental task in the low-criticality taxi condition but not the high-criticality targeting condition (Figure 5, Figure 10). The bad etiquette participants also outperformed good etiquette participants in the stage 3 automation conditions with no differences in the stage 2 automation conditions (Figure 6).

The superior performance in the bad etiquette taxi condition over the good etiquette taxi condition was a surprising finding. A few different theories could explain

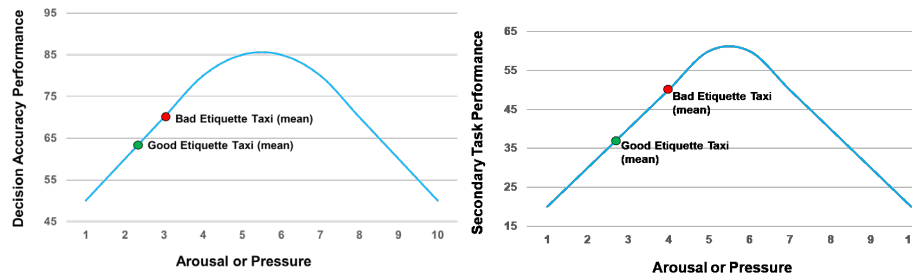


Figure 18. Low-criticality (taxi) task performance mapped onto Yerkes Dodson curve. Left graph represents Decision Accuracy (primary task); right graph these results. First, the Yerkes Dodson law of performance could have had an effect (Yerkes and Dodson, 1908). This law states that as arousal or pressure increases to a certain level, performance will increase. This relationship is usually qualified by a tipping point that represents a decline in performance when arousal or pressure becomes too high. It is possible that the bad etiquette condition elicited a higher level of arousal and pressure leading to superior performance over the good etiquette condition (Figure 18). The increased performance in the bad etiquette condition could be related to arousal, attentional resources, emotional affect, or a combination of all three.

We hypothesize that the bad etiquette taxi condition increased participants’ arousal/motivation toward the task. We believe there are four main explanations for an arousal/motivation increase. First, the bad etiquette condition provided better feedback to participants than the good etiquette condition. Although the different etiquette feedback messages were structured to contain the same information, the good etiquette messages contained extraneous words (e.g., for a correct response, the good etiquette said, “great

job, you matched the best pair, keep it up!” vs the bad etiquette’s “correct answer.”). It is possible the bad etiquette’s direct, no-nonsense feedback led to improved performance through increased arousal/motivation. A second explanation is that the bad etiquette condition was perceived as being more critical than the good etiquette condition (Figure 17). The perception of higher criticality in the bad etiquette condition would explain increased arousal/motivation. A third explanation is the bad etiquette condition reduced the amount of participant complacency while completing the task. The bad etiquette helped keep the participants “on their toes” which led to better performance. A final explanation for the arousal/motivation increase is that users felt a desire to “beat” the bad etiquette automation. Perhaps this condition imposed a sense of competition for the participants to outperform the bad etiquette automation.

Another factor contributing to the improved performance of some bad etiquette sub-conditions could be increased attentional resources. The bad etiquette taxi task could have elicited more goal-driven attentional resources (i.e., controlled or system 2), more stimulus-driven attentional resources (i.e., automatic or system 1), or both (Stanovich & West, 2000; Kahneman, 2011). If increased attentional resources was goal-driven, that would directly relate to the arousal/motivation increases described in the previous paragraph. Another explanation is the bad etiquette condition increased stimulus-driven attentional resources. It is likely that participants only fixated on the first chunk of the etiquette messages and the bright red coloration of the bad etiquette feedback. This would only take milliseconds to perceive and possibly could have created a quick spike in stimulus-driven attention leading to slightly better performance (Figure 19). A final

influential factor that could explain bad etiquette's superior performance is positive and negative affect. Perhaps the negative affect toward the bad etiquette messages was more powerful than the neutral or positive affect toward the good etiquette messages, resulting in better performance.

