The Effects of Mid-range Visual Anthropomorphism on Human Trust and Performance Using a Navigation-based Automated Decision Aid

A Dissertation
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Dara S. Gruber

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Barry Kudrowitz, Thomas Stoffregen

April 2018

Acknowledgements

Committee

Barry Kudrowitz, Thomas Stoffregen, Wilma Koutstaal, and Lana Yarosh

Design team

Henry Nahurski, Megan Peaslee, Jonathan Erbacher, Aleyse McNealy, Lauren Seymour, Max Pederschmidt, Ange Wang

Research assistants

Ashley Aune, Zachary Halfen, Eliza Schumer, Hanna Boleman, Ivan Trubetskoy

Statistics consultants

Lian Hortensius, Jennifer Teves

Editors

Harriet Gruber, Robert Gruber, Hali Gruber, Sarah Gruber, Denise Peaslee

Dedication

In loving memory of my incredible grandparents whose love and support continually guided and inspired me throughout this process.

Mary and John Spiropoulos

Betty and Bernie Gruber

R.I.P.

Abstract

The majority of us use personal assistant technology every day. From calendar alerts to fitness goal reminders, we have come to depend on this automation to provide us with information about our lives and help us to make "better" decisions. Research has been published on how to best represent recommender information to users, but not much has been done in the way of studying decision aids for low risk daily use. This research aims to explore how users of this technology trust computer-generated suggestions and how best to display those suggestions to optimize trust and favorable performance outcomes for continued use.

Table of Contents

List of Tables

List of Figures

Introduction, Purpose, and Aim

Problem Statement

The importance of developing trustworthy and socially acceptable personal assistive aids for everyday use is a tech industry priority. Numerous companies spend billions of dollars on research and development for creating useful automation to offload the burdens of daily life. Such automation can be perceived as creepy and/or invasive. These feelings are frequently associated with too much or too little of the following: Anthropomorphism or how human-like the automation seems, and transparency or knowledge of how the automation works. There simply is a need for a better understanding of how automation generates suggestions and guidance on how to humanize computer-generated aid. More specifically, the following research questions need to be addressed:

1. What are the past and present industry practices as well as research advancements in designing for appropriate use and trust of navigation decision aids?
2. Can level of anthropomorphism affect trust for different aid information reliability?
3. How do humans behaviorally respond to visually anthropomorphized aids in a navigational context?

The aforementioned questions are tested using different levels of anthropomorphic imagery paired with automated advice to see how visual representation influences trust and performance across varied information reliability conditions.

Objective

The objective of the current study is to expand upon the existing research on human trust in automated decision aids by evaluating the effects of mid-range anthropomorphism and information reliability on system trust and performance. Current research efforts in this space are in highly critical systems with high associated risk such as self-driving cars and unmanned aerial systems. It is equally important to explore systems used by millions of people to make seemingly non-trivial life decisions on a daily basis. Implications of this research are massive as machine learning and predictive technologies become more and more ubiquitous. Big tech companies strive to understand how users will understand, trust, act, and perform using technology where automation has increasingly more decision-making power.

Building Trust in Navigation Decision Aids

The aim of this chapter is to provide a chronological account of both trust in human automation interaction literature and in-vehicle routing and navigation systems. The section begins with a robust history of trust in automation research and serves as the main review of relevant trust literature for the project. Then, a product-focused discussion of in-vehicle routing and navigation systems leads to the current navigation decision aid technology used in this project. Showing a progression of products from a technological standpoint is useful for creating a deeper understanding of how the products work and what exactly changed in terms of function and capabilities. Routing and navigation products are presented in parallel to the chronological trust research review to help realize connections between theoretical work and the relevant technological products at that time. A comparison of this nature between product evolution and theoretical frameworks has the potential to encourage similar efforts to help bridge the divide between product research and development (R&D) and human related theories of interaction within a specific domain.

History of Trust in Human Automation Interaction

Function allocation. Affective constructs of human interaction with technology began with a theoretical assessment of function allocation by Jeremy Fitts (1951). Function allocation by definition is a design process that refers to the division of activities between humans and machines in a system. Fitts boldly proposed that man and machine are comparable. He created what is known today as “Fitts List,” a two-column list organized by properties, with one labeled “human” and the other labeled “machine.”

It compares the functions for which man is superior to machines to the functions for which machines are superior to man. For example, Fitts List tells us that men are flexible but not consistent; whereas machines are consistent but not flexible. The list allows designers to choose machines for functions for which they are best suited, and men for the functions for which they are best suited (Fitts, 1951).

Humans appear to surpass present-day machines in respect to the following:	
1. Ability to detect a small amount of visual or acoustic energy	7. Ability to respond quickly to control signals and to apply great force smoothly and precisely
2. Ability to perceive patterns of light or sound	8. Ability to perform repetitive, routine tasks
3. Ability to improvise and use flexible procedures	9. Ability to stare information briefly and then to erase it completely
4. Ability to store very large amounts of information for long periods and to recall relevant facts at the appropriate time	10. Ability to reason deductively, including computational ability
5. Ability to reason inductively	11. Ability to handle highly complex operations, i.e., to do many different things at once
6. Ability to exercise judgment	

Table 1. Fitts List. Adapted from "Human engineering for an effective air-navigation and traffic-control system" by Fitts, P. M.

About a decade later, a second argument was made stating that men and machines are complementary rather than comparable as proposed by Fitts in 1951 (Jordan, 1963). Jordan argued that little progress had been made in the area of allocating functions between human and machine because of this comparative way of thinking. He further criticized man-machine comparability by deconstructing Fitts List. Jordan pointed out the following: machines can be tools used to extend a human's abilities, machines can also be used to extend production methods utilized to replace the human completely,

responsibility can only be assigned to humans, and man degrades gracefully while machine does not.

Further analysis by Jordan (1963) revealed that both humans and machines have a physical environment, but humans also have a psychological environment that encourages motivation and other affective constructs such as trust. By designing systems where humans do less, challenge is eliminated, thus motivation is eliminated as well. Given that humans function best under an optimum level of difficulty, machines can and should be used as tools to bring the perceptual and motor requirements to optimum levels of human performance; hence the complementary relationship.

Sheridan and Verplank (1978) set the stage for the next development with a model of supervisory control, in which a human operator controlled a physical process through an intermediary computer. The Sheridan-Verplank Scale of Human-Machine Task Allocation (SVL) was the first real taxonomy of automation levels aimed at further organizing function allocation into distribution of tasks for various automation types.

Sheridan-Verplank 10 Levels of Human-Machine Function Allocation
1. The Human does all the planning, scheduling, optimizing, etc. and turns task over to computer for merely deterministic execution.
2. Computer provides options, but human chooses between them, plans the operations, and then turns task over to computer for execution.
3. Computer helps to determine options, and suggests one for use, which the human may or may not accept before turning task over to computer for execution.
4. Computer elects options and plans actions, which human may or may not approve, computer can reuse options suggested by human.
5. Computer selects action and carries it out if human approves.

6. Computer selects options, plans and actions and displays them in time for the human to intervene, and then carries them out in default if there is no human input.
7. Computer does entire task and informs human of what it has done.
8. Computer does entire task and informs human only if requested.
9. Computer does entire task and informs human if it believes the latter needs to know.
10. Computer performs entire task autonomously, ignoring the human supervisor who must completely trust the computer in all aspects of the decision-making.

Table 2. Levels of Automation. Adapted from *Human and Computer Control of Undersea Teleoperators,* by Sheridan, T. B., and Verplank, W., 1978.

In the early 1980s, Bainbridge (1983) wrote a paper addressing the ironies of automation, specifically, the claim of how automation can expand rather than eliminate problems for the human operator. Bainbridge argued that humans still need to be designed into the system even if they do not carry out the functions directly. As more functions are able to be automated, operators are taking on more of a monitor role. Bainbridge warned against operator skill deterioration and dissatisfaction for workers due to the shift to majority deskilled tasks. Ironically, humans are bad at vigilance tasks such as monitoring systems. In order to prevent human skill deterioration and dissatisfaction, Bainbridge suggested designing systems in which the operator is provided assistance (i.e., alarms) to help recognize problems in the system. Additionally, adequate training should be provided, and real system usage (i.e., drills) should be practiced to help form and keep accurate mental models of the system. Lastly, Bainbridge referenced Jordan's concept known as "graceful degradation" (Bainbridge, 1983). Graceful degradation refers to failing slowly, allowing the operator to notice the issue and address the problem (Jordan,

1963). Bainbridge believed that automation should fail gracefully just as a person completing a manual task may fail gracefully.

Interpersonal trust adaptations. Engineering psychology up until the late 1980s had been mostly focused on work process automation and ergonomic design. In her early work, Muir (1989) identified a large gap in research around how humans affectively relate to automation. Muir proposed that trust in machines could be understood by adapting existing theories and models of interpersonal trust. Muir used two main perspectives of trust developed by Barber (1983) and Rempel, Holmes, & Zanna (1985) to create a framework for the human-automation relationship.

Basis of expectation at different levels of expertise			
Dimension from Barber (1983)	Dynamic Dimension from Rempel et al. (1985)		
Expectation	Predictability (of acts)	Dependability (of dispositions)	Faith (in motives)
Persistence Natural physical Natural biological Moral social	Events conform to natural laws Human life has survived Human and computers act "decently"	Nature is lawful Human survival is lawful Human and computers are "good" and "decent" by nature	Natural laws are constant Human life will survive Human and computers will continue to be "good" and "decent" in the future
Technical competence	j's behavior is predictable	j has a dependable nature	j will continue to be dependable in the future
Fiduciary responsibility	j's behavior is consistently responsible	j has a responsible nature	j will continue to be responsible in the future

Table 3. Human Computer Trust. Adapted from *Operators' trust in and use of automatic controllers in a supervisory process control task,* by Muir, B. M., 1989.

With "j" signifying computer actions, expectations for this model are loosely defined as follows: persistence refers to the frequency of natural and moral social orders. Technical

competence can be thought of as the ability of the other agent to demonstrate expertise and perform accordingly. Fiduciary responsibility is the expectation that the other agent will have some moral obligation to prioritize the other agent's interest before their own.

Muir's main argument is that trust changes as a result of continued interactions between a human operator and the automated system. Trust in an automated system starts out as a function of the system's predictability and consistency. After moderate use, trust shifts to reflect the operator's attribution of the system's dependability. Over time, the operator is able to gather behavioral evidence of system performance which again alters trust as the operator can infer a broader set of attributions about the nature and motives of the automation. Eventually, the operator trust becomes an ability to project and predict their own use and belief in the system's future actions. This is what Muir refers to as faith in the automation.

Around the same time, Lee and Moray (1992) identified similar factors of trust claiming that performance, process, and purpose formed a foundation of human trust in a system. Performance includes the past and present operation of the machine including its reliability and predictability. Performance, more specifically, is the ability of the automation to achieve the human operator's goals through expertise and execution. This concept is highlighted by Sheridan (1992) who defined robustness as a foundation for trust in human-automation relationships. The operator will tend towards trusting automation that reliably achieves the set goals.

Table 4. Purpose, Process, Performance model of trust

	Barber (1983)	Rempel, Holmes and Zanna (1985)	Zuboff (1988)
Purpose	Fiduciary responsibility	Faith	Leap of faith
Process		Dependability	Understanding
Performance	Technical competent performance	Predictability	Trial and error experience
Foundation	Persistence of natural laws		

Table 4. Purpose, Process, Performance model of trust. Adapted from *Trust in automation: Integrating empirical evidence on factors that influence trust,* by Hoff, K. A, and Bashir, M., 2015.

Trust in automation models. Muir (1994) took the above adaptations of interpersonal trust perspectives and developed the first theoretical qualitative model for trust in automation. The model helped to further define the relationship between humans and machines while focusing on how operators can calibrate their trust in an automated system. Muir and Moray (1996) set out to empirically test if these interpersonal trust adaptations could extend to an applied automation setting. Through simulation, participants in the experiments were asked to optimize milk production by varying the balance between manual and automatic control. Most operators in the first study preferred manual control almost exclusively which created a ceiling effect and made the relationship between trust and automation use particularly hard to understand. The second experiment was arranged in a way to reduce manual control bias found in Study 1 and showed a strong positive relationship between operator trust and use of automatic mode. These results supported the first dimension of the model of trust in automation where competency and responsibility expectations of the system are the main source of trust variance for the operator. There is additional empirical evidence that supports the

model's second dimension. Results alluded to a shift among predictability, dependability, and faith over time which also impacted trust among participants. The data showed that faith may be a better predictor of automation usage early on rather than after continued use which was originally postulated separately by both Muir (1994) and Moray (Lee and Moray, 1992). Overall, findings from these experiments revealed that the same factors influence the development of trust for both interpersonal and automation relationships, but maybe not in the same order.

Riley (1994) created model of automation reliance that shows a different set of influencing factors derived from human psychology literature. Figure 1 shows a collection of factors of which theoretical relationships between them are displayed via dashed line.

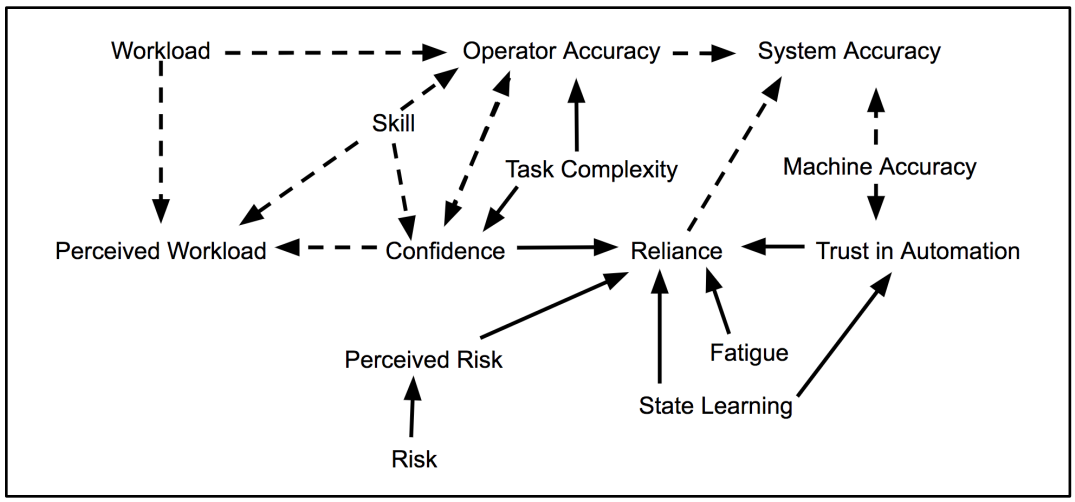


Figure 1. Riley's model of automation use. Adapted from *A theory of operator reliance on automation*, by Riley, V., 1994.

Riley (1994) also chose to empirically investigate the validity of the relationships in his proposed model of automation use. Each lab-based experiment included a computer task where participants had to decide whether or not to rely on the automation

to perform a task. Aspects of workload, task uncertainty, and automation reliability were measured using a computer game task while risk-taking was assessed through a gambling task and series of questionnaires. Results indicated that automation reliability, task uncertainty, and risk do influence an operator's decision to use automation. Another outcome was empirical reasoning to separate trust in automation from uncertainty of automation states in order to better understand the dynamic nature of how humans use an automation. However, results were unable to show a relationship between workload and automation use. Interestingly, Riley found that people are fairly bad at assessing their own expertise which is thought to impact automation use.

Expertise and self-confidence were further evaluated by Kantowitz, Hanowski, and Kantowitz (1997) who measured the relationship of trust and self-confidence to driver acceptance of ATIS. This study was designed to provide data that would aid the ATIS designers in selecting an appropriate level of system accuracy (reliability). Two traffic routes were given to participants: one was a "familiar network" and the other was a "new city." Information accuracy was set at 100%, 71%, or 43%, and the authors found that drivers with low self-confidence in unfamiliar settings would accept and trust less accurate information. Kantowitz et al. (1997) defined confidence as how sure a person was in decision-considering factors such as expertise. Other important trust-related findings from this study include: (a) trust in the system is higher when information is accurate, (b) trust could be recovered if accurate information is presented on subsequent links, and (c) inaccurate information decreases trust more in familiar settings. Essentially, when self-confidence is greater than trust, manual control is preferred (Kantowitz et al., 1997).

Other social constructs traditionally studied within social psychology literature, such as personality and emotion, began to drift into the human-computer interaction (HCI) realm. Research by Reeves and Nass (1996) illustrated that humans respond socially to technology, and that reactions to computers and reactions to human collaborators can be similar. For example, social psychologists predict that people with similar personality characteristics will be attracted to each other; this is known as the similarity attraction hypothesis coined by Nass and Lee in 2001. User acceptance of software is also predicted by this finding. Users tend to accept software more readily when it displays personality characteristics that are similar to their own. Beyond personality, the concept of affective computing suggests that human-computer interaction may significantly improve when computers can sense and respond to the user's emotional states (Picard & Picard, 1997).

Parasuraman and Riley (1997) further addressed theoretical, empirical, and analytical studies pertaining to human use, misuse, disuse, and abuse of automation technology. Understanding the factors associated with each of these aspects of human use of automation can lead to improved system design, effective training methods, and judicious policies and procedures involving automation use. Together, Parasuraman and Riley created a simple four-part classification system human-automation interaction: Use, Misuse, Disuse, and Abuse.

Automation Use refers to the voluntary activation or disengagement of automation by human operators. Trust, mental workload, and risk can influence automation use, but interactions between factors and large individual differences make prediction of automation use difficult (Parasuraman & Riley, 1997).

Automation Misuse refers to overreliance on automation that can result in failures of monitoring or decision biases. Factors affecting the monitoring of automation include workload, automation reliability and consistency, and the saliency of automation state indicators. Some design solutions include emergent displays, system feedback, and adaptive automation (Parasuraman & Riley, 1997).

Automation Disuse is the neglect or underutilization of automation and is commonly caused by alarms that activate falsely. This often occurs because the base rate of the condition to be detected is not considered in setting the trade-off between false alarms and omissions. According to sensitivity theory, omission is the failure to respond which is otherwise known as a miss. (Parasuraman & Riley, 1997).

Automation Abuse deals with the automation of functions by designers and implementation by managers without regard for consequences for human performance. It tends to define the operators' roles as by-products of the automation. Automation Abuse can also promote misuse and disuse of automation by human operators (Parasuraman & Riley, 1997).

Madsen and Gregor (2000) separated cognition from affect in a model of human computer trust (HCT). The emotionally driven, affect-based component plays a greater role in situations where the operator's perceived knowledge in the system and expertise is too low to make a cognitive decision. The operator's perceived knowledge in the system is defined as the cognition-based component of human-computer trust. While the model is void of procedural development and dynamic changes in trust, it incorporates well-studied and familiar factors such as competence, system reliability, and faith.

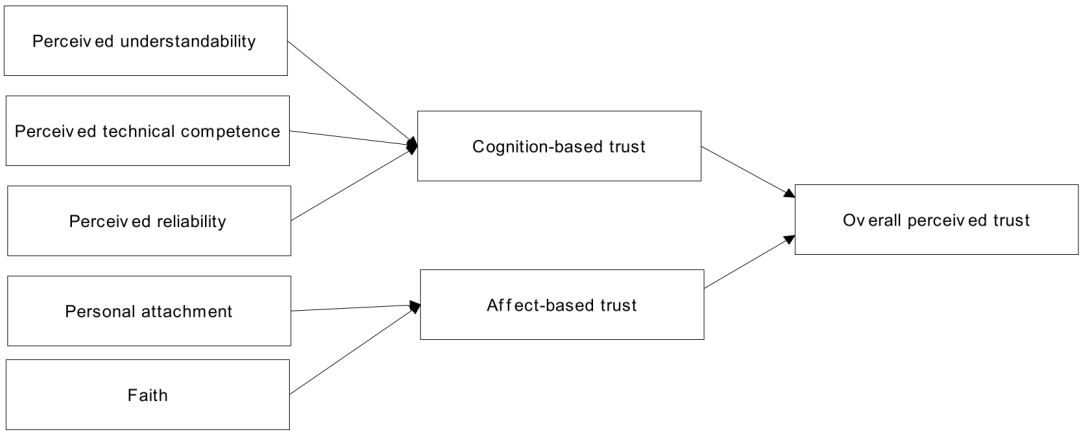


Figure 2. Cognition vs Affect Model for Trust. Adapted from *Measuring human-computer trust*, by Madsen, M., and Gregor, S., 2000.

