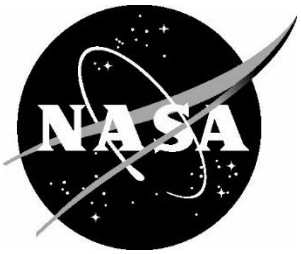
NASA/TM–20205003378



Designing and Training for Appropriate Trust in Increasingly Autonomous Advanced Air Mobility Operations: A Mental Model Approach

Version 1

Eric T. Chancey and Michael S. Politowicz
Langley Research Center, Hampton, Virginia

December 2020

NASA STI Program Report Series

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

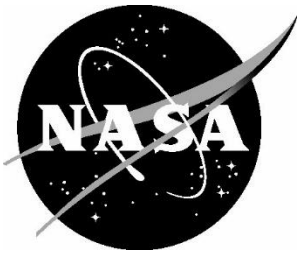
For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- Help desk contact information:

https://www.sti.nasa.gov/sti-contact-form/
and select the "General" help request type.

NASA/TM–20205003378



Designing and Training for Appropriate Trust in Increasingly Autonomous Advanced Air Mobility Operations: A Mental Model Approach

Version 1

Eric T. Chancey and Michael S. Politowicz
Langley Research Center, Hampton, Virginia

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

December 2020

Acknowledgements:

The authors would like to acknowledge the helpful comments provided by Chris Teubert, Transformative Tools and Technologies (T^3) – Autonomous Systems' Technical Lead, Vanessa Aubuchon from the T^3-Revolutionary Aviation Mobility (RAM) leadership team, and Anna Trujillo, the Human Performance and Monitoring Team Lead within the Crew Systems & Aviation Operations Branch. We would also like to thank Staci Altizer for help with technical edits. NASA's T^3-RAM Sub-Project, within the Aeronautics Research Mission Directorate's Transformative Aeronautics Concepts Program, sponsored this work.

Table of Contents

Abstract

To enable effective human-autonomy teaming (HAT) in Advanced Air Mobility (AAM) operations, the current paper presents a theoretical framework to design and train for appropriate trust in automation. The novel contribution of this work resides in connecting the construct of trust to mental models and showing how this method could be used to enable emerging HAT concepts such as Adaptive Trust Calibration. To contextualize this framework, in section 2 we discuss simplified vehicle operations (SVO) and remote vehicle operations (RVO), which are leading operational concepts within AAM. In section 3 we describe our perspective on automation and increasingly autonomous systems and present a brief discussion on human-automation interaction and human-autonomy teaming. In section 4 we provide a detailed discussion on the construct of trust in automation. In section 5 we present a framework that associates mental models with trust through principles of transparent design. Finally, in section 6 we present three descriptive models for designing and training for appropriate trust in increasingly autonomous systems.

1. Background

Advanced air mobility (AAM) represents an ecosystem of emerging aviation technologies and concepts that allows the transportation of people and goods to locations in both rural and urban environments, including those not traditionally served by current modes of air transportation (National Academies of Sciences, Engineering, and Medicine, 2020). Many of the proposed AAM concepts will be supported by increasingly autonomous systems, which will require technologies to take on more responsibilities and fundamentally alter traditional human-automation interaction paradigms. The growing reliance on higher levels of automation will necessitate research and development efforts that identify new and different ways in which humans and machines work together. Recognizing this need, NASA's Transformative Tools and Technologies – Revolutionary Aviation Mobility (T^3-RAM) Sub-project has identified Human-Autonomy Teaming (HAT) as a critical area of research required to enable safe and effective AAM operations. The notion of "teaming" between a human and machine should not focus on how machines can think or act like people, but instead on identifying capabilities and principles that facilitate humans and machines working and thinking better together (Holbrook et al., 2020). Under T^3's Autonomous Systems (AS) Enduring Discipline Area of Research, the HAT Foundational Research Activity has been tasked with providing basic research that advances the field of HAT through theory-development and experimental validation in controlled laboratory studies. An initial focus of this research activity is on trust calibration, which was identified as a key HAT research challenge by the T^3-AS HAT Planning Team (see Holbrook et al., 2020). The purpose of the current work is to introduce the theoretical perspectives of trust adopted by the HAT Foundational Research Activity and then provide an overview of future research. Although this activity is focused on foundational, basic scientific HAT development efforts, this work is geared heavily toward the advancement of AAM applications.