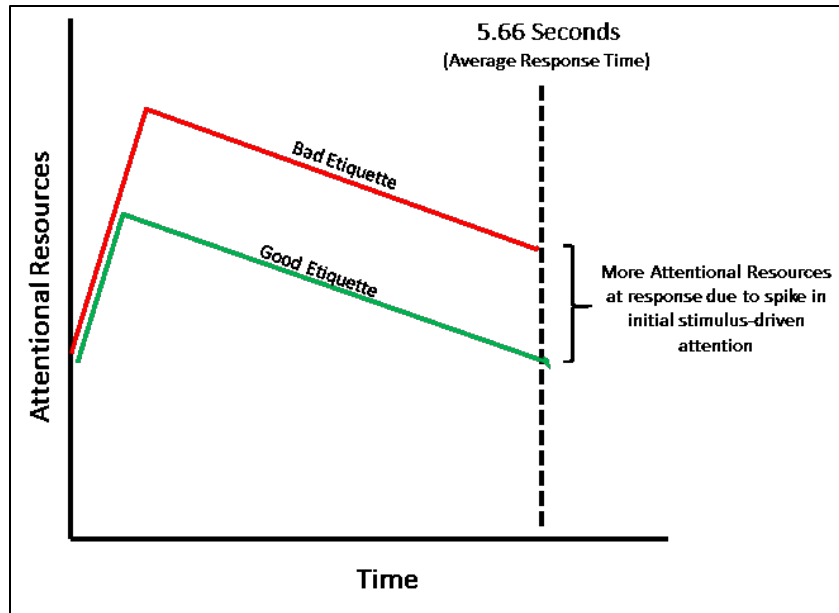


Figure 19. Initial spike in stimulus-driven attention in bad etiquette condition led to higher attentional resources than the good etiquette condition when a response was selected.

The findings regarding improved performance in the stage 3 over stage 2 conditions (Figure 6, Figure 10) and improved performance in the high reliability over low reliability conditions (Figure 7, Figure 9) were expected and support previous research.

Hypothesis 2, higher trust with good automation etiquette, was partially supported. Although there were no differences in total trust between good and bad etiquette conditions, we found differences on specific trust questions and within sub-

conditions. The good etiquette condition participants were more confident that they could successfully perform without the automation aid in the scenario (Trust question 3) than the bad etiquette participants (Figure 11).

These trust findings could have important implications related to training with automation when overconfidence is common with new users. If three human-automation interaction conditions are met, etiquette could potentially be used to calibrate trust. These three conditions include (a) users are unfamiliar and untrained on a particular type of automation or task; (b) automation reliance should be high when users are inexperienced; and (c) overconfidence is likely to occur. If these three conditions exist, our findings indicate etiquette may be a helpful supplement to achieve optimal trust between users and automation. Specifically, keeping etiquette neutral or slightly rude may produce less overconfidence in users in the early stages of training and use. This could be particularly beneficial if users are executing a higher criticality task.

Hypothesis 3, there will be main effects of etiquette on performance in stage 3 automation but not in stage 2 automation, was supported. There were significant differences between the two etiquette conditions in stage 3 automation but not in stage 2 (figure 13). These findings potentially suggest a direct relationship with degree of automation and etiquette – as degree of automation increases, so does the impact of automation etiquette on performance, depending on the automation reliability level (Figure 20). Users may be more sensitive to automation etiquette when the automation is more advanced. Our experiment demonstrates this can be true in non-anthropomorphized automation. The design implications of this finding are simple – etiquette may warrant greater consideration in more complex automation.

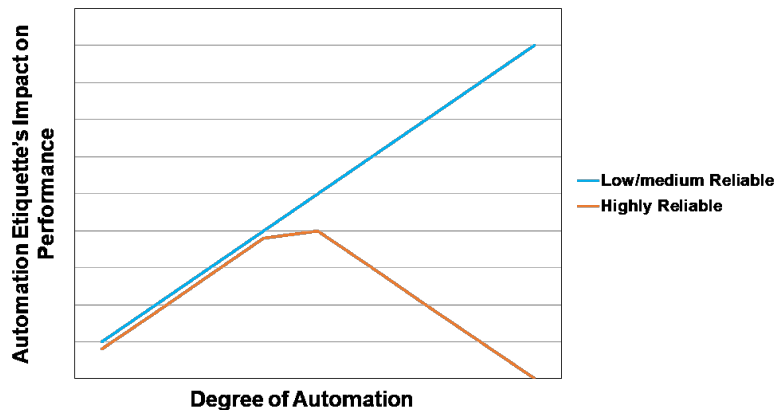


Figure 20. Direct relationship with degree of automation and automation etiquette's impact on performance – qualified by the automation reliability level.