Kelly et al. (2001) developed a subjective measure of trust and an accompanying model of trust in automation that aimed to provide a framework for automation design. It is important to note that the model was developed specific to the air traffic control domain. The model shows relationships among the comprehensive list of factors. According to Kelly et al., the operator's ability to understand the system, the competence of the system itself, and the operator's perceived self-confidence are the three main contributing factors to the overall level of trust in an automated system.

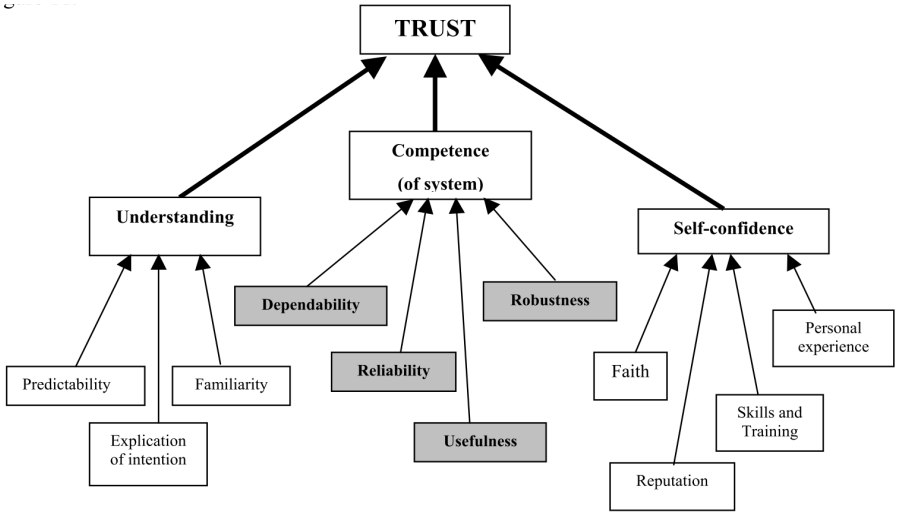


Figure 3. Framework for ATC trust. *Adapted from Principles and Guidelines for the Development of Trust in Future ATM Systems: A Literature Review,* by Kelly et al., 2001

Kelly et al. are some of the first to directly integrate skills and training into a model of trust in automation. However, the relationship to self-confidence is a bit assumptive in that it implies an exclusive impact through the operator's self image when it will likely also directly affect the operator's direct interactions with automation. Another uniqueness of this model lies in the distinctions made between automation performance and the operator's understanding of how the automation works. By viewing trust in this way, Kelly et al. imply that operator trust in automation is relative. In other words, performance may not matter as much in terms of system trust if an automated system is not designed in a way that can be easily understood by operators.

Applied trust models for designing automation. Perhaps the most widely applicable model comes from designing automation with the knowledge of appropriate function allocation that can help to mitigate many of the failures listed above. Refer back to the levels and types of automation proposed by Parasuraman, Sheridan, and Wickens (2000); a continuation of their work comes in the form of an iterative model that steps designers through a decision process (Figure 4).

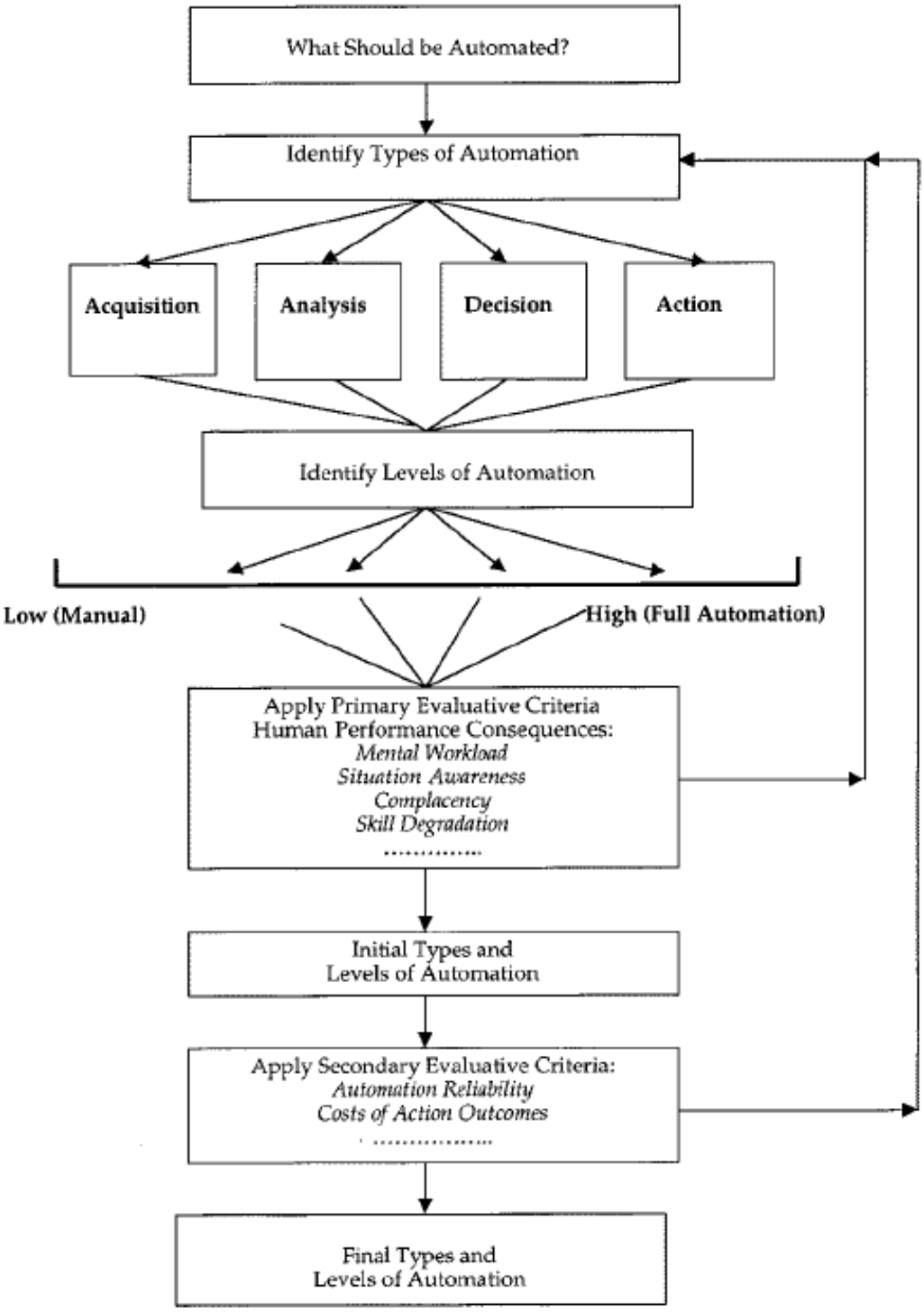


Figure 4: PSW model; designing for appropriate trust in automation. Adapted from *A model for types and levels of human interaction with automation,* by Parasuraman, R., Sheridan, T. B., and Wickens, C. D., 2000.

After identifying the type and level of automation for the given task, the next step is to consider the human performance consequences. Within the model, performance consequences are prioritized and split into two waves to encourage iterative process. The first wave is termed Primary Evaluative Criteria and includes factors such as workload, complacency, and situation awareness. Secondary Evaluative Criteria are to be considered after the Primary Evaluative Criteria. Reliability and Cost of Decision (i.e., action consequences and risk propensity) are the main factors in the second wave. These criteria will be discussed later on as manipulations in research evaluating trust in automation. Automation design is not an exact science and therefore Parasuraman, Sheridan and Wickens suggest the use of bounds in their framework that provide a systematic approach to iterative design.

Lee and See (2004) wrote the first notable review of trust in automation and created a conceptual model of the dynamic process which governs trust and reliance. According to Lee and See (2004), "three critical elements of this framework include: the closed-loop dynamics of trust and reliance, the importance of the context on trust and on mediating the effect of trust on reliance, and the role of information display on developing appropriate trust." Figure 5 expands on concepts exhibited in the following frameworks: Bisantz and Seong (2001) and Riley (1994) whose frameworks showed a dynamic interaction between the operator, context, and automation interface. Dzindolet, Pierce, Beck, and Dawe (2002) focused on the role of cognitive, social, and motivational processes that combine to influence reliance. Dzindolet et al. (2002) targeted changes in motivational processes along with cognitive processes to help explain reliance on automation. The general idea is that factors such as subjective workload, self-confidence,

effort to engage, and perceived risk are most pertinent to intent formation. When the operator has formed a solid use intent, other factors relating to time and configuration constraints may affect the initial reliance on automation. Ultimately, the decision to rely on automation is context dependent. Even so, factors affecting performance such as maintenance and weather, can cause inappropriate reliance.

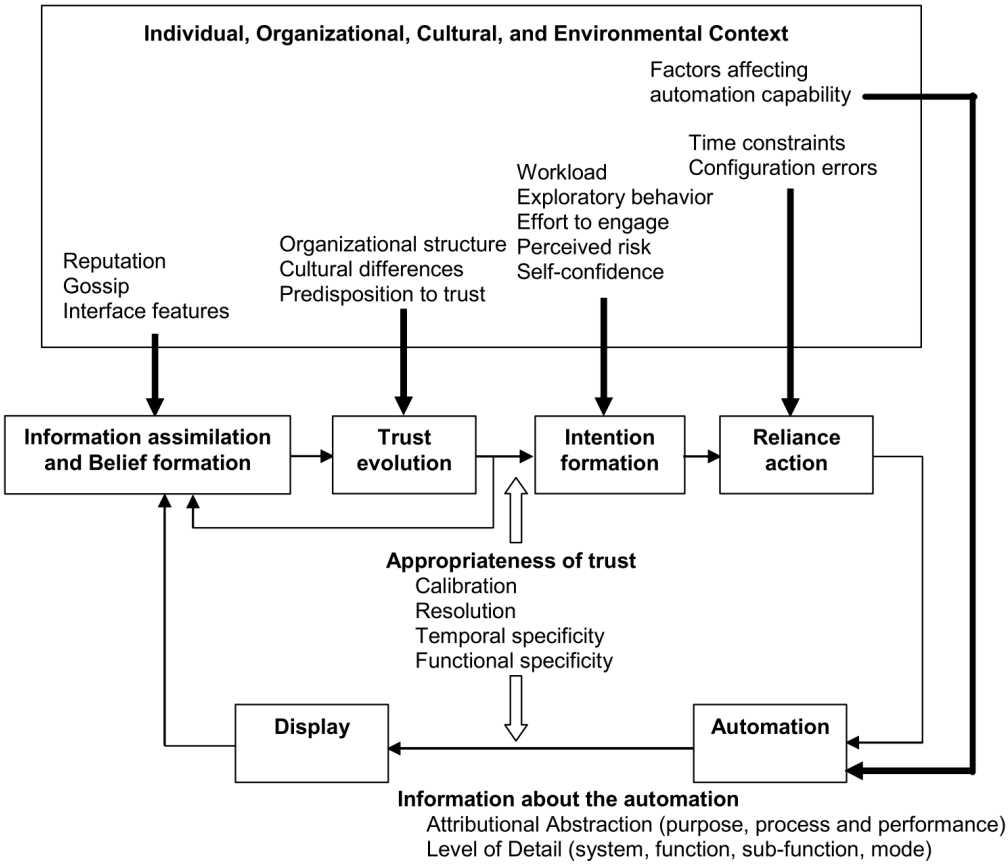


Figure 5. Dynamic interaction for trust in automation. Adapted from *The role of trust in automation reliance,* by Dzindolet, M. T. et al., 2003.

Design features affecting trust. Parasuraman and Miller (2004) studied the effects of automated information delivery in terms of etiquette, or machine politeness. They discovered that non-personified machine etiquette strongly affected human trust, usage decisions, and human-automation system performance. Parasuraman & Miller

(2004) hypothesized that good etiquette could compensate for poor reliability and result in increased usage decisions. They tested this in a 2x2 flight simulator study using low (45%) and high levels (70%) of reliability and low and high etiquette, and found that machine etiquette could strongly affect human trust, usage decisions, and human-automation system performance. Good automation etiquette enhanced diagnostic performance, regardless of automation reliability. The effects of etiquette overcame low reliability, and performance of good etiquette/low reliability was almost as good as high reliability/low etiquette (Parasuraman & Miller, 2004).

Good automation etiquette enhanced diagnostic performance, regardless of automation reliability; its effects overcame low reliability. This builds on etiquette-affected trust in automation work from Nass and Lee (2001). Nass and Lee found that in order to maximize liking and trust, designers should set parameters that create a personality that is consistent with the user and the content being presented. These parameters may include qualities such as words per minute and frequency range.

Other research in the early 2000s on computer etiquette suggested that recognizing and understanding the interactions of social and work context with the roles of the computer and human to specify acceptable behavior, can enhance human-computer interactions (Parasuraman & Miller, 2004). In highly critical domains, etiquette exhibited by non-personified machines and computer-based automation (such as facial expressions, speech, voice tones, and gestures) can overwhelmingly affect users' perceptions and the optimal usage of them (Parasuraman & Miller, 2004).

Machine etiquette can influence trust, which impacts use of automation that is correctly calibrated. While some critics may view etiquette as superfluous or negatable,

consider the following example of failure related to etiquette (miscalibration): The ship, Royal Majesty, ran aground because the GPS failed and did not alert the crew that it had failed. A human using a GPS would have alerted the crew if it had failed, so the crew assumed that the automation would do the same (Parasuraman & Miller, 2004). This is a great example of misuse in the form of automation complacency, inappropriate trust and overreliance on automation.

Below is a summary of applied trust in automation between the years of 2002 and 2012 taken from a systematic empirical review by Hoff and Bashir (2015).

Design Feature	Design Recommendation	Source of Empirical Support
Appearance/ Anthropomorphism	Increase the anthropomorphism of automation in order to promote greater trust	de Visser et al. (2012); Pak et al. (2012)
	Consider the expected age, gender, culture, and personality of potential users because anthropomorphic design features may impact trust differently for diverse individuals	E.J. Lee (2008); Pak et al. (2012)
Ease-of-Use	Simplify interfaces and make automation easy to use to promote greater trust	Atoyan et al. (2006); Gefen et al. (2003); Li & Yeh (2010); Ou & Sia (2010); Zhou (2011)
	Consider increasing the saliency of automation feedback to promote greater trust	Wang et al. (2011)
Communication Style	Consider the gender, eye movements, normality of form, and chin shape of embodied computer agents to ensure an appearance of trustworthiness	Gong (2008); Green (2010); E.J. Lee (2008)
	Increase the politeness of an automated system's communication style to promote greater trust	Parasuraman & Miller (2004); Spain & Madhavan (2009)
Transparency/ Feedback	Provide users with accurate, ongoing feedback concerning the reliability of automation and the situational factors that can affect its reliability in order to promote appropriate trust and improve task performance	Bagheri & Jamieson (2004); Bass et al. (2013); Bean et al. (2011); Beck et al. (2007); Dadashi et al. (2012); Dzindolet et al. (2002); Dzindolet et al. (2003); Gao & Lee (2006); Jamieson et al. (2008); Oduor & Wiebe (2008); Seong & Bisantz (2008); Seppelt & Lee (2007); Wang et al. (2009)

	Evaluate tendencies in how users interpret system reliability information displayed in different formats	Lacson et al. (2005); Neyedli et al. (2011)
	Consider providing operators with additional explanations for automation errors that occur early in the course of an interaction or on tasks likely to be perceived as "easy" in order to discourage automation disuse	Madhavan et al., 2006; Manzey et al., 2012; Sanchez, 2006
Level of Control	Consider increasing the transparency of high-level automation to promote greater trust	Veberne et al. (2012)
	Evaluate user preferences for levels of control based on psychological characteristics	Thropp (2006)

Table 5. Empirical Support for designing trustworthy systems. Adapted from *Trust in automation: Integrating empirical evidence on factors that influence trust,* by Hoff, K. A, and Bashir, M., 2015.

Not listed in the table above is a study by Merritt, Heimbaugh, LaChapell, and Lee (2012) assessing the implicit attitudes of users toward automation, and how this helps predict explicit trust in a specific automated system. Implicit attitudes are unconscious processes that affect behavior automatically and predominantly, whereas explicit processes, including explicit trust, are cognitively effortful and conscious. Prior to this, there were no empirical studies examining how implicit attitudes toward automation might affect user's explicit trust in the system. Participants self-reported a measure of their tendency to trust in machines, and completed an Implicit Association Test to measure their attitude toward automation (Merritt et al., 2012). Participants then completed a within-subjects 30-trial X-ray screening task. X-ray images were provided of luggage and participants were asked to determine whether each image contained a gun or knife. While completing the task, they were placed under cognitive load to more accurately simulate real-world working conditions. The user's explicit propensity to trust and implicit attitude toward automation did not correlate significantly. The researchers

suggest that implicit attitudes can be used to calibrate trust, and further believe that this can be useful for establishing trust in new systems, like GPS technology, that previously may have had low levels of accuracy (Merritt et al., 2012).

Applied Discussion

Sheridan and Verplank's ten levels of automation (1978) can vary across automation types which can be clustered by a four-piece model similar to human information processing. The four types taken from Parasuraman et al. (2000) include acquisition, analysis, decision, and action which map to the following stages of human information processing, respectively: information processing, information analysis, decision selection, action implementation. Advanced traveler information systems (ATIS) at their core are a decision aiding technology. Intelligent Transportation Systems (ITS) encompass a wide range of technologies, one of which is ATIS. It provides real time, in-vehicle information to drivers regarding navigation and route guidance, motorist services, roadway signing, and hazard warnings. Other ITS technologies include Advanced Vehicle Control Systems (AVCS) which take over some or all of the driving tasks particularly in emergency situations, Commercial Vehicle Operations (CVOs) which include things like vehicle identification, location, etc., and Advanced Traffic Management Systems (ATMS) which reduce congestion using vehicle route diversion, automated signal timing, changeable message signs, and priority control systems. For the purpose of this discussion only certain aspects of ATIS relating to in-vehicle information will be used. While the main focus is on In-Vehicle Routing and Navigation Systems (IRANS), other related technologies should be mentioned. They are: In-Vehicle Motorist Services Information Systems (IMSIS), In-Vehicle Signing Information Systems (ISIS),

and In-Vehicle Safety Advisory Warning Systems (IVSAWS). Together, these technologies provide drivers with an on-road experience that promotes safety and efficiency (Barfield and Dingus, 1998).

As time and technology progress, level of automation within decision aid technologies, like in-vehicle routing and navigation systems, vary. In this section, a chronological exposure to routing and navigation product releases is discussed in relation to the trust literature. Features of each navigation product are analyzed in terms of factors affecting trust in the system as reported by researchers in the field.

Predating most research in the human automation interaction realm, the 1932 Iter Avto was not quite a decision technology; it was more of a low-level analysis technology. Based on the speed of the car, the system automatically progressed a paper scrolling map display in the direction the vehicle was moving. This system was the first of its kind and built to help alleviate mental workload by eliminating the need to pull the car over to examine a folding paper map while on the road. With a crude manual display, error recovery was cumbersome at best.