2. Advanced Air Mobility Concepts

The emergence of AAM has been driven largely by advances in electric and hybrid propulsion, energy storage, and increasingly autonomous software systems (National Academies of Sciences, Engineering, and Medicine, 2020). AAM broadly includes both manned and unmanned aircraft of any size with any mission, provided that they leverage the transformative technologies of the AAM ecosystem. The technical, regulatory, and economic paths of least resistance will ultimately determine major applications within AAM, but the industry is moving forward with several key interests. These include, but are not limited to, the following AAM subsets: small Unmanned Aircraft Systems (sUAS), Urban Air Mobility (UAM), Thin-Haul Commuters, and Autonomous Cargo (i.e., large UAS).

Currently, *sUAS* represent the most developed subset of AAM. sUAS (commonly referred to as drones) are typically associated with package delivery in the context of AAM, although the use cases are extensive (see Federal Aviation Administration, 2020, for overview). The concept of *UAM* represents an ambitious subset of AAM, which envisions high frequency, high density transportation of people and goods in an urban environment (Thipphavong et al., 2018). UAM is typically associated with urban passenger transport (i.e., air-taxis) in the context of AAM. *Thin-Haul Commuters* represent an existing market of small, conventional aircraft (5-9 passengers) that will leverage advances in electric propulsion and autonomy to provide reduced cost transportation to and from small cities (Moore & Goodrich, 2015). Lastly, *Autonomous Cargo* (i.e., large UAS) represents an existing market of large, conventional aircraft that will leverage advances in autonomy to shift the pilot from onboard the aircraft to a ground-based location.

Transitioning the pilot to a ground-based location will be challenging, yet all of these AAM subsets (and others that are not yet apparent) will be driven by economic forces to increase the level of automation and, thus, reduce the role of the pilot. This steady transition of duties and location (as applicable) correlates to a spectrum of pilot roles, which Chancey and Politowicz (2020) describe as *Level of Pilot-in-Command (PIC) Distance* (Table 1). This concept provides an accessible method to label the role of a human within remote operations that leverage different levels of automation across a diverse set of automated functions. Level 1 (Onboard Pilot) is the only level with the pilot onboard the aircraft. Levels 2-5 describe a remote operator with varying roles and responsibilities. These two distinguishing categories (onboard and remote) correspond to the concepts of *Simplified Vehicle Operations (SVO)* and *Remote Vehicle Operations (RVO)*, respectively, which are explained in the following sections.

Table 1. Levels of PIC Distance.

PIC Distance Level	*Description*
Level 1: Onboard Pilot	A single onboard pilot will be entirely responsible for the operation of the aircraft.
Level 2: Remote Control Pilot	There will not be a human pilot onboard the aircraft. Instead, a single ground-based, remotely located pilot will be entirely responsible for the operation of the aircraft.
Level 3: Dedicated Remote Operator	There will not be a human pilot onboard the aircraft. Instead, a single ground-based, remotely located pilot will be mostly responsible for the operation of the aircraft, with support from onboard automation.
Level 4: Remote Operator	There will not be a human pilot onboard the aircraft. Instead, a single ground-based, remotely located operator will be responsible for monitoring many automated aircraft. During emergency or challenging situations, however, that operator will have the ability to take control of that aircraft.
Level 5: System Manager	There will not be a human pilot onboard the aircraft. Instead, automation will be entirely responsible for the operation of the aircraft. However, a ground-based, remotely located operator will monitor many automated aircraft and provide situation updates (e.g., weather, traffic, winds) to the automation as needed.

Note. Adapted from "*Public Trust and Acceptance for Concepts of Remotely Operated Urban Air Mobility Transportation"* by E. T. Chancey and M. S. Politowicz, Proceedings of the Human Factors and Ergonomics Society Annual Meeting. (p. 1). *Copyright 2020 by NASA*.