Hypothesis 4, there will be main effects of etiquette on trust in stage 3 automation but not in stage 2 automation, was partially supported. There were significant differences on trust question 2, how much the user relied on the automation. In the stage 3 automation, the users reported that they relied on the bad etiquette automation more than

the good etiquette automation; however, etiquette had no effect in the stage 2 automation (Figure 14). Participants were more likely to use and rely on the automation in the stage 3 conditions and demonstrated more sensitivity to etiquette in these conditions. The key insight here is etiquette may have different impacts on user trust as automation complexity increases or decreases. An important design implication from these findings is to implement different etiquette models into user test.

Hypothesis 5, the bad etiquette automation will produce higher subjective workload than the good etiquette automation; and the high-criticality targeting task will produce higher subjective workload than the low-criticality taxi task, was supported. Participants in the bad etiquette condition reported a higher workload than participants in the good etiquette condition. Additionally, participants in the high-criticality task reported higher workload than those in the low-criticality task. This relationship was qualified by an interaction –bad etiquette workload was higher than good etiquette workload in the targeting task but not the taxi task (Figure 15). These findings indicate that there could be tradeoff to deliberately using bad or neutral automation etiquette to improve performance or help calibrate trust. Bad or neutral etiquette may incur workload cost. Our findings of higher workload with stage 2 over stage 3 automation were expected and support previous research (Figure 16).

We used an exit questionnaire to measure participants' perceptions about the criticality of the task. We asked participants to rate how high the consequences of poor performance were, how important it was to complete the task correctly with minimal errors, and how much pressure the participants felt trying to perform the task. Participants

in the high-criticality targeting task thought their task was more critical than participants in the low-criticality taxi task. Additionally, the bad etiquette conditions were perceived as more critical than the good etiquette conditions in both the taxi and targeting task (Figure 17). The results support the conclusion that our attempt to manipulate task criticality through domain, instructions, and context was effective. This could provide a positive contribution to the methodology of psychological research. Task criticality can be a sensitive manipulation because of the ethical guidelines relating to undue stress or discomfort in experimental settings. The use of the military vs taxi construct proved effective at increasing task criticality without jeopardizing participant well-being. The methods we used for manipulating criticality would be especially beneficial in lower-fidelity, non-simulator environments.

Limitations and Future Direction

Future research should address and expand on the limitations of this study. First, the study introduced different stages of automation and reliability conditions to the experiment. Although this was beneficial to our study, these manipulations may have confounded the true impact of etiquette and criticality. Future research could isolate etiquette and criticality to establish more direct causation. Second, future research should add better measures of affect/motivation to their studies. We measured affect/motivation with a two-question scale derived from Ryan and Deci (2000). Researchers should implement a higher reliability and more comprehensive affect scale. Additionally, we also only used this scale at the end of the experiment. Researchers could take repeated

measurements after each trial block to understand how participants' perceptions change over time.

Third, our etiquette was manipulated through after-response feedback. This is arguably much less powerful than before-response commands (i.e., telling the user what to do instead of how they did). Future experiments should use before-response commands/actions to manipulate etiquette. Fourth, our study only included two stages of automation. Future research should introduce additional stages/levels of automation (Sheridan and Verplank, 1978). Our automation possessed level 3 automation, which narrows the selection down to a few. Level 10 automation decides everything, ignores the human, and acts autonomously. Additional studies could explore etiquette's relationship with higher levels/degrees of automation. Fifth, we used two levels of etiquette that are generally classified as extremely polite (good etiquette) and rude (bad etiquette). More extreme etiquette manipulations should be used to see if worse etiquette might elevate performance on the Yerkes-Dodson Curve. Last, our etiquette contained both goal-driven attentional cues (i.e., through the written text in the messages) and stimulus-driven attentional cues (i.e., through the coloration of the message). Is it possible that etiquette could be manipulated with only stimulus-driven cues (i.e., using only the red and green colors)? Further research could help determine this. Overall, future research should examine additional domains, tasks, etiquette delivery mechanisms, and etiquette scales coupled with varied degrees of automation to better understand etiquette's role in human-automation interaction.