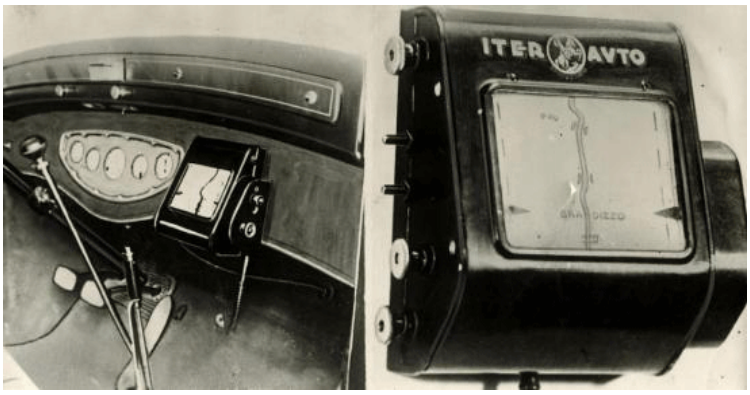


Figure 6. Iter Avto. Reprinted from *Before There was GPS: Personal Navigation in the 1920s and 1930s*, by Dempsey, K., 2013, Retrieved from https://www.gislounge.com/gps-navigation-1920s-1930s/

General Motors created DAIR in the late 1960s as the next step for in-vehicle navigation (Preston, 2013). Among other features including emergency communication, DAIR had the first active communicative display using a network of lights and buzzers to inform the driver which direction to turn. The catch was that drivers needed to have destination-specific cards to insert into the dashboard so that signals buried in the road could be matched and the system could therefore direct them accordingly. While confined to a test area near Detroit, DAIR took the guesswork out of navigating with turn-by-turn instructions effectively eliminating the need for drivers to read and encode street names (Preston, 2013). Helping drivers to make simple navigation decisions like when and where to turn, DAIR was a great example of decision automation. A fair amount of human inputs were still needed for the system to work and despite the directive lights and buzzers, the human had to ultimately carry out the decision to turn. The most interesting tie to research in the 1960s was with Jordan's warning of human motivation loss. With this type of system, what motivates the human to understand routes and be familiar with street names when the machine is capable of doing that process? This small peek at turn-by-turn navigation, albeit never launched, provided great insight into human response to automated turn by turn navigation and issues that may arise.

In the early 1970s, Japanese researchers developed CACS a system that achieved dynamic route guidance using inputs from other ITS technology such as ATMS (Fujii, 1989). Some consider this the first attempt at an integrated intelligent routing and transportation system; using car volume and traffic information to route vehicles appropriately. As early as 1981, Japanese automakers began integrating this concept directly into vehicles. All with slightly different executions, Nissan, Toyota, and Honda

all attempted electronic in-vehicle displays for in-progress routing (Regan, Lee, & Young, 2008). The general point of the display was to communicate location information to the driver in a simple, digestible way. The most abstract display was Nissan's Drive Guide system which showed the driver distance left in bar graph form. This system arguably increased cognitive load by forcing the driver to transform what was seen from the bar graph to the actual route and physical surroundings. Navicom was capable of generating a line between two points and graphically displaying progress on the line abstraction of the route which did not adequately take route shape into account (Newcomb, 2013). With the Navicom display, the line was at least representative of a road which could allow the driver to establish a quicker connection between the display and the real world. Probably the best example of a display was Honda's release of the Electro Gyrocator in 1981 (Newcomb, 2013). The Electro Gyrocator had the first electronic in-vehicle map display. The display showed drivers where they were on a scaled map and which direction they were heading. A few years later, Etak, a U.S. aftermarket system, also offered a robust electronic map display. Maps were stored on cassette tapes and a car icon was placed on the map showing approximate location. In the late 1980s similar beacon-based systems were being tested in London and Berlin (Regan et al., 2008).

Figure 7. Toyota Navicom. Reprinted from *Hand-Cranked Maps to the Cloud: Charting the History of In-Car Navigation,* by Newcomb, D., 2010. Retrieved from https://www.wired.com/2013/04/history-in-car-navigation/

Figure 8. Honda Electro Gyrocator. Reprinted from *Hand-Cranked Maps to the Cloud: Charting the History of In-Car Navigation,* by Newcomb, D., 2010. Retrieved from https://www.wired.com/2013/04/history-in-car-navigation/

Figure 9. ETAK. Reprinted from *Hand-Cranked Maps to the Cloud: Charting the History of In-Car Navigation,* by Newcomb, D., 2010. Retrieved from https://www.wired.com/2013/04/history-in-car-navigation/

Physically placing one's location on an electronic map and displaying progression through an accurately represented space offered a new way for drivers to conceptualize navigation while driving. Thanks to augmented dead reckoning, these products could reliably recreate roadway systems within the vehicle causing an attentional split between the real world and its electronic adaptation. Related research during this time came from Bainbridge (1983) in his ironies of automation paper, warning of skill degradation for humans. This concept introduced the idea that humans can become too dependent on automation and addressed some of the associated dangers.

The aforementioned in-vehicle navigation products of the 1980s all required human inputs. Input methods included keyboards, knobs, or a single button depending on the product. With the exception of Etak, all inputs were operational while driving. Inputs were still necessary but even then posed an interesting driver distraction issue which influenced the direction of driving related research at the turn of the century.

Not long after the debut of electronic in-vehicle navigation systems came the idea that humans may interact with machines similarly to how they interact with other people (Muir, 1989; Lee & Moray, 1992). Using interpersonal trust research as a foundation for a human-automation interaction model, Muir (1994) and Riley (1994) proposed a group of factors contributing to the dynamic relationship of trust between humans and automation. The idea being that factors such as responsibility, competence, and predictability which contribute to how humans trust other humans could be adapted to fit how humans relate to technology.

Muir (1994) focused more on system predictability and consistency as an extension of Barber's (1983) dimensions of technical competence, while Riley (1994) showed how humans could become reliant on automation by organizing associated factors such as trust and perceived risk. Lee and Moray (1992) reflected on the importance of system reliability via performance and system purpose. They argued that system performance was key for establishing trust in an automated system, and that the initial few interactions greatly contributed to the human's overall perception of its reliability.

At the same time as this first real theoretical connection between trust and system reliability, the global positioning service (GPS) began being integrated into in-vehicle

navigation systems as early as 1990, while the first GPS-based system in the U.S. was Oldsmobile's Guidestar in 1995 on a less accurate government scrambled GPS signal (Dunbar, 2015; Mateja, 1995). The accuracy afforded by GPS technology increased the reliability of location services thus affecting navigation system performance. Up until the mid 1990s, factors affecting in-vehicle routing and navigation systems capabilities were the biggest hindrance to the trusting relationship between humans and automation.

Figure 10. Oldsmobile Guidestar. Reprinted from *Automotive navigation systems,* by Arlt, G., 2016. Retrieved from https://www.historicvehicle.org/automotive-navigation-systems/

Realizing the natural inclination for humans to treat technology like other humans, the progression of in-vehicle systems in the 1990s included the addition of GPS and more robust interaction models; some including voice interaction. In 1992, Toyota debuted a voice-assisted system in their Celsior model. The inclusion of voice interaction as a modality for vehicle navigation was important for two main reasons: first, it made the interaction more similar to that of another human, and second, it helped drivers to keep eyes on the road.

Related research in the mid 1990s came from empirical validation and slight modification of the early proposed trust in automation models, Muir and Moray (1996). Contrary to previous thought, findings indicated that faith over system performance was a

better predictor of early automation use. The concept of "faith" is unique in that it is traditionally an affect-rooted term which authors redefined to fit the scope of automation. Faith, or when operator trust becomes an ability to project and predict their own use and belief in the system's future actions, is difficult to measure and better explained by combinations of other variables such as risk-taking, perceived expertise, and self-confidence (Riley, 1994; Kantowitz et al. 1997).

At some point in the mid 1990s with a rise in personal computers, the internet, and technology for the middle class (like GPS), the benefits and accessibility of using automation were realized more widely. In 1996, internet maps and routing directions made an aggressive move into homes and therefore into vehicles via personalized paper maps and routing directions. Drivers would enter start and end destinations and websites like MapQuest would generate a map and list of turn-by-turn instructions for them to print. What seems like a step back in terms of level of automation, was actually a huge step forward into free and accessible ITS for the masses. In the late 1990s most in-vehicle routing and navigation systems products were only afforded by the elite.

The shift from cassette to CD-ROM marked a turning point for both in-vehicle and aftermarket systems. In 1997, Alpine introduced one of the first navigation systems that used GPS and stored maps and the operating system on CDs (Newcomb, 2013). Instead of several cassettes for each city, consumers only had to purchase a disc for a multistate region of the U.S. The early Alpine models used a control that was larger than most modern mobile phones to input characters for a destination address. The first portable GPS systems, including Garmin StreetPilot, debuted in 1998 (Newcomb, 2013).

Figures 11 & 12. Alpine, Garmin StreetPilot. Reprinted from *Automotive navigation systems,* by Arlt, G., 2016. Retrieved from https://www.historicvehicle.org/automotive-navigation-systems/

The emergence of hard-drive navigation systems meant there was no longer a need for multiple discs; the maps were already built in. In 2003, Toyota introduced the first hard disk drive-based system (Newcomb, 2013). Those systems were more convenient and loaded maps and destinations faster, but they were difficult to update, which impacted their reliability for lack of system relevancy.

By the mid 2000s, Garmin, Mio, Navigon, Magellan, TomTom, and others flooded the market with stand-alone GPS devices (Newcomb, 2013). These devices used available operating systems like Embedded Linux or Windows Embedded CE. Some of the higher-end models offered real-time traffic, map upgrades, terrain mapping, and high-resolution screens. Typical consumers bought one device and just used it with multiple vehicles for many years. The most important changes in newly-released models related mostly to map data and the point of interest (POI) database which most consumers were willing to forgo.

Figures 13, 14 & 15. Garmin, Mio, Navigon. Reprinted from *Automotive navigation systems,* by Arlt, G., 2016. Retrieved from http://www.gpsinformation.org/i3/i3.html

Figures 16 & 17. Magellan and TomTom. Adapted from *Automotive navigation systems,* by Arlt, G., 2016. Retrieved from https://www.historicvehicle.org/automotive-navigation-systems/

Research-wise, the concern of automation adoption due to poor system performance and reliability (disuse) quickly turned to a concern of human judgement in using the technology in spite of its failures. Parasuraman and Riley (1997) created a four-part classification of automation use that lists overreliance on technology (misuse) as a major detriment to both performance and ultimately trust over time. Instead of working to understand how to establish and maintain the highest level of trust in a system, researchers moved to look at how to achieve appropriate reliance and trust in an automated system.

The switch from theoretical connections with interpersonal psychology to applied design of automation occurred in the mid 2000s. As in-vehicle routing and navigation systems technology became more reliable and products flooded the market, the focus fell to interface and information display. It was realized that the way in which drivers physically interact with the system may affect psychological judgments of the system. From physical inputs to visual and auditory perception of information, research surfaced on how to design in-vehicle routing and navigation systems for appropriate trust.

Navigation systems at the time (and still today) used three primary guidance display screens to communicate navigation information to drivers (a) maps, (b) direction lists, and (b) turn-by-turn guidance which notifies the driver before a turn. "Maps can be effectively used to plan a route since they provide a pictorial representation of an area or region, while ordered lists of directions can limit information processing and lead to fast and accurate navigation performance." Portable aftermarket systems at this time included Garmin, Magellan, and TomTom, most of which boasted color displays, split screen views for map and turn-by-turn directions, voice commands, and dedicated hard controls for system inputs and repeated functions (Llaneras and Singer, 2003).

Beyond basic usable, functional system design, research in the coming years aimed to find the extent to which increasingly human-like automation influenced trust in the system and optimal performance/continued use over time. Anthropomorphism and etiquette made another strong research surge during this time, claiming that age, gender, culture, personality and politeness may impact trust differently for diverse individuals (Gong, 2008; Lee, 2008). Alternatively, on the system level, providing continuous and accurate feedback promotes trust and improves performance (Dzindolet et al., 2002;

Wang, Jamieson, & Hollands, 2009). Results from this era of research yielded in-vehicle routing and navigation systems features such as personalizable voice output, more natural language searches by grouping (i.e., restaurants, pharmacy), and even confidence intervals for estimated time of arrival (ETA).

2007 began the smartphone's rise in popularity as GPS was readily integrated and applications such as Google Maps gained traction. Google Maps was free for iPhone users at the start and remains free today on a wider variety of platforms (Gruber, 2006). Most of the changes through the turn of the decade centered around social inclusion. Technology at this time shifted from mostly functional to highly social. Companies like Waze developed systems leveraging social altruism and aspects of play to create a trusting community of users (Empson, 2013). Following the social trend, people no longer only wanted to navigate to a destination, they wanted to navigate to a person (who may not always be stationary). Location sharing features from personal smartphone devices began appearing and continue to be used for routing and navigation among other things.

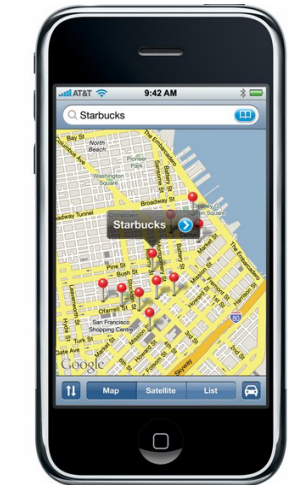


Figure 18. Google Maps. Adapted from *Automotive navigation systems,* by Arlt, G., 2016. Retrieved from https://www.historicvehicle.org/automotive-navigation-systems/

Computing restrictions in the early years of navigation products did not allow for robust psychologically based interaction research. The refinement of electronic displays in the 1990s seemed to jumpstart the research on trust in technology. Information reliability and predictability was important, but how information was displayed to the human had the most impact on affective judgments such as trust. The emerging pattern, at least for this domain, is that the development of new technology raises societal questions further propagated by previously existing theoretical models. As more is understood about the applications of the new technologies, theoretical models are adapted to fit the more specific use cases and then empirically supported becoming applied models. Based on results from empirical studies, changes are made for new products using similar technologies in order to improve the experience or performance. With the increasing reliability and adoption of technology, finer details within design and technology implementation were able to be researched thus extending applied models in various directions.

The latest technology trend for in-vehicle routing and navigation systems combines advanced traveler information systems with advanced vehicle control systems in semi autonomous and fully autonomous vehicles. This effectively shifts in-vehicle routing and navigation systems from a decision automation to an action implementation technology (Parasuraman et al. 2000). This combination of intelligent transportation technologies and subsequent switch to action implementation automation is outside the scope of this review, but should be considered when designing future generation in-vehicle routing and navigation systems that are fully integrated into autonomous vehicle systems. Recent studies on trust in autonomous vehicles focus on the combined effect of

this technological merge (Waytz, Heafner, & Epley, 2014; Koo *et al*., 2015). There may be a benefit to isolating advanced traveler information systems and advanced vehicle control systems to determine their effects more accurately separately and combined. As the field continues to expand and semi autonomous vehicles become mainstream and accessible, it is expected that new models for trust will continue to appear and be empirically validated per the recognized pattern above.

Important Related Works

Decision Aids

Decision support systems are fundamental to decision aiding in a number of complex decision environments, and are especially critical in high-risk military command and control tasks (Crocoll & Coury, 1990). How aids are presented to the user on the interface level is still in question. Should the decision aid select alternatives and determine the action to be taken; or should the decision aid present status information about the situation and leave the decision in the hands of the human? Researchers across multiple domains have tested various types and levels of these systems. Most decision support aids fall within the mid-range levels of the Parasuraman, Sheridan, and Wicken's framework, indicating that the operator must decide on the action with status information or a suggested course of action from automation.

Crocoll and Coury (1990) revealed the influence of decision aids and how each type, status, and recommendation impacts performance and strategy. Their research demonstrated that decision aids do have an impact on decision-making performance. They found that the presentation of status or recommendation information significantly reduced the time required to identify the target and make a decision. The decrease was significant and the groups who were assisted by the decision aid were approximately three times faster than the control group. Given that decision aiding was found to be significantly better than no aid, the question of selecting the appropriate decision aiding information rose in importance (Crocoll & Coury, 1990).

The following three areas were examined in an attempt to determine the best type of decision information: (a) accuracy and response times for each information type at the

different levels of accuracy of the decision aid, (b) the subject's strategy selection, and (c) the subject's comments. There were four decision aid conditions: a control group where no decision-aiding information was provided; a second group that received only status information; a group that received only recommendation information; and a fourth group that received both status and recommendation information. Results indicated that, in general, providing decision-aiding information reduced the time required to complete the task. Differences among the three types of decision-aiding information occurred under those conditions when the decision aid was incorrect. When the decision aid provided inaccurate information, the group receiving only status information was least affected by the decision aid and was best able to correctly identify the aircraft (Crocoll & Coury, 1990).

This study provides us with pivotal information about the importance of decision aid accuracy (henceforth known as reliability), and opens the door for further research on trust in decision support systems for highly critical tasks where responsibility of decision error still falls with the human operator.

When discussing decision aids in the context of trust, it is important to mention works on workload, complacency, and situation awareness. Workload in this instance is synonymous with cognitive load and is essentially a form of cognitive arousal, whether it be memory and/or attention based. Workload must be designed to be appropriate for the task that is to be performed. Biros, Daly, & Gunsch (2004) wanted to know how a user's task load level affects the relationship between user trust and automation use. They believed task load would be a moderator to the positive relationship of trust and automation use. Through recording performance of command and control tasks, they

found that the use of system's automation is directly and positively related to level of perceived trust the person puts in the system. Increased task load may lead to user over-reliance on automation decision support systems despite lower perceived trust in the system (Biros et al., 2004). This shows that users rely on information technology even when its veracity is in question.

Overreliance on automation as a decision solution to minimize workload leads to issues such as complacency. If reliability is high but not perfect, human operators may not monitor the automation and its information systems and therefore may not detect failures (Parasuraman, Sheridan, & Wickens, 2010). Complacency is said to be a contributing factor in situation awareness. Situational awareness refers to how automation of decision-making functions may reduce a human operator's awareness of the system. A lack of situation awareness is related to what Endsely and Kiris (1995) call the "out-of-the-loop performance decrement". Researchers ran a 30-trial within subjects study during which participants participated in a dynamical decision-making task under various conditions of automation support that would fail at random. Researchers found that subjects were more confident in their decisions when they were operating with increased automation assistance. Upon automation failure, lower levels of performance and situation awareness were recorded (Endsely & Kiris, 1995). This was thought to be caused by complacency, thus confirming its relationship to the out-of-the-loop performance problem.

Cognitive processes in trusting decision aids. Trust could possibly be obtained through the proper design of training and user interfaces that would provide information

regarding the automation process performance and purpose. The simple act of having this information is not enough to ensure that the appropriate amount of trust is gained. This is due to the fact that trust is developed via an underlying cognitive process and must be presented in a way that coincides with said process. It is through Lee and See (2004) that we can ascertain three of these processes: analogical, analytic, and affective (or emotional). Just as in name, each differs in ways of cognitive processing (Duez, Zuliani, & Jamieson, 2006).

When considering the analytic process, we find that the user's previous experience, knowledge, and mental model provide the basis for information processing. The motives, interests, behavior, and capabilities of the other party are all factored into this decision making process. The highest amount of cognitive resources is required from the human operator when adjusting for analytical trust. In contrast, analogical judgement is procedural and rule oriented, leaving the human operator with less cognitive demand. The rule-based expectations of analogical judgement pairing with the situation can increase a person's overall expectation of satisfactory performance. Reputation and gossip also hold bearing when contemplating analogical trust (Atoyan, Duquet, & Robert, 2006).

Finally, affective aspects of trust represent how feeling and behavior can play a part in the core influence of trust. The feelings and emotions of an individual can influence the base judgement of any technology or activity. Affective association an impression can be equally, if not more, efficient that the simple weighing of pros and cons based off memory of relative information. This is especially true when mental resources are limited or the complexity of a decision or judgement is increased (Slovic et al., 2004).

Trust building also holds a temporal dimension. The development of trust can happen via, training, self-experience, hearsay, or the experience of others. It is in these instances that Miller (2005) recognizes that the analytic process can falter in importance when compared with the human interaction of the affective and analogical processes, especially when the system is novel in nature. Affective information may be the only source when considering the first time use of a novel system. With no background information of the agent's past behavior or motives, trust can only be ascertained through what affect, or feeling, a system provides. If the affect on the human operator is intolerably negative, prompting a response of "do not trust", the odds of further information gathering are significantly decreased. If only a small or moderate amount of affective trust tuning is needed, an operator will continue to use the system. This further use and familiarity will allow trust to be developed through the analogical process. It is through assertion and archiving of these two trust processes that analytical trust can begin to build (Miller, 2005). Essentially, the adjusting of analogic and affective processes will play a larger part than that of memory, logic, and experience associated with the analytical process.