2.1. Remote Vehicle Operations

Chancey and Politowicz (2020) define RVO as a concept of operations where "aircraft are remotely controlled by some combination of one or more humans piloting a single aircraft or operating/monitoring many aircraft, with varying degrees of automation support" (p. 1). This definition covers a range of remote roles (Levels 2-5) in the Level of PIC Distance concept. Remote Control Pilot (Level 2) is a single pilot controlling one aircraft remotely with full responsibility for the operation of the aircraft, which is most closely related to a military UAS remote pilot. Dedicated Remote Operator (Level 3) indicates a single human is responsible for controlling (or directing) one aircraft remotely and is similar to a current sUAS Ground Station Operator (GSO). The Dedicated Remote Operator, however, differs from the remote-control pilot in that direct control of the aircraft is significantly augmented by automation. Remote Operator (Level 4) specifies that the human operator is responsible for controlling or coordinating more than one highly automated aircraft. System Manager (Level 5) represents a shift of responsibility to the automated aircraft, which opens up the potential for a team of remote operators to manage a large number of aircraft. This role represents the end goal for AAM operations. It will be important to understand the roles, responsibilities, and technologies necessary for maximizing the ratio of aircraft to operators (see Holbrook et al., 2020).

2.2. Simplified Vehicle Operations

The General Aviation Manufacturers Association (GAMA) defines *Simplified Vehicle Operations (SVO)* as "the use of automation coupled with human factors best practices to reduce the quantity

of trained skills and knowledge that the pilot or operator of an aircraft must acquire to operate the system at the required level of operational safety" (GAMA, 2019, p. 2). The goal is to facilitate increased access to pilot certification while simultaneously maintaining safety, particularly as the demand for highly skilled pilots is expected to surpass availability within the scope of the AAM vision. This concept only applies to onboard pilots and is considered to be the transition phase between current pilot roles and future RVO roles. A similar concept, the Naturalistic Flight Deck, was explored in the context of enabling single-pilot operations for very light jets (VLJs; Schutte et al., 2007) prior to the emergence of AAM, and many of the ideas are applicable to SVO. However, there is limited research that addresses this topic as it applies to the AAM ecosystem (cf. Feary, 2018; Wing, Chancey, Politowicz, & Ballin, 2020).

3. Automation and Increasingly Autonomous Systems

Under both RVO and SVO concepts, a significant increase in automation would be required to enable the range of proposed AAM operations. In many domains, automation is implemented extensively to reduce human errors and workload, enhance efficiency, and provide economic advantages (Nickerson, 1999; Wickens, 2018). Yet to some, increasingly autonomous systems promise to surpass current automation capabilities in furthering AAM goals. This section provides a discussion to clarify the terms automation and "autonomy," and concludes with a brief discussion on human-automation interaction (HAI) and human-autonomy teaming (HAT).

3.1. A Framework for Increasingly Autonomous Systems

Hancock (2017) describes automation as a system designed to accomplish a set of largely deterministic steps to achieve a limited set of pre-defined outcomes. Alternatively, autonomy is a *characteristic* of system capabilities that "independently assume functions typically assigned to human operators, with less human intervention overall and for longer periods of time" (Pritchett, Portman, & Nolan, 2018, p. 4). Hancock (2017) also proposes that autonomous systems "are generative and learn, evolve and permanently change their functional capacities as a result of the input of operational and contextual information" (p. 284). From a technical perspective, the algorithms supporting autonomous system decisions and actions are (or would be) non-deterministic in many cases. Yet, somewhat contrasting with the word "autonomy" itself, an autonomous system may still require human supervision, direction, and cooperation (cf. RVO and PIC Distance). Clearly, the idea of an aircraft, or network of aircrafts, acting independent of any human input is neither a desired outcome, nor technically feasible (cf. Pritchett et al., 2018; Endsley, 2017). In mature AAM operations both machine and human agents will be responsible for a range of functions that will require varying degrees of inter-dependency (e.g., HAT; Pritchett et al., 2018).