Conclusion

When designing this experiment, we hoped to answer a few critical questions relating to etiquette's role in human-automation interaction. First, at which stage/level/degree of automation does etiquette become important and how do user's etiquette expectations differ as automation complexity increases? We found evidence that etiquette mattered in stage 3 (decision) but not in stage 2 (information analysis) automation. We believe our experiment provided support for a direct relationship between the impact of etiquette and automation complexity. Second, we hoped to find out if etiquette can be systematically scaled to calibrate user trust? Our study provided evidence that certain aspects of trust can be targeted with etiquette. Automation reliance (Trust question 2) and user confidence (Trust question 3) can be sensitive to etiquette. Specifically, bad etiquette can increase user reliance and decrease user confidence in automation; both can be positive attributes in many situations.

Third, we aspired to determine if automation etiquette matters less in critical, stress-inducing task. Our study found mixed results. Etiquette did matter less in higher criticality task when looking at task performance. However, etiquette mattered more in higher criticality task when looking at user workload. Etiquette did not matter more in higher criticality task for user trust. Last, we hoped to establish how multitasking situations and workload are impacted by etiquette and how could designers systematically alter etiquette in known multitasking environments. Our study demonstrated that secondary task performance can actually improve with bad etiquette. However, this improvement is not without a cost. The bad etiquette condition imposed a

higher subjective workload on participants than the good etiquette. Perhaps in times of increased activity or crises, etiquette alterations could help users manage multiple tasks.

This study aimed to improve the understanding of automation etiquette on performance and trust. The impact and prevalence of human-automation interaction will continue to occupy an increasing role in society and will only become more important as technology advances at an exponential rate. The relationship of automation etiquette and task criticality on human-automation interaction should continue to be explored.

APPENDICES

Appendix A





Automation Induced Complacency Scale adopted from Merritt et al. (2019). Scale rated on a 7-point Likert scale where 1=strongly disagree and 7=strongly agree.

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

Appendix B



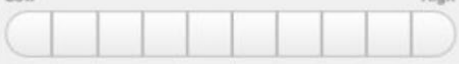



Trust measure adopted from Lee and Moray (1994)

Please answer the questions below about the computer aid (automation) by clicking on the scale:

<p>To what extent did you trust (i.e. believe in the accuracy of) the automation aid in this scenario?</p> <p>Not at all Extremely</p> 	<p>To what extent did you rely on (i.e. actually use) the automation aid in this scenario?</p> <p>Not at all Extremely</p> 
<p>To what extent were you self-confident that you could successfully perform without the automation aid in this scenario?</p> <p>Not at all Extremely</p> 	<p>To what extent do you think the automation improved your performance in this scenario compared to performance without the automation?</p> <p>Not at all Extremely</p> 

Appendix C

NASA-TLX adopted from Hart and Staveland (1988).

<p>How mentally demanding was the task?</p> <p>Low High</p> 	<p>How successful were you in accomplishing what you were asked to do?</p> <p>Not very Very</p> 
<p>How physically demanding was the task?</p> <p>Low High</p> 	<p>How hard did you have to work to accomplish your level of performance?</p> <p>Low High</p> 
<p>How hurried or rushed was the pace of the task?</p> <p>Not very Very</p> 	<p>How insecure, discouraged, irritated, stressed, and annoyed were you?</p> <p>Low High</p> 

Appendix D

Motivation, affect, and criticality questionnaire adopted from modified Ryan and Deci Inventory. Scale rated on a 7-point Likert scale where 1=strongly disagree and 7=strongly agree.