Miller and Parasuraman's findings (2004) show that the user's trust in a system is greatly due to the machine etiquette itself, especially when dealing with complex systems. We define machine etiquette as the accepted behavior, predefined roles, and the interactions of the system with operators, humans and intelligent agent participants, in a common area (Miller, 2005). These experiments show that, even in reliable systems, the trust and performance are lowered when the automation etiquette is poor. It is pointed out by Lee and See (2004) that the good etiquette of a system can support both analogical and

affective trust tuning development.

Importance of Information Reliability

Information and system aid reliability are often difficult to calculate. This is because it is often based on mean automation accuracy over time and designed for contexts other than the real context in which the operator will use it. This factor is important to consider, because it is known to impact user trust in the system and can therefore undermine its performance benefits.

Wang, Jamieson, and Hollands (2009) studied the effectiveness of using system reliability information to support appropriate trust and reliance on a Combat Identification System (CID) aid. The simulated command and control task is meant to closely resemble the high-risk, highly critical decision environment of combat. In the study, aid reliability was manipulated and three conditions were tested: no aid, 67%, and 80% aid. Aid reliability disclosure was also considered, that is, whether the participants were informed or not informed of the aid's reliability level. Trust and reliance were collected via questionnaire and performance was measured by error rate and response time. Ultimately, Wang et al. (2009) found that low cost of reliance means less disuse. The appropriateness of reliance is dependent on reliability knowledge (resulting from aid disclosure) of the system. This means that users will adjust reliance more appropriately when informed of the reliability levels and imperfect automation can actually improve CID performance.

In a slightly different domain, Madhavan and Phillips (2010) had participants complete a luggage screening visual search task with the assistance of an automated

decision support system that varied in reliability from moderately reliable (70%) to highly reliable (90%). The second independent variable of interest in this study was computer self-efficacy (CSE). This concept is a subset of expertise in that it is defined as a person's judgment of his/her ability to use a computer that is essentially perceived as directed expertise (Madhavan & Phillips, 2010). The goal was to examine the relationship between CSE, and trust and utilization of the system. High-CSE participants trusted the system more, complied with it more, and generated significantly more hits than low-CSE participants, particularly on trials in which the aid was highly reliable. The results indicated that high-CSE levels led to a better ability to gauge the true capabilities of the system. However, all participants consistently underestimated the actual reliability of the system at both levels of CSE (Madhavan & Phillips, 2010).

In a less critical domain, Kantowitz, Hanowski, and Kantowitz (1997) measured the relationship of trust and self-confidence to driver acceptance of automated traffic information systems (ATIS). This study was designed to provide data that would aid the ATIS designers in selecting an appropriate level of system accuracy (reliability). Two traffic routes were given to participants; one was a "familiar network" and the other was a "new city." Information accuracy was set at 100%, 71%, or 43%, and the authors found that drivers with low self-confidence in unfamiliar settings would accept and trust less accurate information. Kantowitz et al. (1997) defined confidence as how sure a person was in decision-considering factors such as expertise. Other important trust-related findings from this study include: (a) trust in the system is higher when information is accurate, (b) trust could be recovered if accurate information is presented on subsequent links, and (c) inaccurate information decreases trust more in familiar settings. Essentially,

when self-confidence is greater than trust, manual control is preferred (Kantowitz et al., 1997).

Anthropomorphism

Anthropomorphism occurs when any entity is given human-like qualities. Implementation of anthropomorphism within technology can be manifested in many ways, including: visually (Pak, Fink, Price, Bass, and Sturre, 2012), through language and sound (Parasuraman & Miller, 2004; Nass & Lee, 2001), and physical shape (Lee, Jung, Kim, & Kim, 2006).

Pak et al. (2012) found that for younger adults, subjective trust and objective behavioral trust were significantly lower when presented with a non-anthropomorphic aid compared to an anthropomorphized aid. Objective behavioral trust was measured through peek behavior, where participants had the option to reveal four possible answers. This behavior was specifically related to the authors' defined concept of dependency as a form of reliance on automation. Pak et al. also discovered that participant answer time was reduced with the use of an anthropomorphic aid. The authors concluded that increased trust leads to increased dependence on the aid, which in turn leads to faster performance (Pak et al., 2012).

Parasuraman and Miller (2004) studied the effects of automated information delivery in terms of etiquette, or machine politeness. They discovered that non-personified machine etiquette strongly affected human trust, usage decisions, and human-automation system performance. Good automation etiquette enhanced diagnostic performance, regardless of automation reliability; its effects overcame low reliability.

Another example of etiquette-affected trust in automation comes from Nass and Lee (2001). Nass and Lee found that in order to maximize liking and trust, designers should set parameters that create a personality that is consistent with the user and the content being presented. These parameters may include qualities such as words per minute and frequency range.

Research in the area of anthropomorphism has also demonstrated that a robot's personality may be a trust factor in complex domains such as manufacturing and aviation. Humans trust a polite and friendly automated system; this can compensate for low reliability in terms of the development of trust in high criticality automation (Oleson et al., 2011). Part of automation taking on human-like qualities is the idea of the operator craving a sense of social presence in the interaction. This begs the question, is physical embodiment required for successful social interaction between humans and social robots? In a more physical interpretation of anthropomorphism, Lee (2006) discovered that physical embodiment of a social agent without tactile interaction can negatively affect participants' feelings of social presence. While trust was not measured in Lee's study, there is some merit in tactile interaction with an agent so strongly affecting psychosocial constructs.

More recently, computer etiquette suggests that recognizing and understanding the interactions of social and work context with the roles of the computer and human to specify acceptable behavior, can enhance human-computer interactions (Parasuraman & Miller, 2004). In high criticality domains, etiquette exhibited by non-personified machines and computer-based automation (such as facial expressions, speech, voice tones, and gestures) can overwhelmingly affect users' perceptions and the correct,

optimal usage of them (Parasuraman & Miller, 2004). Parasuraman & Miller (2004) hypothesized that good etiquette could compensate for poor reliability and result in increased usage decisions. They tested this in a 2x2 flight simulator study using low (45%) and high levels (70%) of reliability and low and high etiquette, they found that machine etiquette could strongly affect human trust, usage decisions, and human-automation system performance. Good automation etiquette enhanced diagnostic performance, regardless of automation reliability. The effects of etiquette overcame low reliability, and performance of good etiquette/low reliability was almost as good as high reliability/low etiquette (Parasuraman & Miller, 2004).

Perhaps the greatest example of a failed anthropomorphic aid to date is Microsoft's "Clippy". Clippy was a semi-anthropomorphic paper clip that provided tips and help to Microsoft Office users in the 1990s and early 2000s. Stanford professor Cliff Nass was hired by Microsoft to determine why their automated helper aid was so hated. In their book, "The Man Who Lied to His Laptop" Nass and Yen (2010) explain that Clippy defied most acceptable social conventions in trusting human to human relationships. Clippy gave frequent and unhelpful information, was persistently annoying, and used formal language to address users. Nass created and tested a "Clippy 2.0" which employed the use of a scapegoat bonding Clippy and the user against the common enemy of Microsoft. Research found a drastic change of heart in Microsoft Office users between the two versions of Clippy. Clippy 2.0 was extremely well-received, but due to the self-deprecating nature of Nass' solution to making Clippy more lovable, it was never implemented (Nass & Yen, 2010). Generally speaking, productivity and acceptance are likely enhanced with designs that consider affect (Norman, Ortony, & Russell, 2003). But

in the case of Clippy, the designed affect was not implemented correctly and was therefore detrimental to use (Nass & Yen, 2010). Together this research suggests that the emotional and attitudinal factors that influence human-human relationships may also contribute to human-automation relationships. Additionally, properly executed anthropomorphism makes all the difference in how humans interact with automation.

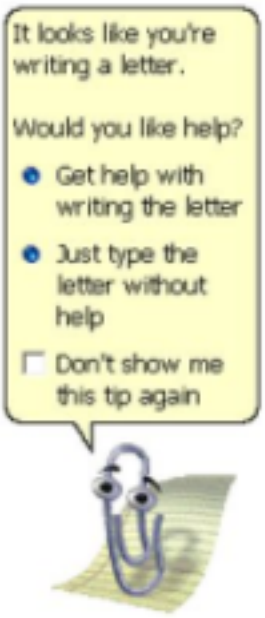


Figure 19. Clippy. Adapted from *The man who lied to his laptop: What we can learn about ourselves from our machines*, by Nass, C., 2010.

Realism

In trying to conceptualize realism as a graphical communication form, knowledge was sought from the domain of comics and the works of Scott McCloud. In his book *Understanding Comics,* McCloud says, "When pictures are more abstracted from reality they require greater levels of perception, more like words". McCloud addressed the many categorizing factors associated with realism, stating that imagery can be placed on many scales from realistic to iconic, complex to simple, objective to subjective, and specific to

universal. Ultimately, he claimed that words are the ultimate abstraction leaving most to the reader's mind (McCloud, 1991).

McCloud continued to parse out communicative imagery in a thought provoking way. A photorealistic picture of a human is a close approximation of what the retina would receive if looking at a real person. The meaning of the photo is therefore attained by way of resemblance to reality. While cartoon imagery is stripped of details or has exaggerated features which moves the perceiver away from resemblance, it still manages to convey that basic meaning. According to McCloud (1991), the continuum from realistic to cartoon images represents increasing levels of what he calls iconic abstraction. Iconic abstraction is defined as "removing an image from its retinal source, but still retaining its basic meaning". When iconic abstraction is expressed at its fullest, it results in a compilation of pure shapes, colors, and lines, almost completely devoid from meaning. When you connect the lines, you get a "Big Triangle" that aims to systematically place imagery based on the three aforementioned components of realism, abstraction, and meaning.

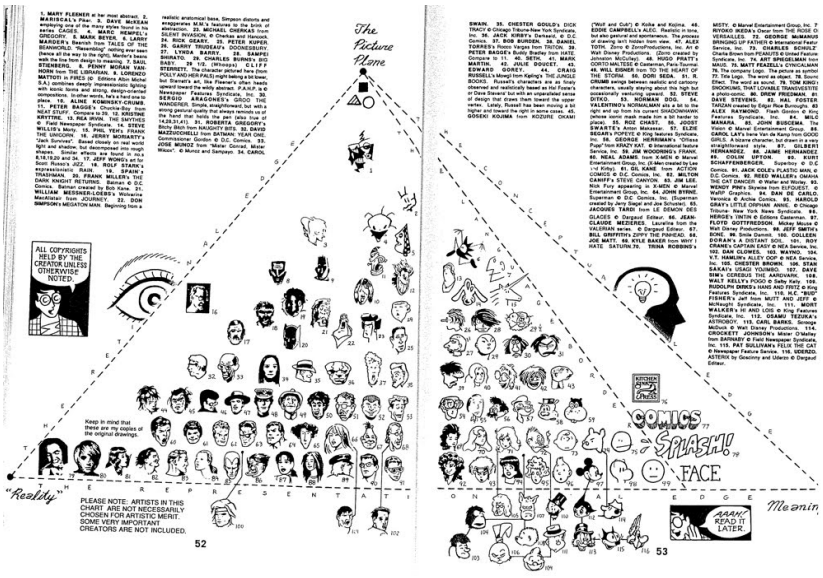


Figure 20. Big Triangle. Reprinted from *Understanding comics: The invisible art* (52), by McCloud, S., 1993.

Bringing image realism back to the field of Human Factors is a study by Yeh and Wickens (2001) that evaluated scene realism in an augmented reality cue detection task. The realism of the graphic images used were hypothesized to influence reliance on automation cuing. Few studies have empirically studied how the distribution of attention within a display is affected by image or scene detail, or how the reality with which the information is displayed biases operators to trust one data source over another. One example of empirically tested image quality also with a cueing task is a study by MacMillan, Entin, and Serfaty (1994) who showed that when image quality is poor, operators under trust cuing information. For Yeh and Wickens (2001) high realism did not ultimately promote greater trust. In the highly realistic setting, participants were more conservative in their responses and less willing to report a target. Most importantly, it was found that realism did influence reliance on the automated cuing information. The findings replicated by MacMillan et al. (1994), suggested that operators would over rely on automation when the image quality was high.

Study 1

Methods

Automated decision aids span multiple domains and risk environments. For this study, low risk routing and navigation decisions were used to evaluate the effect of anthropomorphism on various aspects of trust and performance using the automated decision aid. Specifically, a prototyped automated traveler information system (ATIS) will provide status and decision information on route selection for dynamic navigational environments.

Experimental design. This study was a 2×3×2 mixed factor design. Independent variables included the level of information reliability (2 levels: 72%, 90%), anthropomorphic aid image (3 levels: non, low, high) and advice type (2 levels: good, bad). We also included a separate control group that had no automated decision aid. Dependent variables were performance (task time, proportion of correct responses, confidence in decision) and trust (objective, subjective trust). Separate mixed factor ANOVAs were run to test the below null hypotheses for the two experimental dependent variables:

H₁ Level of information reliability alone does not influence human trust (and performance)

H₂ Level of anthropomorphic image alone does not influence human trust (and performance)

H₃ Type of aid advice alone does not influence human trust (and performance)

H_4 Levels of anthropomorphic imagery and information reliability do not interact to influence human trust (and performance)

H_5 Levels of information reliability and aid advice type do not interact to influence human trust (and performance)

H_6 Levels of anthropomorphic imagery and aid advice type do not interact to influence human trust (and performance)

H_7 Levels of anthropomorphic imagery, information reliability, and aid advice type all do not interact to influence human trust (and performance)

Anthropomorphism. The main manipulation of interest is anthropomorphism, which is defined as having human-like qualities. This is most commonly portrayed through visual and auditory channels. It can include anything from human-like shapes, sounds, and facial features, to etiquette and humor. First we must consider the base form of what is taking on human characteristics. Characters in animated films are excellent examples of object-based anthropomorphism. In the 1991 Disney feature film Beauty and the Beast, household objects are given voices and faces, making them appear human-like. Pixar similarly personified automotive vehicles in the 2006 movie, Cars. Human-like traits can also be imposed on animals. Again, animated films such as Disney's Robin Hood (1973), and Pixar's Finding Nemo (2003) both exemplify animal-based anthropomorphism. Human cartoonism is an additional base form or anthropomorphism in which human drawings are given life-like human characteristics. The extent to which these attributes are humanized determines the level of anthropomorphism on a linear spectrum of realism.

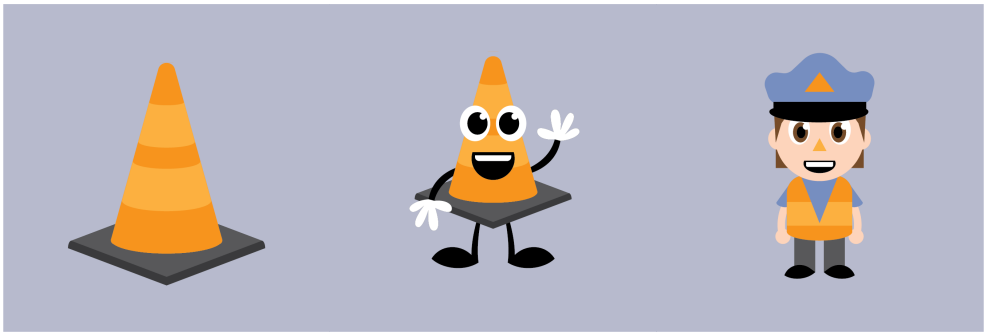
For the purpose of this study, to be fully anthropomorphized is to be human. In terms of visual imagery, this is most commonly represented as a photograph of an actual person. A non-anthropomorphic image is one where no human attributes are found. Previous studies have evaluated the full presence and absence of anthropomorphism in imagery on trust in information decision aids (Pak et al., 2012). None have evaluated any mid-range visual anthropomorphism. Societal exploitation of mid-range anthropomorphism over the years could lead to more accepting socio-emotive feelings towards anything with human-like features. A real human image is not relatable because it is clearly not the viewer. An inanimate object is not relatable because it is non-human. However, something in the middle may be easier for a viewer to map themselves to. This study focuses on the evaluation of mid-range visual anthropomorphism with human-computer trust. Through this we can better understand the space of visual anthropomorphism and the intricacies of how humans relate to and accept anthropomorphized automation decision aids.

The current study evaluates three anthropomorphic conditions: (a) non-anthropomorphic, (b) low anthropomorphic, and (c) high anthropomorphic. The non-anthropomorphic condition includes imagery of an object relating to the domain; in this case, a traffic cone. The low anthropomorphic condition is an object-based cartoon abstraction of the traffic cone to include a face, arms, and legs. The high anthropomorphic condition is a human-based cartoon loosely based on a gender ambiguous traffic cop. The goal is to keep domain specificity as a constant across all image conditions which was not apparent in the design of Pak et al. The implied expertise

of domain-specific imagery is an important confounding factor which should be controlled.

A pilot test was conducted on a series of images in an internet distributed survey (n = 40). The survey including other domain specific image variations such as a traffic light and cartoon chauffeur, and tested for perceived levels of anthropomorphism, gender neutrality, and intelligence. Based on the results, we selected three images for non, low, and high anthropomorphic aid images as depicted in Figure 21.

Figure 21. Anthropomorphic images (non, low, high)



Reliability. Two levels of reliability are used that assume a practical rather than theoretical model of ATIS testing. Pursuant to previous studies of trust in decision aids, a slightly greater than 70% reliability condition was selected to represent the average of moderate-level reliability domains. This number encompasses both a realistic minimal viable value for adequate decision aid use (i.e., minimizing disuse and misuse), and is thought to promote appropriate levels of human trust in the system. The second reliability condition will be set at 90% in keeping with the Federal Highway Administration's (FHA) estimated system accuracies of current ATIS (Toppen & Wunderlich, 2003).

Aid reliability refers to the system accuracy over time, therefore 72% reliability condition includes 11 failures out of 40 trials, and the 90% reliability condition includes four failures out of 40 trials. Failures are an output of inaccurate decision aid information. Participants will be given feedback on whether they are "on time", or "late" which will serve as route choice confirmation.

Good vs bad advice. While reliability is viewed holistically as the total percentage of non-failures in a condition (e.g., 72% and 90%), within each condition there are a certain number tasks designed to succeed and fail. These tasks will be referred to as good advice and bad advice, respectively. Good advice occurs the majority of the time in both reliability conditions and feeds the participant truthful advice on which route to take. Agreeing with good advice will always get the participant to the destination on time. Bad advice is when the decision aid feeds the user wrongful advice. That is, it tells the participant to take a route that is not the most efficient. Complying with wrong advice will always cause the participant to be late for that given task. Therefore, an advantageous performance outcome for bad advice trials occurs when the participant disagrees with the bad advice. The order of good and bad advice trials was determined randomly, but within a few trust building parameters following best practices for human interaction with new technology. All training trials contained good advice, the first 12 trials contained only good advice to allow for initial trust in the system, no more than two bad advice trials occurred in a row, no more than ten good advice trials occurred in a row (aside from the original 12 used to build initial trust in the system).

Dependent measures. Dependent variables were measured after each trial. Performance-based metrics included task time, proportion correct responses, and confidence in the chosen answer. In addition, for the conditions with a decision aid, subjective and objective trust in the aid were assessed.

Participant's behavior (whether they agree, disagree, or peek) was analyzed and used as an 'objective' measure of trust. The rationale is that if participants immediately agree with the aid without peeking, that could be considered a high level of trust in the aid. However, if participants disagree without peeking, it would indicate a complete lack of trust. If participants peek before making a decision it could represent moderate levels of trust. Behavioral trust is a scale from one to four used by Pak et al (2012). If participants immediately click disagree, a value of one (no trust) will be given. If they peek and eventually click disagree, that trial will be assigned a two (moderate distrust). Peeking and agreeing will be assigned a three (trust but verify), and clicking agree assigned a four (trust).

If trust is the attitude, reliance is the behavior and can be measured using objective task performance measures. The combination of time on task and decision selection are frequently used to determine reliance. Reliance may be appropriate when operators trust automation that is either reliable or more reliable than manual operation, or inappropriate when operators trust automation that is either inaccurate or less reliable than manual operation (Dzindolet et al., 2003).