Although we acknowledge fully autonomous systems represent a level of technological sophistication that surpasses many (currently all) forms of automation, we argue that it is simply a characteristic of technology and not in and of itself something wholly different from automation (see Wing et al., 2020). To this point, a "system" described as autonomous could be either a human or technology (cf. Hancock's 2017 definition above). Clearly, many in the research and development community have begun to refer to "autonomy" as a technology or set of technologies, rather than a characteristic of the technology (e.g., the term "human-autonomy teaming"). To remedy this, we adopt Parasuraman, Sheridan, and Wickens' (2000) widely accepted definition of automation, which is "a device or system that accomplishes (partially or fully) a function that was

previously, or conceivably could be, carried out (partially or fully) by a human operator" (p. 287; see also Wickens, 2018). The inclusion of "partially or fully" encompasses the notion of level of automation (LOA), which can range from fully manual (Level 1) to fully autonomous (Level 10) (Table 2). This is a more parsimonious strategy to frame conversations about "automation" versus "autonomy," than traversing the vast philosophical perspectives available in the literature. We would also argue that what is often labeled as "autonomy" or an "autonomous system" may more accurately be labeled "increasingly autonomous" (Pritchett et al., 2018) or "semi-autonomous" (Endsley, 2017). Yet, for simplicity, we do use the term "human-autonomy teaming" and in some instances simply autonomy.

Table 2. Levels of automation of decision and action selection.

HIGH	10. The computer decides everything, acts autonomously, ignoring the human.
	9. informs the human only if it, the computer, decides to
	8. informs the human only if asked, or
	7. executes automatically, then necessarily informs the human, and
	6. allows the human a restricted time to veto before automatic execution, or
	5. executes the suggestion if the human approves, or
	4. suggests one alternative
	3. narrows the selection down to a few, or
	2. The computer offers a complete set of decision/action alternatives, or
LOW	1. The computer offers no assistance: human must take all decisions and actions.

Note. Adapted from "*A Model for Types and Levels of Human Interaction with Automation" by R.* Parasuraman, T. B. Sheridan, and C. D. Wickens, IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30 (3), p. 287. *Copyright 2000 by IEEE.*

In addition to the parsimonious quality and incorporation of LOA, Parasuraman et al.'s (2000) perspective emphasizes the human's role in determining overall system performance. Specifically, the definition proposes that automation replaces, partially or fully, *functions previously carried out by a human*. On this point, Parasuraman et al.'s (2000) framework encompasses LOA's underlying inputs (i.e., information-based automation) and outputs (i.e., decision selection and action implementation automation), which are mapped to a simplified version of human information processing stages (Figure 1; see also Onnasch, Wickens, Li, & Manzey, 2014). Stage 1, sensory processing, corresponds to information acquisition automation, which augments or replaces aspects of human selective attention and sensors (e.g., eyes, ears, skin), by selecting, registering, and filtering input data. Stage 2, perception and working memory, corresponds to information analysis automation, which augments or replaces cognitive processes used to integrate information, assess situations, and provide diagnoses. Stage 3, decision making, corresponds to decision selection automation, which augments or replaces cognitive processes associated with deciding among alternatives and selecting appropriate actions. Stage 3 automation departs from information analysis by making assumptions about the costs and values of the decision impact, in a probabilistic and uncertain environment. Stage 4, response execution corresponds to control and action execution automation. Generally, Stage 4 automation replaces human actions and manual control (e.g., hand, foot, voice), to some degree. Although some forms of automation may represent a single stage, automated systems may incorporate more than one stage at various LOAs.