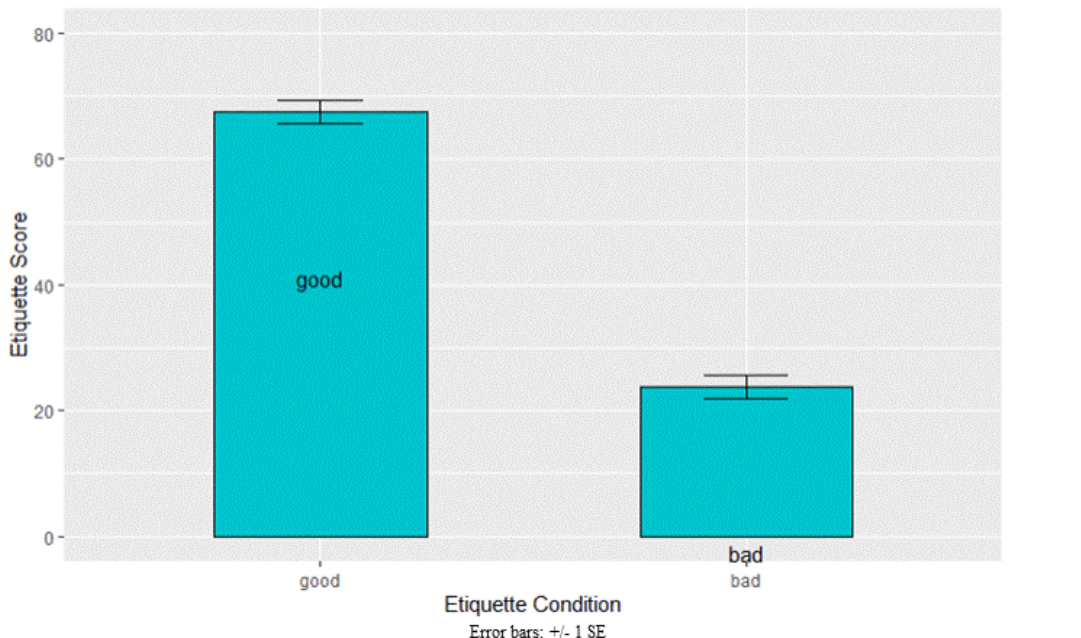
Subscale	Item
Motivation/Affect	I enjoyed working with this automation very much. It is an enjoyable activity
	I would be willing to use this automation again because it was beneficial to my score
Criticality/perceived importance	The consequences of poor performance on this task were high
	It was important that this task is completed correctly with minimal errors
	I felt pressure trying to perform well while completing this task

Appendix E

Etiquette Pilot Test

Results

A pilot test revealed the etiquette manipulations were effective. 13 participants who knew nothing about the study were recruited. A one way ANOVA comparing the two etiquette conditions of the pilot test was statistically significant ($F(1,24)=271.5, p<.0001, R^2=.92$). The good etiquette condition was perceived as more polite ($M=67.5, SD=6.9$) than the bad etiquette condition ($M=23.8, SD=6.5$). See graph below.



Pilot Test

The pilot test consisted of 21 etiquette questions where participants rated the rudeness or politeness of a message. The messages in the pilot test were nearly identical to messages used in the study. The only difference was a domain change where a sales scenario was used instead of a taxi or targeting scenario. Pilot scenario and sample question below.

Intro: Please read the background information below and answer the questions accordingly:

You are a telephone-based insurance salesperson who makes sales calls throughout the day. The company you work for provides a computer-based program that allows you to track your sales. Another function of

this computer-based program is the give you feedback after each sales call to inform you if you made the sale, did not make the sale, or ran out of time. Running out of time means you did not close the sale in the allotted 5-minute time frame imposed by the company.

2 The following statements consist of the feedback messages that the computer-based sales tracker provides you after each sale. For example, after a successful sale, the computer-based sales tracker would give you a message that says “Great job! You made the sale. Keep it up!” **Rate the politeness of each of the following messages from the computer-based sales tracker on a scale from 1 to 7 with 1 being very rude and 7 being extremely polite.**

3 *Great job! You made the sale. Keep it up!*

- 1 - Very rude (1)
- 2 - Moderately rude (2)
- 3 - Slightly rude (3)
- 4 - Neither rude nor polite (4)
- 5 - Somewhat polite (5)
- 6 - Moderately polite (6)
- 7 - Extremely polite (7)

Appendix F

Experimental Counterbalance Sample Size

Block #	Automation Stage	Reliability (.60)	Reliability (.80)
Block 1	Stage 2	49	62
	Stage 3	57	48
Block 2	Stage 2	58	56
	Stage 3	48	46
Block 3	Stage 2	54	46
	Stage 3	45	58
Block 4	Stage 2	43	42

*All numbers represent sample size for specific condition.