A series of surveys, including a brief demographic survey, followed the empirical testing. Substantial evidence demonstrates that trusting tendencies, considered as a personality trait, can be reliably measured and can influence behavior in a systematic

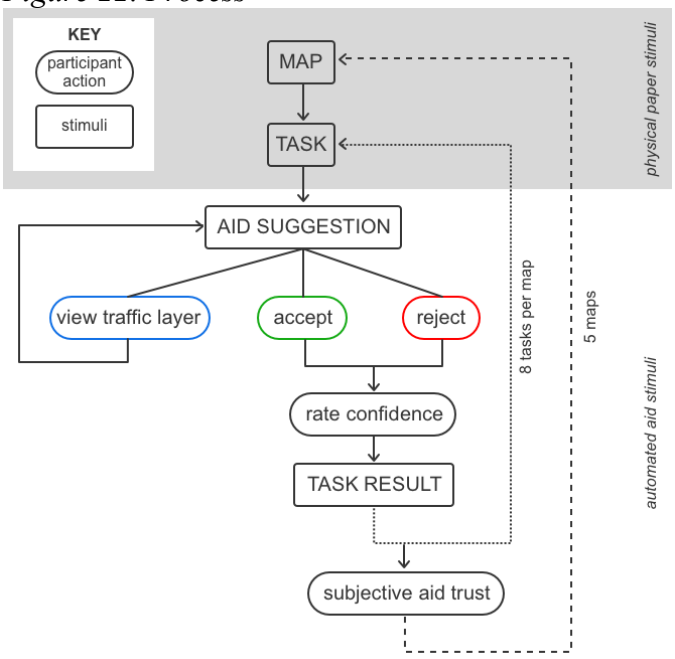
manner. Rotter's Interpersonal Trust Scale was administered to help differentiate people on their propensity to trust others. People with a high propensity to trust fared better in predicting others' trustworthiness than those with a low propensity to trust (Kikuchi, Wantanabe, & Yamasishi, 1996). Since trust in automation theorists have shown interpersonal trust to be related but not equivalent, the Jian et al. trust in automation scale was used to measure subjective trust in automation for this study. Additionally, a baseline of perceived anthropomorphism was assessed via survey adapted from Epley, Akalis, Waytz, & Cacioppo (2008). People who identify as being lonely are more likely to have a higher baseline for perceived anthropomorphism. To help sort through individual differences, a personality assessment was given. The BFAS was used to identify some big five traits related to the decision actions and preferences of participants (i.e., neuroticism and performance).

Participants, procedure, and analytic plan. The study took place at the University of Minnesota – Twin Cities campus. A total of 106 university students participated (34 male, 72 female, average age 20.5) for course credit or cash compensation. The data from one male participant was excluded on the basis of self reported city familiarity discovered in the demographic survey. The data were replaced with that of another male participant resulting in 105 usable participants (33 male, 72 female, average age 20.5). All participants were tested individually in a single experimental session lasting about 90 minutes.

Figure 22 illustrates the procedural flow. Participants in each condition were provided with a total of 40 tasks, each pertaining to a unique map of an unfamiliar

metropolitan area. There were five different maps used throughout the study, with eight consecutive tasks per map. Participants received a physical map with realistic information about the area including general traffic trends. Then participants received a task notecard with a brief written scenario such as the following: *"It's currently 8:40AM, you have an exam on campus at 9:00AM. Which route should you take?"* A practice map and four practice tasks were used as training for the experimental trials.

Figure 22. Process

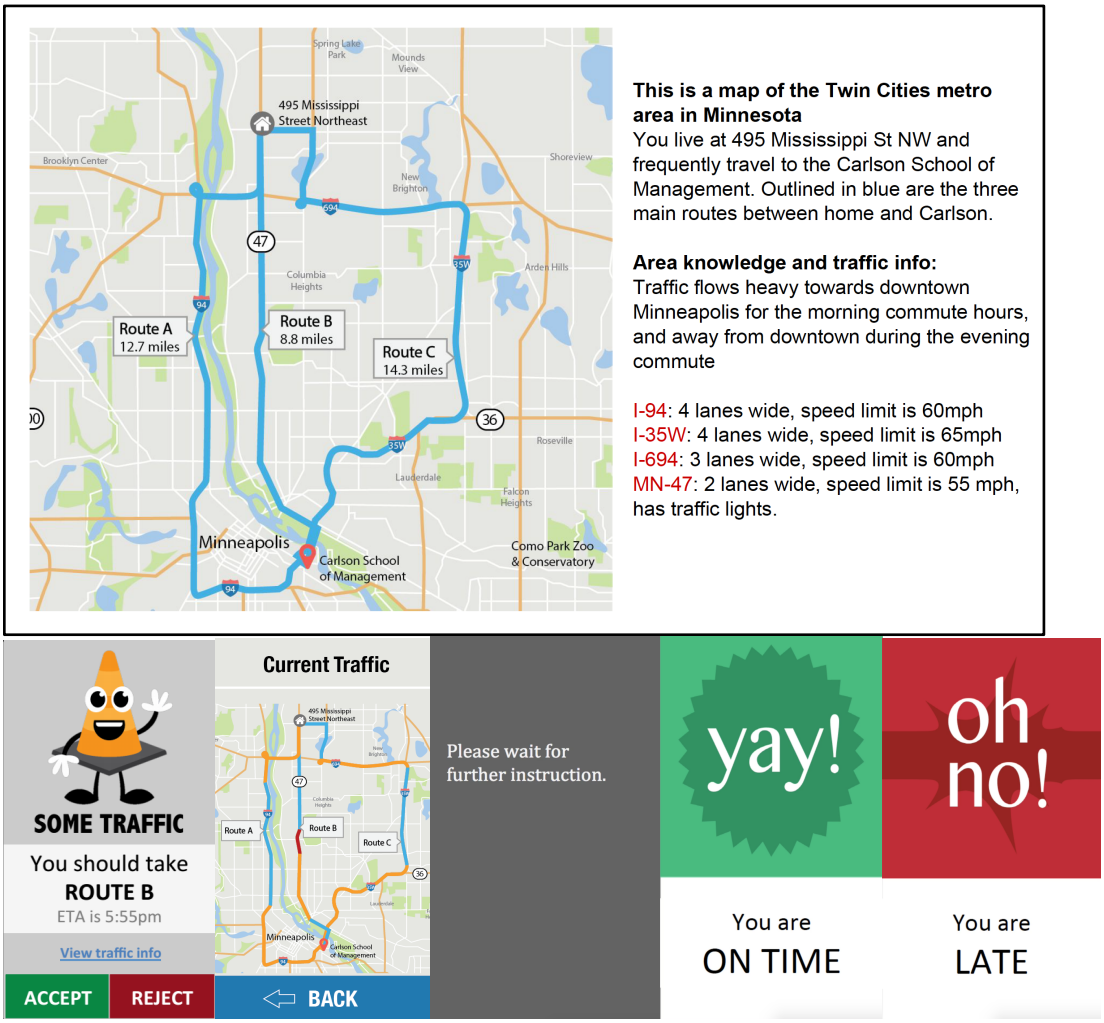


In all conditions except the control, the participant was directed to click through the prototype on the iPhone (automated decision aid) and decide to accept or reject the suggested route from the aid. The aid presented an accurate or inaccurate recommendation (termed "good advice" or "bad advice" respectively) to the participant paired with some level of anthropomorphic image. The participant could accept or reject the aid suggestion or choose to view more information. The "view more information" provided the participant with a visual traffic overlay for the routes which always reflected

accurate traffic information for the task at hand. After accepting or rejecting, the participant was directed to verbally rate the degree of confidence they had in their decision on a scale of 1-10. The next screen on the aid then visually displayed either an "on time" or "late" confirmation message. After all eight tasks for one map were completed, participants were asked to rate their overall level of trust in the automated decision aid on a scale of 1-10. This process was repeated for the remaining four map scenarios.

Once all 40 tasks were completed, participants were asked to complete post-task surveys including but not limited to demographic information, questions concerning the perceived anthropomorphic mental state qualities (e.g., good intent) of each participant's assigned smartphone image (not applicable for the no-aid control condition), and personality assessments. Sessions involving the smartphone were both audio and screen recorded using QuickTime and saved locally to a secured lab computer. The stimuli were created with InVision prototyping software. Researchers entered experimental data in real time as participants received feedback on the decisions, indicated confidence levels, as well as rated trust in the automation aid. After each participant, researchers coded the time data by watching and listening to the session recordings.

Figure 23. Stimuli in order of appearance



Results

Dependent variables were organized into two main categories pertaining to anthropomorphic image manipulation (no-aid control, non-anthropomorphic aid, low anthropomorphic aid, and high anthropomorphic aid). Performance measures (time to decision, correctness of decision, confidence in answer) were assessed in all four conditions. Trust variables (subjective and objective trust) only applied to the three aid

conditions. Both behavioral and self-report data were evaluated for each of the two dependent variable categories.

Performance. A 2 (reliability, between) ×3 (anthropomorphism, between) ×2 (advice type, within) mixed factor ANOVA was run using the *nlme* package in R. This analysis revealed a significant interaction of reliability and advice ($F(1, 84) = 9.03$, $p = .0035$). Subjects in both reliability conditions performed equally well when receiving good advice, but subjects in the 90% reliability condition performed worse when receiving bad advice than subjects in the 72% reliability condition. As shown in Figures 3 and 4, the effects of bad advice were more pronounced for the 90% reliability condition ($M = .92$ and $M = .55$ for good vs bad advice, or a difference of .37) than for the 72% reliability condition ($M = .91$ and $M = .71$ for good vs bad advice, or a difference of .20). We found no significant interactions of reliability, anthropomorphism, and advice, $F < 1$, reliability and anthropomorphism, $F < 1$, or anthropomorphism and advice, $F < 1$. The control condition showed $M = .40$ for proportion correct across all trials without the help of an automation aid.

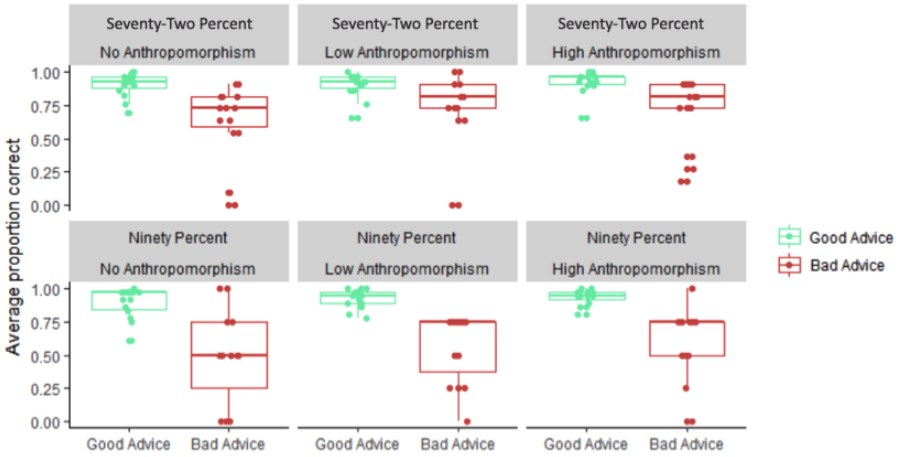
A significant main effect of advice type was seen ($F(1, 84) = 108.77$, $p < .001$) with higher correct responses for good advice ($M = .92$) than for bad advice ($M = .63$). There was no effect of image type ($F = 1.51$) and no reliability condition and image type interaction ($F < 1$). However, overall performance was numerically greater in the high anthropomorphic condition than in the no anthropomorphism condition both for the 72% reliability condition ($M = .83$ vs $M = .78$ for high and no anthropomorphism respectively)

and for the 90% reliability condition (M = .77 vs M = .69 respectively). Medians are also reported due to the relatively low number of trials in which bad advice was given.

Table 6. Performance Summary Tables

Performance Summary – MEANS (SD)				
Reliability	Aid info type	Non-Anthro	Low-Anthro	High-Anthro
Seventy-Two Percent	Good Advice	0.91 (.09)	0.91 (.09)	0.93 (.09)
	Bad Advice	0.65 (.27)	0.76 (.24)	0.72 (.24)
Ninety Percent	Good Advice	0.90 (.11)	0.92 (.07)	0.94 (.06)
	Bad Advice	0.48 (.35)	0.57 (.26)	0.60 (.25)

Figure 24. Proportion Correct Data



There was a significant rank order correlation ($r_s = -.38$, $p < .001$) between the behavioral time to decision data and self-reported confidence. Participants who required more time to decide were significantly less confident in the decision to accept or reject all aid advice types.

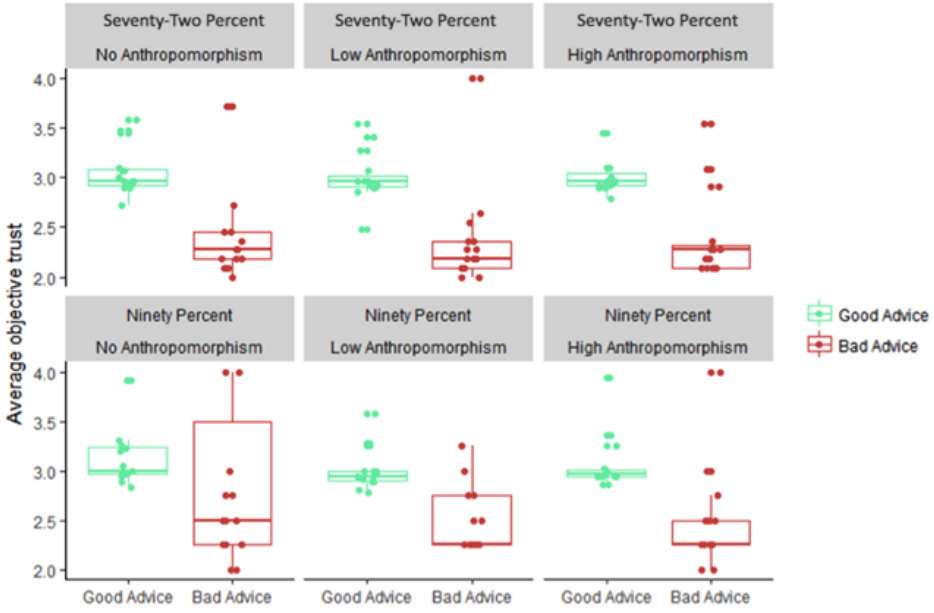
Trust. Objective trust is a behavioral measure of trust (whether participants agree, disagree or view traffic layer screen) as first used by Pak et al. Objective trust was coded on a scale from 1 to 4. If participants immediately clicked disagree that was given a value of 1 (distrust). If participants viewed the hint screen and eventually clicked disagree, that trial was assigned a 2 (moderate distrust). Viewing the hint screen and agreeing was assigned a 3 (trust but verify) and immediately clicking agree was given a value of 4 (trust).

Analyses revealed a significant main effect of advice ($F(1, 84) = 173.01$, $p < .001$) showing higher objective (behavioral) trust for good advice trials ($M = 3.06$) than for bad advice trials ($M = 2.37$). No other main effects or interactions were observed. Numerically, the non-anthropomorphic condition showed higher objective trust ($M = 2.82$) than the other conditions. This pattern was found for both the 90% reliability condition (non-anthropomorphic $M = 2.93$ compared with $M = 2.67$ and $M = 2.70$ for high and low anthropomorphic respectively) and also for the 72% reliability condition (non-anthropomorphic $M = 2.70$ compared with $M = 2.63$ for both the high and low anthropomorphic images).

Table 7. Objective Trust Summary Tables

Objective Trust Summary – MEANS (SD)				
Reliability	Aid info type	Non-Anthro	Low-Anthro	High-Anthro
Seventy-Two Percent	Good Advice	3.07 (0.63)	3.01 (0.59)	3.02 (0.54)
	Bad Advice	2.47 (0.87)	2.35 (0.70)	2.39 (0.68)
Ninety Percent	Good Advice	3.16 (0.79)	3.01 (0.49)	3.06 (0.49)
	Bad Advice	2.85 (0.99)	2.50 (0.68)	2.48 (0.75)

Figure 25. Objective Trust Data



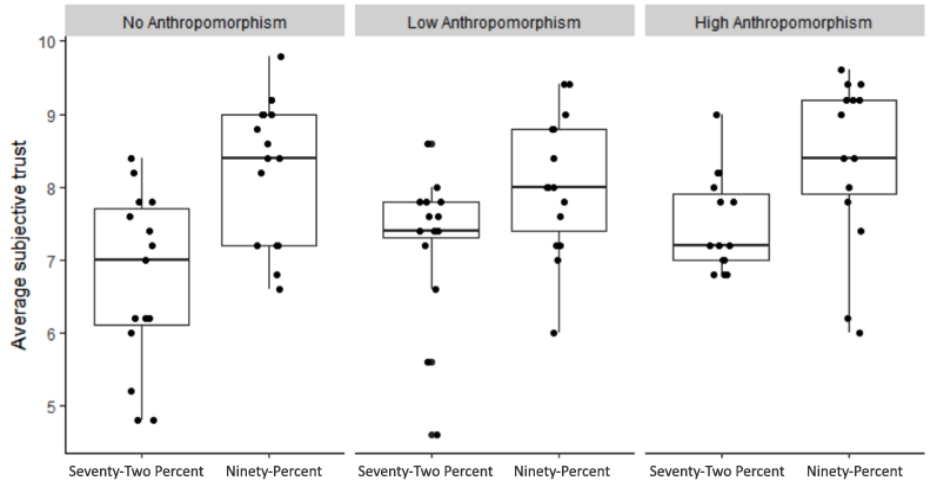
An analysis of subjective trust was conducted using a 2×3 mixed factor ANOVA, set up with a random within-subjects effect of trial. Subjective trust was measured by means of verbal self-report after completion of every eight tasks thus eliminating the aid advice factor (see Figure 25). This analysis revealed a significant effect of reliability ($F(1, 84) = 24.78$, $p < .001$). Participants in the 90% reliability condition reported

significantly higher subjective trust than participants in the 72% reliability condition. No significant effect of anthropomorphism ($F(2, 84) = 1.72$, $p = .185$) or interaction ($F(2, 84) = 1.19$, $p = .31$) were found.

Table 8. Subjective Trust Summary Table

Subjective Trust Summary – MEANS (SD)			
Reliability	Non-Anthro	Low-Anthro	High-Anthro
Seventy-Two Percent	6.72 (1.19)	7.25 (0.99)	7.52 (0.65)
Ninety Percent	8.23 (0.98)	8.04 (0.96)	8.37 (1.13)

Figure 26. Subjective Trust Data



Additionally, a correlation was conducted on the objective and subjective measures of trust revealing a significant positive correlation $r = .23$, ($t(88) = 2.26$, $p = .027$).

In summary, the null hypotheses and results are listed below:

H1 Level of information reliability alone does not influence human trust (and performance) *was accepted*

H2 Level of anthropomorphic image alone does not influence human trust (and performance) *was accepted*

H3 Type of aid advice alone does not influence human trust (and performance) *was rejected*

H4 Levels of anthropomorphic imagery and information reliability do not interact to influence human trust (and performance) *was accepted*

H5 Levels of information reliability and aid advice type do not interact to influence human trust (and performance) *was rejected*

H6 Levels of anthropomorphic imagery and aid advice type do not interact to influence human trust (and performance) *was accepted*

H7 Levels of anthropomorphic imagery, information reliability, and aid advice type all do not interact to influence human trust (and performance) *was accepted*

Discussion

The main results of this study show that without including an image of a person, participants do not significantly alter trust in an automated anthropomorphized aid even when there is a 18% difference in aid reliability. However, we found a strong performance interaction of advice type and information reliability, which will be referred to as "compliance". Compliance with bad advice significantly increased for participants in the 90% condition which supports literature in the field on the prevalence of

automation overreliance for highly accurate systems (Kantowitz et al, 1997; Lee & Moray, 1992; Muir & Moray, 1996). Overreliance is a term that suggests higher levels of compliance for bad advice due to human complacency after complying with a large sum of good advice. This finding confirms that overreliance is less prevalent for systems of lesser overall reliability.