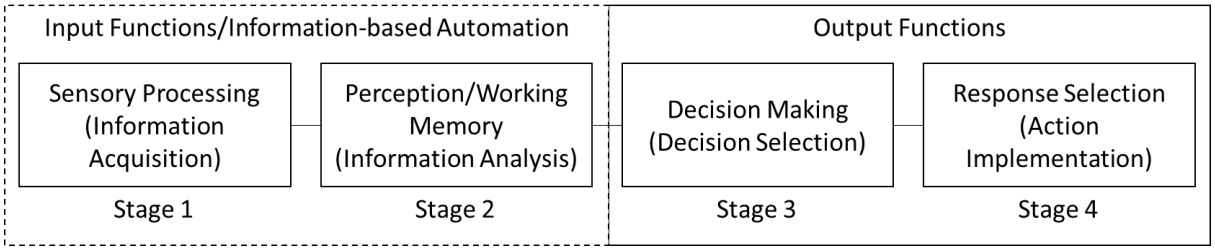


Figure 1. Simplified model of human information processing system mapped to stages of automation, based on Parasuraman et al. (2000). *Note*: System functions automated by processing stage are in parentheses.

The purpose of adopting this framework is to provide a method to structure human-automation interaction and teaming discussions and is not intended to serve as a MABA-MABA list (Men-Are-Better-At/Machines-Are-Better-At; e.g., Fitts, 1951[1]). Indeed, MABA-MABA lists can give the impression that technology and humans have fixed strengths and weakness, which provides a plausibly oversimplified method to base function allocation decisions on (Dekker & Woods, 2002; though see de Winter & Dodou, 2014, for alternative perspective). A lesson that is repeatedly reported (and often relearned) is that automation does not simply replace the human, it changes the human's tasks in unexpected and often unintended ways (Parasuraman et al., 2000). No matter the function, even highly automated system performance cannot be well predicted by the functionality of the technology alone.

3.2. Lessons from Human-Automation Interaction

The HAI literature is replete with examples that illustrate the "pitfalls of automation" (see Lee & Seppelt, 2012, for review). One such pitfall is *out-of-the-loop unfamiliarity*, where, because the task is highly automated, a human operator has a diminished ability to detect automation failures and effectively intervene in a timely manner (Endsley & Kris, 1995). Failures to detect automation errors have been attributed to lack of feedback from passive monitoring (Lee & Seppelt, 2012), vigilance (Warm, Parasuraman, & Matthews, 2008), inadequate situation awareness (Endsley & Kris, 1995), and complacency (Parasuraman & Manzey, 2010). There is a clear desire in the AAM community to adopt increasingly autonomous systems (e.g., Holden & Goel, 2016), yet this strategy runs the risk of introducing out-of-the-loop unfamiliarity issues encountered in similarly highly automated domains. Indeed, although referring to highly automated ground vehicles, Hancock's (2019) maxim holds in the context of SVO as well: "*If you build vehicles where drivers are rarely required to respond, then they will rarely respond when required.*" (p. 485). A related pitfall is *clumsy automation*, which refers to automation that tends to make easy tasks easier and hard tasks harder. For example, flight management systems (FMS) tend to make the low-workload phase of flight easier (straight/level flight, routine climb), whereas the high workload phases tend to be more difficult, such as preparation for landing, where pilots must share time among landing procedures, air traffic control (ATC) communications, and programing the FMS (Lee & Seppelt, 2012). Because the easy task is highly automated, the operator has a diminished ability to respond effectively in off-nominal/difficult situations (i.e., out-of-the-loop unfamiliarity) and impoverished skills and lack of experience to respond appropriately. Often these pitfalls are the

[1] Note: Fitts et al. presented this list as a general guideline, rather than the "gospel" of function allocation, for which it has been the target of various criticisms across the decades (see Sheridan, 2000, and de Winter & Dodou, 2014 for discussions).

result of having the ability to automate some functions, leaving the human with the "leftover" aspects of the tasks that were too difficult to automate or too challenging to certify. These issues are indicative of a "machine-centered" approach to system design that often neglects the human at the expense of safety, sometimes excising the capacity for resilient system performance (i.e., the human).