REFERENCES

- Brown, P., Levinson, S. C. (1987) *Politeness, Some universals in language usage*.
Cambridge, UK: Cambridge University Press.
- Hayes, C. C., Miller C. A. (2011) *Human-Computer Etiquette*. Boca Raton, FL:
Auerbach Publications. New York, NY: Farrar
- Hart, S.G., Staveland, L.E. (1988). Development of NASA-TLX (task load index):
Results of empirical and theoretical research. *Advances in Psychology*, 52, 3-4.
- Kahneman, D. (2011). *Thinking Fast and Slow*. New York, NY: Farrar, Straus and
Giroux
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to
automation. *International journal of human-computer studies*, 40(1), 153-184.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate
reliance. *Human Factors: The Journal of the Human Factors and Ergonomics
Society*, 46(1), 50-80.
- Merritt, S.M., Ako-Brew, A., Bryant, W.J., Staley, A., McKenna, M., Leone, A., &
Shirase, L. (2019). Automation-induced complacency potential: Development and
validation of a new scale. *Frontiers in Psychology*, 10(225).
- Miller, C., Wu, P., Funk, H. (2008) A Computational Approach to Etiquette:
Operationalizing Brown and Levinson's Politeness Model. *Smart Information
Flow Technologies*, July/August 2008, 28-35.

- Miller, C., Wu, P., Funk, H., Wilson, P., and Johnson, W.L. (2006). A computational approach to etiquette and politeness: initial test cases. *In proceedings of 2006 BRIMS Conference*, May 15-18, 2006, Baltimore, MD.
- Mosier, K., & Fischer, U. (2012). Impact of Automation, Task and Context Features on Pilots' Perception of Human-Automation Interaction. *In Proceedings of the HFES annual meeting*, 56 (1), 70-74.
- McAuley, E., Duncan, T., & Tammen, V. V. (1987). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, 60, 48-58.
- Nass, C. (2004). Etiquette Equality: Exhibitions and Expectations of Computer Politeness. *Communications of the ACM*, 47(4), 35-37.
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1), 81-103.
- Nass, C., Steuer, J., & Tauber, E. R. (1994, April). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 72-78). ACM.
- Pak, R., Rovira, E., Mclaughlin, A., Baldwin, N. (2016). Does domain of technology impact user trust: investigating trust in automation across different consumer-oriented domains in young adults, military, and older adults. *Theoretical Issues in Ergonomics Science*, 18(3), 199-220.

- Pak, R., McLaughlin, A. C., Leidheiser, W., & Rovira, E. (2016). The effect of individual differences in working memory in older adults on performance with different degrees of automated technology. *Ergonomics*, 1-15.
- Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, 47(4), 51-55.
- Parasuraman, R., Sheridan, T., Wickens, C. (2000). A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics*, 30(3), 286-297.
- Reeves, B. Nass, C. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge, UK: Cambridge University Press.
- 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.
- Rovira, E., Pak, R., McLaughlin, A. (2017). Effects of individual differences in working memory on performance and trust with various degrees of automation, *Theoretical Issues in Ergonomics Science*, (18)6, 573-591.
- Ryan, E., Deci, E. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology* (25), 54-67.
- Sheridan, R. M., Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. (Technical Report, Man-Machine Systems Laboratory, Department of Mechanical Engineering). Cambridge, MA: MIT Press.

- Spain, R. D., & Madhavan, P. (2009). The role of automation etiquette and pedigree in trust and dependence. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 53(4), 339-343.
- Stanovich, K. R., West, R. W. (2000). Individual Differences in reasoning: Implications for the Rationality Debate. *Behavior and Brain Sciences* 23, 645-65.
- Yang, E., Dorneich, M. (2018). Affect-Aware Adaptive Tutoring Based on Human-Automation Etiquette Strategies. *Human Factors*, 60 (4), 510-52
- Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18(5), 459–482.