Other interesting results are related to analysis of good and bad advice. There was significantly higher objective trust for good advice trials than for bad advice trials across all reliability and anthropomorphic conditions. Participants exhibited more frequent peeking behavior to verify the advice with a visual traffic map when exposed to bad advice. As discussed above, participants also showed high levels of compliance with bad advice. While this was especially true for the 90% reliability condition as seen in the significant interaction between advice type and reliability, it was also true in the 72% condition and across all anthropomorphic conditions. The aforementioned significant main effect of advice type for both performance and objective trust suggests that for bad advice, participants were more likely to check the traffic condition for themselves and take the automated suggested route. They were not preemptively notified of advice type nor was the exposure pattern predictable. This indicates that while people may be good at detecting bad advice (lower objective trust), they may underestimate their expertise or understanding of the situation (greater time on task and lower confidence in decision) and choose to comply with the advice anyway.

Findings for the measures of trust indicate that all levels of semi anthropomorphism as represented in this study are not held to the same perceptual and trust standards as an image of a real person. This aligns with results from a previous

study in another domain (that of medicine) that only observed a significant effect of trust when contrasting a fully non-anthropomorphic image with a photograph of a human image (Pak et al., 2012). The null results across all levels of imagery used in the present study raise some questions regarding the similarities and differences between anthropomorphism and realism as mentioned in the introduction of this paper. To explain the non-monotonic results among anthropomorphic conditions, we consider a possible perceived weirdness of the low anthropomorphic image condition. The distinctions of the image's portrayed visual etiquette (e.g., it could be perceived as waving or more animated) may account for the peaks in the data. We partially attribute poor trust findings from this study to the design of the image being object-based and loosely associated with Microsoft's Clippy, according to comments made by at least three participants in that condition. Microsoft's animated anthropomorphized paper clip helper was discontinued after a plethora of negative feedback from Microsoft Office users (Nass & Yen, 2010). Specifically, literature on trust and etiquette indicates that users tend to trust and accept software more readily when it displays personality characteristics that are similar to their own (Reeves & Nass, 1996). From this perspective, some instances of object-based anthropomorphism may not be thought of as human at all despite the inclusion of human-like features.

Upon further exploration, the trust data show somewhat higher objective trust and relatedly poorer performance under bad advice trials for the non-anthropomorphic condition when compared to high anthropomorphic condition under 72% information reliability. The non-anthropomorphic aid had higher subjective trust ratings in the 72% reliability condition. One plausible explanation for this reversed pattern is that the non-

anthropomorphic image of a cone is perceived as being more real than the high anthropomorphic depiction of a cartoon traffic cop. Additionally, an image of a human (not tested in this study) and a traffic cone may be considered as involving the same level of realism, but different levels of anthropomorphism. The disparities that exist in the trust literature across domains for incremental scales of anthropomorphism and realism point to a clear need to further understand this representational space.

There is considerable value in these findings because behavioral and self-report measures were used in the study design. However, not many researchers have used the 4-point metric of objective trust. Thus, the non-anthropomorphic condition demonstrated numerically greater objective trust in conjunction with numerically poorer behavioral performance. We realize a limitation in the 4-point objective trust as comprising a comparatively coarse-grained behavioral outcome measure.

While confirming that automation of medium and high reliability is beneficial to human performance (when compared to no automation at all), we conclude that high levels of information reliability in an automated system contribute to the previously studied overreliance problem (Kantowitz et al, 1997; Lee & Moray, 1992; Muir & Moray, 1996; Pak et al., 2012; Parasuraman et al., 2000). People are able to better identify false information and make better advice judgements in moderately reliable automation environments compared with exceedingly reliable automation environments. Current findings on trust in visually anthropomorphic aids are likely to extend across domains (e.g., from a healthcare context to navigation contexts) although more research is needed to evaluate semi anthropomorphism with full anthropomorphism in order to definitively make such a claim.

Study 2

Methods

The second study was essentially a replication of process and extension of the first study. There were three main goals to running Study 2 as designed: First, we wanted to see if the Pak et al. (2012) findings held true when removed from the health/medical domain by including both a non-anthropomorphic and full anthropomorphic image. Second, we aimed to replicate findings from Study 1 between the non-anthropomorphic and high anthropomorphic stimuli. Third, by using the same methodology as Study 1, we were able to further extend the scale of anthropomorphism to encompass non-anthropomorphic, semi anthropomorphic, and full anthropomorphic imagery. Given the consistency in procedure, stimuli, and population, we did not opt to include a control group for Study 2.

Experimental design. This study was a 3×2 mixed factor design. Independent variables included anthropomorphic aid image (3 levels: non, high, full) and advice type (2 levels: good, bad). Dependent variables were performance (task time, proportion of correct responses, confidence in decision) and trust (objective, subjective trust). Separate mixed factor ANOVAs were run to test the below null hypotheses for the two experimental dependent variables:

H1 Level of anthropomorphic image alone does not influence human trust (and performance)

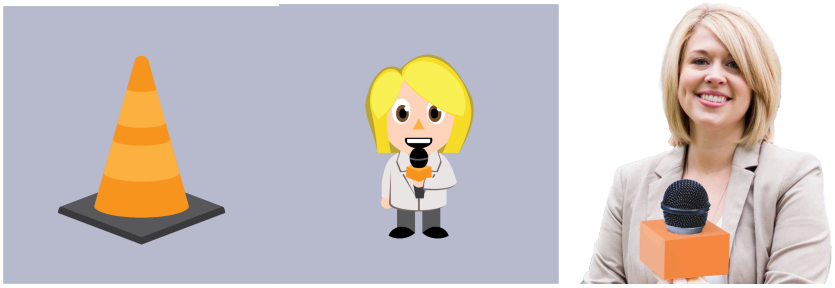
H2 Type of aid advice alone does not influence human trust (and performance)

H3 Levels of anthropomorphic imagery and aid advice type do not interact to influence human trust (and performance)

Anthropomorphism. Study 2 evaluated three anthropomorphic conditions: (a) non-anthropomorphic, (b) high anthropomorphic, and (c) full anthropomorphic. The non-anthropomorphic condition included imagery of an object relating to the domain; in this case, a traffic cone. The full anthropomorphic condition was a domain-specific photo of a female traffic reporter also considered to be domain specific. A female traffic reporter was selected in order to make the results more directly comparable with that of Pak et al. (2012). The high anthropomorphic condition was a domain-specific cartoon with a human base, visually derived from the human photo used in the full anthropomorphic condition.

We once again pilot tested the series of images for perceived levels of anthropomorphism, gender neutrality, and intelligence. Based on the results, we selected three images for non, high, and full anthropomorphic aid as depicted in Figure 27.

Figure 27. Anthropomorphic images (non, high, full)



Reliability. Results from the first study showed that an automation with 90% reliability was essentially too high to see effects for anthropomorphism or advice. The sensitivity of nuanced imagery was deemed more appropriate to study under the 72%

condition, where bad advice trials exceeded four data points. Therefore, this study kept reliability as a constant 72% effectively eliminating an entire condition from the first study. The 72% reliability condition includes 11 failures out of 40 trials where failures are an output of inaccurate decision aid information. Participants are given feedback on whether they are "on time", or "late" which also serves as route choice confirmation.

Good vs bad advice. Trials were separated into truthful information or "good advice" trials, and wrongful information or "bad advice" trials. This separation is important to note for performance reasons. Agreeing with good advice will always get the participant to the destination on time. Bad advice is when the decision aid feeds the user wrongful advice. That is, it tells the participant to take a route that is not the most efficient. Complying with wrong advice will always cause the participant to be late for that given task. Therefore, an advantageous performance outcome for bad advice trials occurs when the participant disagrees with the bad advice. The order of good and bad advice trials was determined randomly, but within a few trust building parameters following best practices for human interaction with new technology. All training trials contained good advice, the first 12 trials contained only good advice to allow for initial trust in the system, no more than two bad advice trials occurred in a row, no more than ten good advice trials occurred in a row (aside from the original 12 used to build initial trust in the system).

Dependent measures. Dependent variables were measured after each trial. Performance-based metrics included task time, proportion correct responses, and

confidence in the chosen answer. In addition, for the conditions with a decision aid, subjective and objective trust in the aid were assessed.

Participant's behavior (whether they agree, disagree, or peek) were analyzed and used as an 'objective' measure of trust. The rationale is that if participants immediately agree with the aid without peeking, that could be considered a high level of trust in the aid. However, if participants disagree without peeking, it would indicate a complete lack of trust. If participants peek before making a decision it could represent moderate levels of trust. Behavioral trust is a scale from one to four used by Pak et al (2012). If participants immediately click disagree, a value of one (no trust) will be given. If they peek and eventually click disagree, that trial will be assigned a two (moderate distrust). Peeking and agreeing will be assigned a three (trust but verify), and clicking agree assigned a four (trust).

If trust is the attitude, reliance is the behavior and can be measured using objective task performance measures. The combination of time on task and decision selection are frequently used to determine reliance. Reliance may be appropriate when operators trust automation that is either reliable or more reliable than manual operation, or inappropriate when operators trust automation that is either inaccurate or less reliable than manual operation (Dzindolet et al., 2003).

A series of surveys, including a brief demographic survey, followed the empirical testing. Substantial evidence demonstrates that trusting tendencies, considered as a personality trait, can be reliably measured and can influence behavior in a systematic manner. Rotter's Interpersonal Trust Scale (RITS) was administered to help differentiate people on their propensity to trust others. People with a high propensity to trust fared

better in predicting others' trustworthiness than those with a low propensity to trust (Kikuchi, Wantanabe, & Yamasishi, 1996). The Jian et al. trust in automation scale was used to measure subjective trust in automation for this study. Additionally, a baseline of perceived anthropomorphism was assessed by survey adapted from Epley et al. (2008). People who identify as being lonely are more likely to have a higher baseline for perceived anthropomorphism. To help sort through individual differences, a personality assessment was given. The Big Five Aspect Survey (BFAS) was used to identify some big five traits thought to be related to the decision actions and preferences of participants (i.e., neuroticism and performance).

Participants, procedure, and analytic plan. The study took place at the University of Minnesota – Twin Cities campus. A total of 47 university students participated (14 male, 33 female, average age 19.5) for course credit or cash compensation. The data from two female participants were excluded on the basis of exceeding the age requirement for the study, and city familiarity which were discovered in the demographic survey. The data were replaced with two other female participants resulting in 45 usable participants (14 male, 31 female, average age 20.5). All participants were tested individually in a single experimental session lasting about 90 minutes. Figure 13 illustrates the procedural flow. Participants in each condition were provided with a total of 40 tasks, each pertaining to a unique map of an unfamiliar metropolitan area. There were five different maps used throughout the study, with eight consecutive tasks per map. Participants received a physical map with realistic information about the area including general traffic trends. Then participants received a task notecard

with a brief written scenario such as the following: *"It's currently 8:40AM, you have an exam on campus at 9:00AM. Which route should you take?"* A practice map and four practice tasks were used as training for the experimental trials.

In all conditions except the control, the participant was directed to click through the prototype on a smartphone and decide to accept or reject the suggested route from the aid. The aid presented an accurate or inaccurate recommendation (termed "good advice" or "bad advice" respectively) to the participant paired with some level of anthropomorphic image. The participant could accept or reject the aid suggestion or choose to view more information. The "view more information" provided the participant with a visual traffic overlay for the routes which always reflected accurate traffic information for the task at hand. After accepting or rejecting, the participant was directed to verbally rate the degree of confidence they had in their decision on a scale of 1-10. The next screen on the aid then visually displayed either an "on time" or "late" confirmation message. After all eight tasks for one map were completed, participants were asked to rate their overall level of trust in the automated decision aid on a scale of 1-10. This process was repeated for the remaining four map scenarios.

Once all 40 tasks were completed, participants were asked to complete post-task surveys including but not limited to demographic information, questions concerning the perceived anthropomorphic mental state qualities (e.g., good intent) of each participant's assigned smartphone image and personality assessments. Sessions involving the smartphone were both audio and screen recorded using QuickTime and saved locally to a secured lab computer. The stimuli were created with InVision prototyping software. Researchers entered experimental data in real time as participants received feedback on

the decisions, indicated confidence levels, as well as rated trust in the automation aid. After each participant, researchers coded the time data by watching and listening to the session recordings.

Results

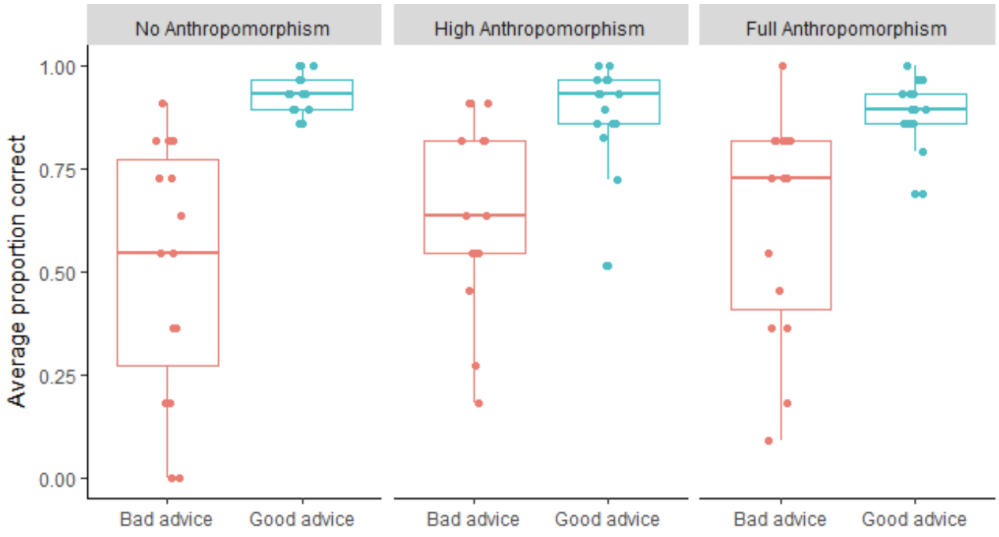
Performance. A 3 (anthropomorphism, between) ×2 (advice type, within) mixed factor ANOVA was run using the *nlme* package in R. This analysis revealed no significant interaction of anthropomorphism and advice ($F(2, 42)=2.282$, $p=.115$). There is however a slight trend towards an interaction of anthropomorphic image condition and advice type. As shown in Figure 15 and Table 9, similar performance accuracy is seen across all levels of anthropomorphic image conditions for good advice trials with means for full, high, and non-anthropomorphic at .89, .92, and .93 respectively. Comparatively poorer performance is seen for the no anthropomorphism condition than either of the anthropomorphic conditions on bad advice trials. (M = .62, .64, and .51 for high, low, and non-anthropomorphic in that order).

There was however a main effect of advice type, ($F(1, 42)=72.851$, $p<.0001$) with higher correct responses for good advice (M = .91) than for bad advice trials (M = .59). No main effect of anthropomorphic condition is seen. There are similar combined advice means for full, high, and no anthropomorphic conditions (M= .76, .78, .72 respectively).

Table 9. Performance Summary Tables

Performance Summary – MEANS (SD)				
Reliability	Aid info type	Non-Anthro	High-Anthro	Full-Anthro
Seventy-Two Percent	Good Advice	0.93 (0.05)	0.88 (0.12)	0.89 (0.08)
	Bad Advice	0.51 (0.31)	0.64 (0.23)	0.72 (0.27)

Figure 28. Proportion Correct Data



There was a significant rank order correlation ($r_s = -.370$, $p < .0001$) between the behavioral time to decision data and self-reported confidence. Participants who required more time to decide were significantly less confident in the decision to accept or reject all aid advice types.

For Study 2 we additionally ran a 3 (anthropomorphic image condition) ×2 (advice type) mixed-factor ANOVA on response time to gain a better understanding of time as a variable. Anthropomorphic image condition was a between-subjects factor and advice type was a within-subjects factor. Results showed a main effect of advice type,

$F(1, 42) = 15.16$, $p < .001$, with response time on good advice trials ($M = 10.61$) significantly faster than response time on bad advice trials ($M = 12.03$). There were no other significant effects of interactions to report, however, response time was longer in the full anthropomorphism ($M= 12.07$) and high anthropomorphism ($M = 11.55$) than in the non-anthropomorphism ($M = 10.34$).

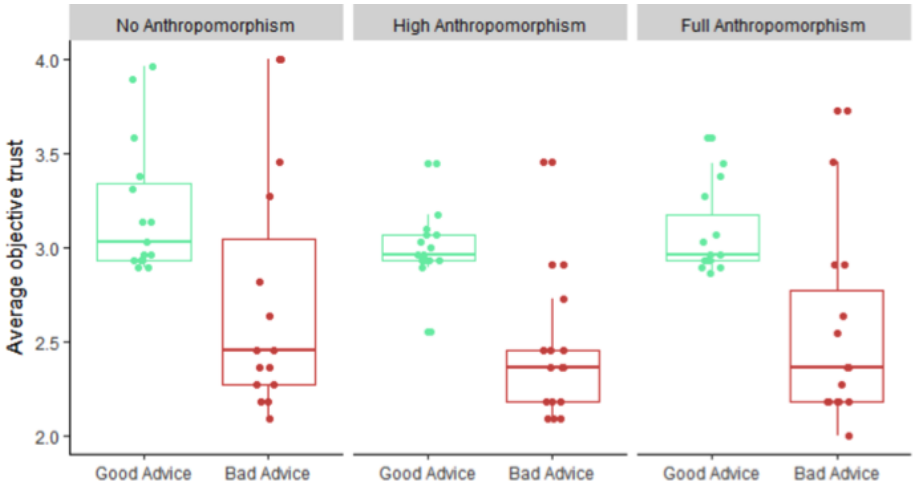
Trust. Objective trust is a behavioral measure of trust (whether participants agree, disagree or view traffic layer screen) as first used by Pak et al. Objective trust was coded on a scale from 1 to 4. If participants immediately clicked disagree that was given a value of 1 (distrust). If participants viewed the hint screen and eventually clicked disagree, that trial was assigned a 2 (moderate distrust). Viewing the hint screen and agreeing was assigned a 3 (trust but verify) and immediately clicking agree was given a value of 4 (trust).

Analyses revealed a significant main effect of advice type ($F(1, 42) = 84.821$, $p < .0001$) showing higher objective trust for good advice trials ($M = 3.09$) than for bad advice trials ($M = 2.56$). No other main effects or interactions were observed. For both good and bad advice, the non-anthropomorphic condition showed numerically higher objective trust ($M = 2.96$) than the high anthropomorphic ($M = 2.71$) or full anthropomorphic ($M = 2.81$) as seen in figure 16. Given numerical difference objective trust data, effect size was calculated to be $\omega^2 = .027$. The reported effect sizes are Omega squared which corrects for the bias inherent in Eta squared, especially with small sample sizes. Interpretation of Omega squared is: Small = .01, medium = .06, and large = .14.

Table 10. Objective Trust Summary Tables

Objective Trust Summary – MEANS (SD)				
Reliability	Aid info type	Non-Anthro	High-Anthro	Full-Anthro
Seventy-Two Percent	Good Advice	3.20 (0.60)	3.00 (0.69)	3.07 (0.71)
	Bad Advice	2.72 (0.84)	2.42 (0.90)	2.54 (0.91)

Figure 29. Objective Trust Data

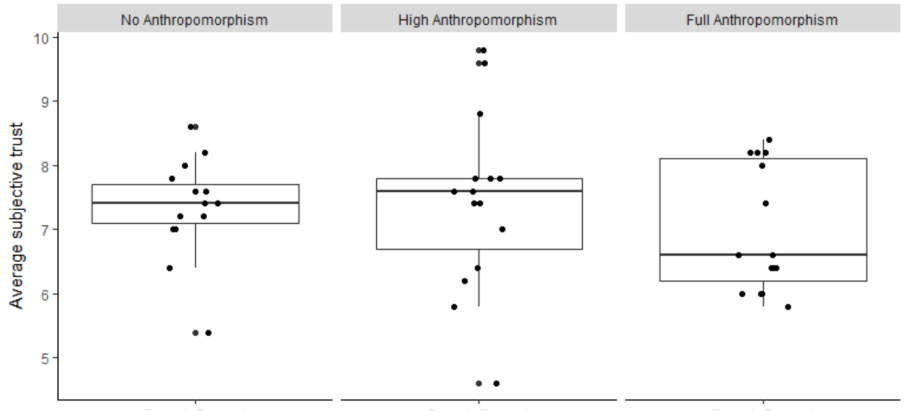


An analysis of subjective trust was conducted using a one-way ANOVA. Subjective trust was measured by means of verbal self-report after completion of every eight tasks thus eliminating the aid advice factor (see Figure 29). This analysis revealed no significant effect of Anthropomorphism ($F(2, 42) = 0.794$, $p = .459$).