3.3. Human-Automation Teaming and Resilient Performance

The purpose of reviewing the lessons learned from HAI is not intended to cast blame on past design strategies. Noticeably, there is a tendency in the human factors and ergonomics community to sometimes exaggerate the *pitfalls of automation* narrative, without recognizing that increased automation, in many domains, has delivered on the promises of increased efficiency, safety, and economic advantages (de Winter, 2019). Similarly, however, the human's capacity to significantly increase the likelihood of resilient performance should also be recognized. Resiliency is defined as follows:

> *Resilience* is the systematic capacity to change as a result of circumstances that push the system beyond the boundaries of its competence envelope. The system may have to amend some, or even all of its goals, procedures, resources, roles, or responsibilities. As a result of those changes, the work system then expresses a revised competence envelope. In effect, it becomes a different system (Hoffman & Hancock, 2017, pp. 565-566).

Clearly, the ability to achieve resilient performance should be a goal in AAM operations. Holbrook et al. (2019) highlight the numerous ways that humans greatly increase the probability for resilient performance and provide a compelling analysis of *what goes right* in daily "nominal" civil aviation operations because of consistent human interventions in highly automated procedures. Proposed AAM design concepts should support the ability for a system (used broadly to include both technology and human components) to monitor, respond, learn, and anticipate (Hollnagel, 2015), and those abilities should manifest from a thoughtful analysis for how humans and technology can work and think better together (Holbrook et al., 2020). An emerging concept that embodies this perspective is HAT.

The concept of HAT adopts the perspective that the benefits of increasingly autonomous technologies will more likely manifest when humans and technologies partner as a *team*. Here a team is defined as "a distinguishable set of two or more *agents* who interact, dynamically, interdependently, and adaptively toward a common and valued goal/objective/mission" (Salas, Dickinson, Converse, & Tannenbauem, 1992, p. 7; the term humans was changed to agents in this definition). The progression from lower levels of automation to increasingly autonomous systems illustrates the capability to fundamentally shift human-automation pairings from interactions to teaming (i.e., HAI to HAT). Whereas low-level automation is designed to accomplish pre-specified steps to achieve a limited set of outcomes, increasingly autonomous systems are characterized by the ability to independently assume functions with less human intervention overall and for longer periods of time (Pritchett et al., 2018). It is these increasingly sophisticated characteristics that may allow more complex interpersonal teaming principles, particularly those related to trust, to become applicable to human-automation partnerships, which is often absent in simple interactions (e.g., it is difficult to imagine how a pilot is teaming with an FMS in any meaningful way). This is not to dismiss the calls for collaborative and complementary pairings between humans and automation that have been issued over the years (e.g., Jordan, 1963; Dekker

& Woods 2002). Instead, the HAT concept opens the human-automation trade-space to incorporate more sophisticated pairing strategies. One such concept that has received much attention in the HAI domain is human-automation trust. The concept of HAT, however, broadens the usefulness of the trust construct to novel and interesting teaming applications (particularly aspects of the construct that were previously only applicable to interpersonal relationships).

4. Trust in Automation and Increasingly Autonomous Systems

Over the past several decades, researchers have invoked the construct of trust to predict and describe human interactions, and more recently teaming, with various technologies (e.g., Sheridan & Verplank, 1978; Sheridan, Fischhoff, Posner, & Pew, 1983; Sheridan & Hennessy, 1984; Muir, 1987; Sherridan, 1988; Lee & See, 2004; Hoff & Bashir, 2015; Sheridan, 2019ab; de Visser et al., 2019). Muir (1987, 1994; Muir & Moray, 1996) provided one of the first formal attempts to model human-automation trust. She proposed a two-dimensional framework to study "human-machine" relationships based on taxonomies of interpersonal trust (i.e., Barber, 1983; Remple, Holmes, & Zanna, 1985), which led to a multitude of theoretical perspectives (Table 3; see Adams, Bruyn, Houde, & Angelopoulous, 2003, for extensive review of early human-automation trust research and Hoff & Bashir, 2015, for recent review). Among these perspectives, Lee and See's (2004) human-automation trust model has emerged as the most influential and widely accepted. As of March 2020, a Google Scholar search shows the article has 2,359 citations. Because of the comprehensive and integrative merits of Lee and See's (2004) model, it serves as the main organizing theoretical lens we use to describe the nature of trust in the current work. Moreover, recent theoretical developments have built off of this model to offer a greater understanding of factors that affect trust (i.e., Hoff & Bashir, 2015), the trust calibration process for interacting and teaming with increasingly autonomous systems (i.e., de Visser, et al., 2019), and theoretically-grounded strategies for transparent design (i.e., Chen, Procci, Boyce, Wright, Garcia, & Barnse, 2014; Lyons, 2013). This section outlines our conceptualization of trust by first providing a definition, a description of appropriate (calibrated) trust, and then describes the trust formation process.