Table 11. Subjective Trust Summary Table

Subjective Trust Summary – MEANS (SD)			
Reliability	Non-Anthro	Low-Anthro	High-Anthro
Seventy-Two Percent	7.44 (1.37)	7.44 (1.37)	6.97 (0.97)

Figure 30. Subjective Trust Data



Additionally, a correlation was conducted on the objective and subjective measures of trust revealing no correlation of average trust by person ($r = -0.052$) which is not significant ($t(43) = -.344$, $p = .733$).

In summary, the null hypotheses and results are listed below:

H₁ Level of anthropomorphic image alone does not influence human trust (and performance) accepted

H₂ Type of aid advice alone does not influence human trust (and performance) rejected

H₃ Levels of anthropomorphic imagery and aid advice type do not interact to influence human trust (and performance) accepted

Discussion

As previously stated, there were three main goals for this study: First, we wanted to see if the Pak et al. (2012) findings held true beyond the health and medical domain. Second, we aimed to replicate findings from Study 1 between the non-anthropomorphic

and high anthropomorphic stimuli. Third, by using the same methodology as Study 1, we are able to further extend the scale of anthropomorphism to encompass non-anthropomorphic, semi anthropomorphic, and full anthropomorphic imagery. These items will be discussed in reverse order.

The anthropomorphic scale was extended to include three categories of anthropomorphism (non, high, full) to be more comparable to the categorization of imagery used by Pak et al. (2012). Even though a slightly different high anthropomorphic image was used in this study than in the first study, results from both "replicated" group conditions (non and high anthropomorphism) showed similar results to Study 1. This not only contributes to the validity of the high anthropomorphic image group, but shows that we were consistent enough in our image creation to mitigate effects of gender bias. A non-monotonic outcome was still observed across the three groups; the full anthropomorphic condition consistently ranked between non-anthropomorphic and high anthropomorphic for both performance outcomes and objective trust. Based on results from Study 1 and now Study 2, anthropomorphism as defined by the scale used in this study may not actually matter or may be how aid imagery is conceptualized and organized by the average person.

The most impressive results statistically are again related to analysis of good and bad advice. There was significantly higher objective trust for good advice trials than for bad advice trials across all anthropomorphic conditions. Participants exhibited more frequent peeking behavior to verify the advice with a visual traffic map when exposed to bad advice. Participants also showed high levels of compliance with bad advice. A significant main effect of advice type for both performance and objective trust suggests

that for bad advice, participants were more likely to check the traffic condition for themselves and take the automated suggested route. They were not preemptively notified of advice type nor was the exposure pattern predictable. This indicates that while people may be good at detecting bad advice (lower objective trust), they may underestimate their expertise or understanding of the situation (greater time on task and lower confidence in decision) and choose to comply with the advice anyway.

Looking more closely at the data, performance outcomes were worse for the non-anthropomorphic group when exposed to bad advice than any other image condition. Additionally, objective trust and response times were highest for the non-anthropomorphic group. With system reliability no longer a factor, we are able to explore the directional effects of these other contributing variables. Results begin to paint a more holistic picture of how imagery may contribute to the overreliance problem. The data suggest that when exposed to a non-anthropomorphic image, participants seem to show evidence of overreliance by making quicker and less accurate judgments on bad advice. Alternatively, lower objective trust was seen for both anthropomorphic conditions (high and full) leading to less compliance with bad advice which is a more favorable performance outcome. Based on the trends observed in these data, anthropomorphic imagery may lead to a more appropriate calibration of trust for humans. Given its statistical insignificance, more data is needed to investigate the claim.

The claim that full anthropomorphism significantly increases trust in younger adults (Pak et al. 2012) when compared to a non-anthropomorphic image does not hold true according to the present study. While we found no significant effect of anthropomorphism, results showed tendencies for increased objective trust for the non-

anthropomorphic condition. Described here are some possible explanations for difference in findings. Pak et al. (2012) used an image of gears to represent non-anthropomorphism in their study. This is problematic for a few reasons: First, an image of gears on an automated system may be associated with a system error (Nielsen, 1995) which could have skewed the results towards lower levels of trust. Second, gears have very little to do with the medical or health care domain. In terms of perceived expertise in a medical advice system, an image of gears and an image of a doctor are not comparable. Assuming perception of expertise is in fact an influencing factor for trust (Kantowitz et al., 1997), this expertise discrepancy may have also negatively affected both subjective and objective trust scores. From this analysis we realize that expertise may carry more weight than nuances in visual anthropomorphism as an influencing factor of automated aid trust. Perceived importance of expertise between navigation and medical domains could also contribute to the difference in findings. Additionally, expertise is equally attributable to human and non-human imagery and should be controlled for in future studies.

The major takeaway from Study 2 is that participants who were presented with a non-anthropomorphic image showed (insignificant) evidence of overreliance on automated advice. This trend did not support the results of Pak et al. (2012) and more importantly showed that people may view non-anthropomorphism as being more accurate or knowledgeable than the full anthropomorphic human. Such a societal switch acknowledges the acceptance of a machine's intelligence outweighing the advice of a human expert and full anthropomorphism more similarly than what was originally theorized. Results also further support our original call to consider regrouping anthropomorphic imagery as levels of realism in Study 1.

Self Report Survey Results and Discussion

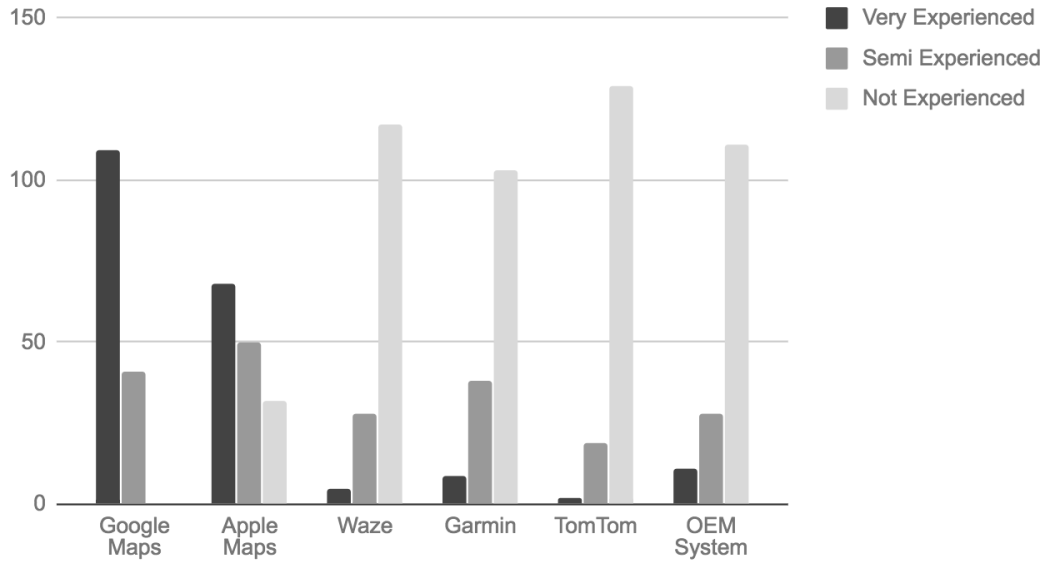
With most psychologically based research, it is hard to control for extraneous variables within the study design. There were a few individual differences variables flagged at the onset of research as possibly being so confounding that they would overshadow the manipulations. These variables included: one's propensity to trust both other people and technology, how likely one is to view an object as human-like, and personality aspects that would affect one's motivation for performance. The surveys act as an a posteriori check of random group assignment while also providing a chance to further substantiate the generalizability of our sample to the greater population by comparing results to that of other studies.

Participants filled out a series of self report surveys upon total task completion for each study. The first of which was a demographic survey that served as both a collective of previous navigation system experience and a disqualifying screener for participants who may have lied about their eligibility for the study. Of all 153 participants between both studies, only three people were disqualified; one for age, and two for familiarity driving in one or more of the cities used in the study.

Eligible participants from both studies had an average age of 20.2 with more participation from women than men; 68.7% self identified as females. All 150 participants identified as being semi experienced with a mobile navigation application. To assess the breakdown of familiarity with more commonly used navigation and routing systems in the U.S., participants rated their experience on a three-point scale. Figure 18 shows the combined results of this survey. All participants claimed some experience with Google Maps which helped to even familiarity being that the color and organization

scheme used for the map and traffic design of the studies were most similar to Google Maps.

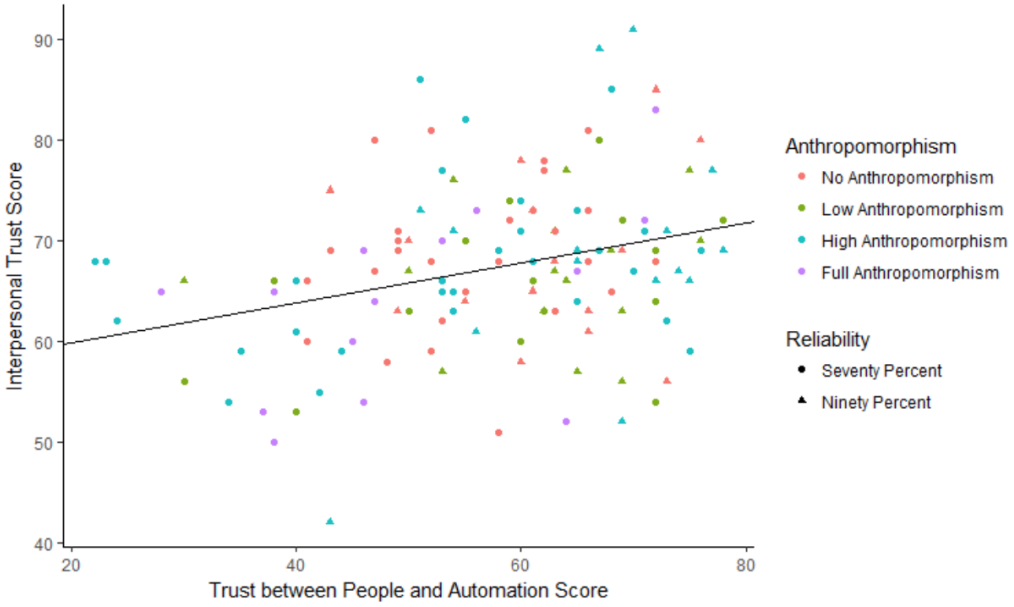
Figure 31. Navigation Product Experience



Interpersonal Trust vs Technological Trust

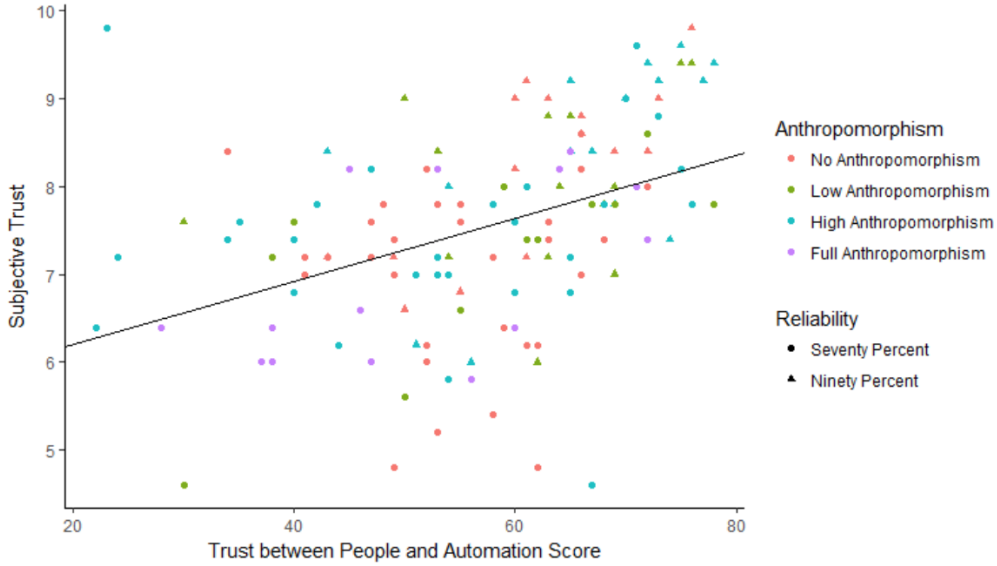
Next, Rotter's Interpersonal Trust Score (Wrightsman, 1991) and Jian et al.'s Trust Between People and Automation (TBPA) scale were administered to the 135 participants who received the smartphone aid across both studies (excluding the 15 control group participants). People with a high propensity to trust fared better in predicting others' trustworthiness than those with a low propensity to trust (Kikuchi, Wantanabe, & Yamasishi, 1996). Trust in automation theorists have shown interpersonal trust to be related but not equivalent, to trust in automation (Muir, 1989, 1994, 1996; Riley, 1994). A significant correlation between these two surveys is realized with $r = .304$, ($t(132) = 3.671$, $p < .001$) showing that people with a high propensity to trust also tend to have higher levels of trust in an automated device.

Figure 32. Trust Survey Correlation



Further analysis was conducted on the TBPA scale for both studies comparing it to the empirical measure of subjective trust throughout tasks (Figure 32). There was a significant positive correlation between the TBPA survey and verbally measured subjective trust ($r = .409$, $t(133) = 5.172$, $p < .0001$). The strong correlation shows evidence of rating consistency across self report measures used in the study.

Figure 33. Trust Survey and Subjective Trust Correlation



Anthropomorphic Mental State

An Anthropomorphic Mental State survey was adapted from Epley et al. (2008). A baseline of perceived anthropomorphism was assessed because Epley et al. (2008) found that people who identified as being lonely were more likely to have a higher baseline for perceived anthropomorphism. How people naturally attribute human characteristics to nonhuman images is important to know within the context of these studies. Based on the instructional set for this survey administration, how human-like participants think the image was used to help them form a system mental state judgement. Therefore, a significant effect of Anthropomorphic Mental State in any one anthropomorphic image condition could uncover unforeseen issues with the image design or sampling bias. A 2 (study number, between) ×2 (reliability, between) × 4 (anthropomorphism, between) ANOVA using Type II Sums of Squares was run using survey composite scores. There were no effects or interactions to report effectively ruling

out baseline anthropomorphic mental state as a confounding variable for the tasks completed.

Table 12. Anthropomorphic Mental State Summary Tables

Anthropomorphism	No	Low	High	Full
Avg Score	23.4	23.6	24.1	20.3

Anthropomorphism	No	Low	High	Full
72% Reliability	23.4	22.7	23.5	20.3
90% Reliability	23.5	24.4	25.3	

Personality

Participants completed the 100-item Big Five Aspect Scale (BFAS). It was predetermined to look for correlations between neuroticism (contributing aspects include volatility and withdrawal) and the proportion correct performance measure. Personal achievement falls under neuroticism which is consistently related to poor job performance due to its associations with low self confidence, high anxiety, hostility and vulnerability (Judge & Ilies, 2002). Given the performance-based motivation and reward structure, we felt strongly about collecting this data. Separate correlations were done for both good and bad advice trials. There was no correlation between neuroticism and performance for good advice trials ($r = -.103$, $p = .365$), or bad advice trials ($r = -.064$, $p = .573$). These null results further validate the measurement of performance through being unaffected by neuroticism.

Discussion of Measured Trust

Four measures of trust were used in Study 1 and Study 2: Objective trust, subjective trust, interpersonal trust, and trust in technology. Objective trust is the behavioral measure of needing to verifying advice through map peeking behavior during

each task. Subjective trust is the verbal rating of trust in the system's advice on a scale of 1-10 after every eight tasks. Interpersonal trust is a 24-item scale assessing one's propensity to trust other people. Trust in technology is a 12-item questionnaire that is used to assess human trust in a new technological system.

Correlation results in both studies revealed interesting relationships among the four trust measures. Interpersonal trust positively correlated with technological trust. Technological trust positively correlated with subjective trust. However, subjective trust did not significantly correlate with objective trust. The measure we call objective trust is more realistically measuring reliance. Here it is important to reestablish the difference between the terms reliance and trust. Trust is a purely psychological state while reliance is dependent on action or performance. A person can rely on automation even if it is not trusted. For example, the user can know that the automation is likely to fail but use it anyway. Frequent accounts of reliance occur when the human is physically or cognitively overloaded. The objective measure used here is also not very adaptable beyond decision aid technology as it relies on a binary decision matrix for data coding, but it is a step in the right direction.

More researchers should focus on finding valid and reliable behavioral measures for trust in automation rather than relying on various self report surveys. Assessing trust is similar to assessing cognitive workload, where the judgments occur and frequently change during a series of tasks. Asking participants to self report at the end of a task series relies too heavily on memory. The concept of objective trust has the potential to be a more accurate representation of how we make decisions to trust automated advice if it can truly be separated from performance.

Combined Empirical Analysis

In order to explore differences between anthropomorphism and realism, image conditions in both studies were retroactively reclassified in terms realism. Realism was dichotomized into non-realistic imagery and realistic imagery. Non-realistic imagery for this analysis included the androgynous high anthropomorphic cartoon cop from Study 1, and the high anthropomorphic female cartoon traffic reporter from Study 2. Realistic imagery encompassed the non-anthropomorphic cone used in both studies and the full anthropomorphic photo of a woman traffic reporter. Data from a total of 120 participants were used across both studies; 75 participants from Study 1 (excluding the control group and low anthropomorphic group), and all 45 participants from Study 2.

Reliability remained a constant at 72% (taken from Study 2) while good and bad advice were treated identically to the first two analyses. In order to rule out any confounding differences between studies (i.e., time, participant pool, etc.) that could be confused as an effect of realism, a "dummy" variable of study number was introduced into the analysis.

This yielded an unbalanced 2×2×2 mixed factor design. Independent variables included the level of image realism (between: non-realistic, realistic), study number (between: Study 1, Study 2) and advice type (within: good, bad). Because there are equal numbers of high realism in each study, A Type II Sums of Squares is used to test for realism after accounting for study number, and then test for study after accounting for realism. Dependent variables were performance (task time, proportion of correct responses, confidence in decision) and trust (objective, subjective trust). Separate mixed

factor ANOVAs were run to test the below null hypotheses for the two experimental dependent variables:

H1 Level of image realism alone does not influence human trust (and performance)

H2 Study number alone does not influence human trust (and performance)

H3 Type of aid advice alone does not influence human trust (and performance)

H4 Level of image realism and advice type do not interact to influence human trust (and performance)

H5 Study number and aid advice type do not interact to influence human trust (and performance)

H6 Level of image realism and study number do not interact to influence human trust (and performance)

H7 Level of image realism, study number, and aid advice type all do not interact to influence human trust (and performance)

Results

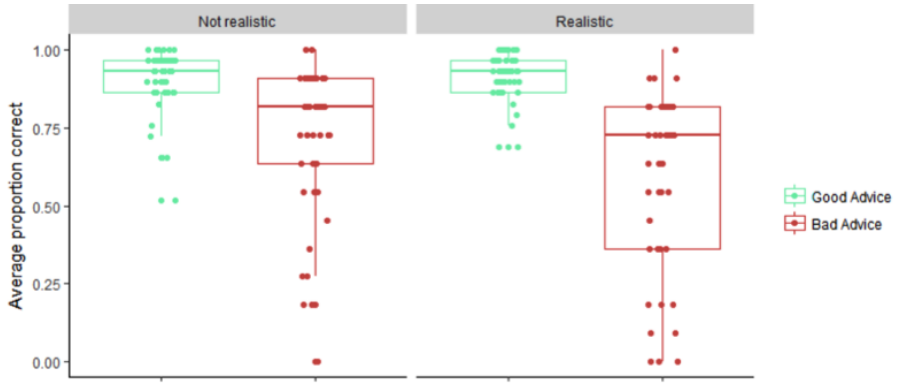
Performance. A 2 (realism, between) ×2 (study, between) ×2 (advice type, within) mixed factor ANOVA was run on proportion correct responses using the *nlme* package in R. Results indicated no significant effect of study ($F(1, 86) = 2.087$, $p = .152$) therefore all figures in this section exclude study number as a variable. No significant interactions are reported. However, there is a weak and insignificant trend towards an interaction of realism and advice ($F(1,86) = 2.696$, $p = .104$) showing poorer performance for the realistic group than for the non-realistic group ($M = .75$ and $.81$, respectively).

Again, we saw a significant main effect of advice ($F(1, 86) = 99.644$, $p < .0001$) where performance was significantly higher on good advice trials ($M = .91$) than bad advice trials ($M = .65$). No main effect of realism was discovered with similar means for non-realism and realism especially within good advice.

Table 13. Performance Summary Tables

Performance Summary – MEANS (SD)			
Reliability	Aid info type	Non-Realistic	Realistic
Seventy-Two Percent	Good Advice	0.91 (0.10)	0.91 (0.07)
	Bad Advice	0.71 (0.24)	0.59 (0.28)

Figure 34. Proportion Correct Data



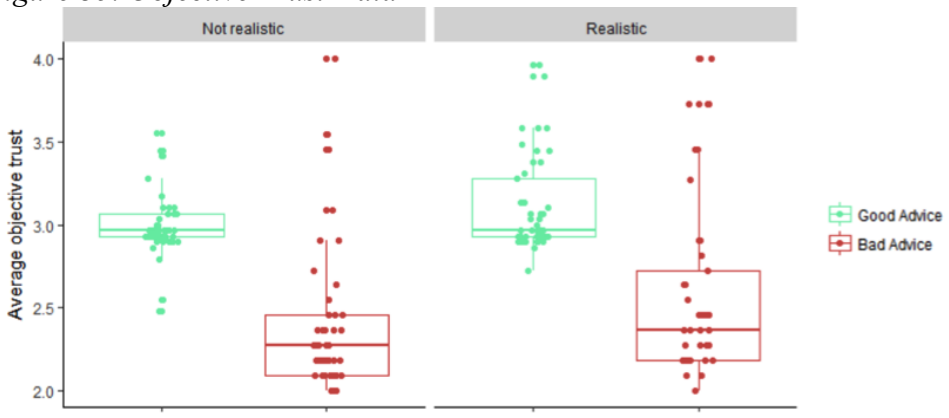
Trust. Another 2 (realism, between) ×2 (study, between) ×2 (advice type, within) mixed factor ANOVA was run for objective trust using the *nlme* package in R. Once again, it is important to first report no effect of study. Analyses revealed a significant main effect of advice ($F(1, 86) = 223.729$, $p < .0001$) showing higher objective trust for good advice trials ($M = 3.06$) than for bad advice trials ($M = 2.49$). No other main effects or interactions were observed. An insignificant trend toward a main effect of realism was

shown (F(1, 86) = 2.462, p = .120) where the realistic condition over combined advice type showed higher objective trust (M = 2.85) than the unrealistic condition (M = 2.7). Given the trending objective trust data, effect size was calculated to be $\omega^2 = .010$. The reported effect sizes are Omega squared which corrects for the bias inherent in Eta squared, especially with small sample sizes. Interpretation of Omega squared is small = .01, medium = .06, and large = .14.

Table 14. Objective Trust Summary Tables

Objective Trust Summary – MEANS (SD)			
Reliability	Aid info type	Non-Realistic	Realistic
Seventy-Two Percent	Good Advice	3.01 (0.61)	3.11 (0.65)
	Bad Advice	2.39 (0.76)	2.58 (0.88)

Figure 35. Objective Trust Data



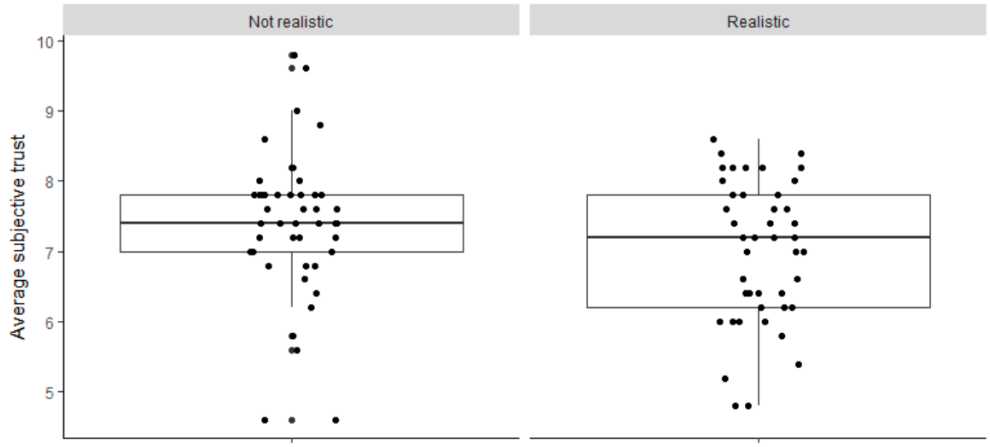
An analysis of subjective trust was conducted using a 2 (realism, between) ×2 (study, between) ANOVA. Subjective trust was measured by means of verbal self-report after completion of every eight tasks thus eliminating the aid advice factor. This analysis

revealed no main effect of study, and no interaction between study number and realism. Interestingly, there is a significant main effect of realism ($F(1, 86) = 4.410$, $p = .039$). Subjective trust is lower for realistic aids ($M = 7.01$) than non-realistic aids ($M = 7.4$).

Table 15. Subjective Trust Summary Table

Subjective Trust Summary – MEANS (SD)		
Reliability	Non-Realistic	Realistic
Seventy-Two Percent	7.40 (1.03)	7.01 (1.00)

Figure 36. Subjective Trust Data



In summary, the null hypotheses and results are listed below:

H₁ Level of image realism alone does not influence human trust (and performance)

H₂ Study number alone does not influence human trust (and performance)

H₃ Type of aid advice alone does not influence human trust (and performance)

H_4 Level of image realism and advice type do not interact to influence human trust (and performance)

H_5 Study number and aid advice type do not interact to influence human trust (and performance)

H_6 Level of image realism and study number do not interact to influence human trust (and performance)

H_7 Level of image realism, study number, and aid advice type all do not interact to influence human trust (and performance)

Discussion

The non-realistic grouping was comprised of two human-based cartoon variations. One was a gender ambiguous traffic cop, and the other was a woman traffic reporter. Homogeneity of data when combined indicated comparable treatment between the cartoon variations which could indicate a similarity in implied domain expertise. We realize that given its composition, calling the group non-realistic is not really accurate. Instead we will refer to the grouping as "cartoon". As discussed earlier, expertise as a factor may outweigh visual image treatments such as anthropomorphism and even realism as indicated by insignificant results across the board. However, with consistency in perceived expertise across image conditions we see more impressive trends when the data are organized by realism.

While still not significant when grouping images by realism, performance and objective trust results trended more closely toward significance than the results from Study 1 or Study 2. Again, we see higher objective trust in realistic imagery (cone and female reporter photo) and poor performance when given bad advice. Participants'

decisions were affected by the high levels of objective trust in the system which led to greater aid compliance during bad advice trials.

Subjective trust is also significantly lower for realistic images than non-realistic images which aligns well with the previous reports of subjective trust. Given the tendency for participants in the realistic image group to be late when given bad advice, they could have perceived higher fault in the system thus resulting in lower post task subjective trust ratings.

The higher objective trust seen in the cartoon image condition raises some questions. Do humans trust a cartoon in the same way as something that exists in the natural world? Furthermore, do humans attribute expertise with cartoons in the same way as "real" humans or images of the "real" tools they use? While inconclusive, we argue that at least in American culture, people tend to treat (and trust) cartoons in a different manner than real objects. We see evidence of this in the relationship observed between objective trust and performance. High objective trust yielded low proportion correct scores among bad advice trials (as seen with the cartoon image group), and low objective trust showed a higher performance metric for bad advice trials (seen in the realistic image group). Something is unexpected about how humans inherently trust and behave toward cartoon imagery within the context of automated navigation decision aids. The next chapter will include a summary findings throughout all three analyses along with research supported explanations for theories resulting from this body of research.

Summary and Recommendation for Future Work

The objective of this project was to expand upon the existing research on human trust in automated decision aids by evaluating the effects of mid-range anthropomorphism and information reliability on system trust and performance. The following three main research questions were addressed and conclusions will be organized accordingly:

1. What are the past and present industry practices as well as research advancements in designing for appropriate use and trust of navigation decision aids?
2. Can level of anthropomorphism affect trust for different aid information reliability?
3. How do humans behaviorally respond to visually anthropomorphized aids in a navigational context?

Limitations of the studies will be discussed along with practical implications, overall contributions, and directions for future research.

What are the past and present industry trends as well as research advancements in designing for appropriate use and trust of navigation decision aids?

Based on a side-by-side temporal comparison of in-vehicle routing and navigation systems (IRANS) and trust in automation research, there seems to be a cyclical trend in research and new in-vehicle routing and navigation systems development. First comes the development of the new technology which immediately raises societal questions further propagating by existing theoretical models. As more is understood about the applications

of the new technology, theoretical models are adapted to fit specific use cases, which are then empirically supported and become applied models. Based on results from empirical studies, changes are made for new products using similar technologies in order to improve the experience or performance. With the increasing reliability and adoption of technology, finer details within design and technology implementation are able to be researched thus extending applied models in various directions. Looking forward, as autonomous vehicles reach the general population, new models for trust in these systems containing increasing levels of automation will soon appear.

Can level of anthropomorphism affect trust for different aid information reliability?

The simple answer is no. Level of anthropomorphism does not interact with information reliability for navigation decision aid tasks. The main results of Study 1 show that without including an image of a person, participants do not significantly alter trust in an automated anthropomorphized aid even when there is a 18% difference in aid reliability. The study did confirm that automation of medium and high reliability is beneficial to human performance when compared to using no automation aid. The conclusion is that high levels of information reliability in an automated system contributes to human overreliance. This is evidenced by the ability of participants to better identify false information in moderately reliable compared with exceedingly reliable automation environments.

How do humans behaviorally respond to visually anthropomorphized aids in a navigational context?

Key takeaways are that people tend to view non-anthropomorphic imagery differently than forms of anthropomorphic imagery, and that the high anthropomorphic cartoon condition may help to better calibrate human trust in a navigation-based automated decision aid. Null results across all levels of imagery used in Study 1 and Study 2 raise some questions regarding the similarities and differences between anthropomorphism and realism. Anthropomorphism as defined by the scale used in this study may not actually matter, or it may be dependent on how aid imagery is conceptualized and organized by the average person. When grouping the data by realism, we see higher objective trust in realistic imagery (cone and female reporter photo), and poor performance when given bad advice. While still not significant, results for both performance and objective trust showed more prominent trends than the results from Study 1 or Study 2 suggesting that realism may be a better framework by which to evaluate decision aid imagery.

Future Research, Limitations, and Implications

Further research is needed to explore the effects of iconic abstraction as a visual accompaniment to information decision aids. Results of the conducted research yielded interesting data trends but no conclusive evidence that level of anthropomorphism or bifurcated realism had a significant effect on system trust. In spite of repeated pilot image perception testing, visual image aspects may have been overlooked. All images portrayed Caucasian figures that were meant to be gender ambiguous or female. A known limitation

of these studies is with the inclusion of gender and ethnic specific images. Visual styling such as gender and ethnicity have been studied in non-navigational contexts and have shown that people tend to trust images with ethnic and gender traits similar to their own. Trait matching to each participant was not a feasible solution for this study given the limited development resources. Other perceived aspects like playfulness and expertise remain generally understudied in the realm of automated decision aids. Based on these results, perceived domain expertise of an image may hold more weight than the realistic nature of the image itself, even in low risk domains such as daily routing and navigation. Image realism and expertise may not be mutually exclusive, but understanding how each impacts trust could help to focus applied efforts beyond navigation in creating personas to engender appropriate trust in AI applications for other domains.

Another limitation of this body of work is the isolation of visual from auditory manifestations of anthropomorphism in navigation which is a voice dominant field. The tasks were all set prior to actual driving making visual information a safe and appropriate option, however perceived anthropomorphism and realism in automated voice is too important to ignore. To avoid a convoluted study design and to unpack the visual aspects of personifying navigation aid advice, all aspects of automated voice remained untouched. This however provides an opportunity for researchers to evaluate similar traits such as playfulness and expertise within the auditory expression of information decision aid advice. The way in which purely auditory information is conveyed to a driver is thought to affect both attention and trust. As the attentional demands of "drivers" change with more autonomous vehicles entering the market, it will be interesting to see how

navigation decisions are communicated and whether or not there will be a shift back to visual.

The greatest benefit of this work lies in the abstraction of results into understanding how people are making decisions prior to the act of driving. This includes knowing what information is needed in order to be confident in a decision; how much stake is put in the advice provided by the decision aid; and ultimately what contributes to trusting the aid and compliance with the advice it offers. For example, communication and visual verification of the current traffic conditions, as expressed through the frequency of peeking behavior of the objective trust measure, is important to help the human understand and predict how the system is generating the advice.

As tech companies are moving towards personal assistant models for the home and vehicle, the information presented in this paper can help researchers and companies to frame decision aid-type advice in a way that can achieve the greatest level of compliance without endangering the appropriateness of trust in the system.

References

Atoyan, H., Duquet, J. R., & Robert, J. M. (2006, April). Trust in new decision aid systems. In *Proceedings of the 18th Conference on l'Interaction Homme-Machine* (pp. 115-122). ACM.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775-779.

Barber, B. 1983. The logic and limits of trust. New Brunswick, NJ: Rutgers University Press.

Barfield, W., & Dingus, T. A. (2014). *Human factors in intelligent transportation systems*. Psychology Press.

Biros, D. P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*,*13*(2), 173-189.

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

Crocoll, W. M., & Coury, B. G. (1990, October). Status or recommendation: Selecting the type of information for decision aiding. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 34, No. 19, pp. 1524-1528). SAGE Publications.

Duez, P. P., Zuliani, M. J., & Jamieson, G. A. (2006, October). Trust by design: information requirements for appropriate trust in automation. In *Proceedings of the 2006 conference of the Center for Advanced Studies on Collaborative research* (p. 9). IBM Corp.

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

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79-94.

Dunbar, B. (2015, May 05). Global Positioning System History. Retrieved September 20, 2017, from https://www.nasa.gov/directorates/heo/scan/communications/policy/GPS_History.html

Empson, R. (2013, June 11). WTF Is Waze And Why Did Google Just Pay A Billion For It? Retrieved September 20, 2017, from https://techcrunch.com/2013/06/11/behind-the-maps-whats-in-a-waze-and-why-did-google-just-pay-a-billion-for-it

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(2), 381-394.

Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. *Psychological science*, *19*(2), 114-120.

Fitts, P. M. (1951). Human engineering for an effective air-navigation and traffic-control system.

Fujii, H. (1989). The Cacs Project: How Far Away are We from the Dynamic Route Guidance System?. In *Transportation for the Future* (pp. 145-159). Springer, Berlin, Heidelberg.

Gruber, F. (2006, April 17). Comparing the Mapping Services. Retrieved September 17, 2017, from https://techcrunch.com/2006/04/17/comparing-the-mapping-services/

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

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53-71.

Jordan, N. (1963). Allocation of functions between man and machines in automated systems. *Journal of applied psychology*, *47*(3), 161.

Judge, T. A., & Ilies, R. (2002). Relationship of personality to performance motivation: a meta-analytic review. *Journal of applied psychology*, *87*(4), 797.

Kantowitz, B. H., Hanowski, R. J., & Kantowitz, S. C. (1997). Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 164-176.

Kelly, C., Boardman, M., Goillau, P. and Jeannot, E. (2001). Principles and Guidelines for the Development of Trust in Future ATM Systems: A Literature Review. *European Organisation for the Safety of Air Navigation*, 48-62.

Kikuchi, M., Wantanabe, Y., & Yamasishi, T. (1996). Judgment accuracy of other's trustworthiness and general trust: An experimental study. *Japanese Journal of Experimental Social Psychology, 37*(1), 23-36.

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.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243-1270.

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents? : the effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International journal of human- computer studies*, 64(*10*), 962-973.

Llaneras, R. E., & Singer, J. P. (2003). In-vehicle navigation systems: Interface characteristics and industry trends. In *Proceedings of the Second International*

Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design. Iowa City, IA: University of Iowa.

MacMillan, J., Entin, E. B., & Serfaty, D. (1994, October). Operator reliance on automated support for target recognition. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 38, No. 19, pp. 1285-1289). Sage CA: Los Angeles, CA: SAGE Publications.

Madhavan, P., & Phillips, R. R. (2010). Effects of computer self-efficacy and system reliability on user interaction with decision support systems. *Computers in Human Behavior*, *26*(2), 199-204.

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

McCloud, S. (1993). Understanding comics: The invisible art. *Northampton, Mass*.

Mateja, J. (1995, February 06). Oldsmobile's $1,995 Talking Map. Retrieved September 17, 2017, from http://articles.chicagotribune.com/1995-02-06/business/9502060009_1_screen-victor-ide-guidestar-navigation

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2012). I trust It, but I Don't Know Why Effects of Implicit Attitudes Toward Automation on Trust in an Automated System. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 0018720812465081.

Miller, C. A. (2005, July). Trust in adaptive automation: the role of etiquette in tuning trust via analogic and affective methods. In *Proceedings of the 1st international conference on augmented cognition* (pp. 22-27).

Muir, B. M. (1989). Operators' trust in and use of automatic controllers in a supervisory process control task. Unpublished doctoral dissertation. University of Toronto, Toronto, Ontario, Canada.

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

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429-460.

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, *7*(3), 171.

Nass, C., & Yen, C. (2010). *The man who lied to his laptop: What we can learn about ourselves from our machines*. Penguin.

Newcomb, D. (2013, April 10). From Hand-Cranked Maps to the Cloud: Charting the History of In-Car Navigation. Retrieved September 15, 2017, from https://www.wired.com/2013/04/history-in-car-navigation/

Nielsen, J. (1995). 10 usability heuristics for user interface design. *Fremont: Nielsen Norman Group.*

Norman, D. A., Ortony, A., & Russell, D. M. (2003). Affect and machine design: Lessons for the development of autonomous machines. *IBM Systems Journal*, *42*(1), 38-44.

Oleson, K. E., Billings, D. R., Kocsis, V., Chen, J. Y., & Hancock, P. A. (2011, February). Antecedents of trust in human-robot collaborations. In *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* (pp. 175-178). IEEE.

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059-1072.

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

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230-253.

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

Picard, R. W., & Picard, R. (1997). *Affective computing* (Vol. 252). Cambridge: MIT press.

Preston, B. (2013, August 05). G.M. Had a Version of OnStar in 1966. Retrieved September 15, 2017, from https://wheels.blogs.nytimes.com/2013/08/05/g-m-had-a-version-of-onstar-in-1966

Reeves, B., & Nass, C. (1996). *How people treat computers, television, and new media like real people and places* (pp. 19-36). Cambridge, UK: CSLI Publications and Cambridge university press.

Regan, M. A., Lee, J. D., & Young, K. (Eds.). (2008). *Driver distraction: Theory, effects, and mitigation*. CRC Press.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of personality and social psychology*, *49*(1), 95.

Riley, V. (1994). A theory of operator reliance on automation. *Human performance in automated systems: Current research and trends*, 8-14.

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American psychologist*, *35*(1), 1.

Sheridan, T. B. (1992). Musings on telepresence and virtual presence. *Presence: Teleoperators & Virtual Environments*, *1*(1), 120-126.

Sheridan, T. B., & Verplank, W. (1978). Human and Computer Control of Undersea Teleoperators. Cambridge, MA: Man-Machine Systems Laboratory, Department of Mechanical Engineering. *MIT*.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk analysis*, *24*(2), 311-322.

Toppen, A., & Wunderlich, K. (2003). *Travel time data collection for measurement of advanced traveler information systems accuracy*. Mitretek Systems.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system: The role of aid reliability and reliability disclosure. *Human Factors: The Journal of the Human Factors and Ergonomics Society*.

Wrightsman, L. S. (1991). Interpersonal trust and attitudes toward human nature. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of social psychological attitudes, Vol. 1. Measures of personality and social psychological attitudes* (pp. 373-412).

Yeh, M., & Wickens, C. D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, *43*(3), 355-365.