Table 3. Selected Influential Human-Automation Trust Theories

Trust Theory References	Description
Sheridan (1988, 2019a, 2019b)	Proposed several quantitative human-automation trust models for supervisory controls (i.e., Signal detection, statistical parameter estimate, model-based control) and the notion of applying attributes of human morality to affective trust criteria for increasingly intelligent automation
Muir (1987, 1994; Muir & Moray, 1996)	Proposed a framework that integrated bases of trust (persistence, technical competence, and responsibility) and dynamics of trust (predictability, dependability, and faith).
Lee and Moray (1992; 1994)	Proposed modified version of Muir's framework, which added leap of faith, understanding, and trial-and-error experience (Zuboff, 1988). Related this updated framework to the concepts of purpose, process, and performance.
Parasuraman and Riley (1997; Riley, 1996)	Cited trust as one of the key components in determining automation use, along with other variables such as workload, perceived risk, and self-confidence.
Cohen, Parasuraman, and Freeman (1998)	Proposed the Argument-based Probabilistic Trust (APT) model, which introduced the use of event-trees that probabilistically model decisions to determine automation dependence.
Seong and Bisantz (2000; Seong, Bisantz, & Gattie, 2006)	Proposed a trust model based on Brunswik's (1952) Lens model, which accounts for trust calibration.
Dzindolet, Pierce, Beck, Dawe, and Anderson (2001), Dzindolet, Pierce, Beck, and Dawe (2002)	Proposed a conceptual model of automation use, which cited trust as a key component. Loosely based on concepts proposed by Parasuraman and Riley (1997).
Lee and See (2004)	Proposed a qualitative model that specified how to design trustable automation and presented a review of both interpersonal and human-automation trust theories.
Madhavan and Wiegmann (2007)	Proposed a model of sequential development of trust for automation and humans and, additionally, a framework of factors that affect the development of trust in automation.
Hoff and Bashir (2015)	Building off Lee and See (2004), proposed a three-layer trust model consisting of dispositional, situational, and learned trust.
Hoffman (2017)	Proposed a cognitive systems engineering taxonomy of emergent trust in human-machine relationships
de Visser et al. (2019)	Proposed an iterative human-agent trust model; introduced the concept of *relationship equity* to the human-technology paradigm

Note: References ordered chronologically by first publication.

4.1. Defining Trust

One of the ongoing struggles of trust research (interpersonal and human-automation) is establishing the appropriate characterization of the construct. Trust has been conceptualized as a belief (e.g., Kramer, 1999), an attitude (e.g., Barber, 1983), an intention (e.g., Mayer, Davis, & Schoorman, 1995; Schoorman, Mayer, & Davis, 2007), and as a behavior (e.g., Deutsch, 1960; Meyer, 2001). To resolve these conflicting perspectives, Ajzen and Fishbein (1977, 1980) propose a process where attitudes toward a target object are based upon beliefs about that target, which then leads to the adoption of an intention to act out a behavior toward that target. Specifically, a *belief* provides the basis for attitudes, which are formed by experiences, knowledge, and the availability of information. An *attitude* is "a learned predisposition to respond in a consistently favorable or unfavorable manner with respect to a given object" (Fishbein & Ajzen, 1975, p. 6). Lee and See (2004) describe attitudes as "affective evaluations of beliefs that guides people to adopt a particular intention" (p. 53). Regardless, attitudes develop from the beliefs about an object by associating it with particular attributes (Ajzen, 1991). *Intentions*, based on attitudes, lead to behaviors. Intentions are regulated by environmental and cognitive variables and capture the motivational factors that influence behavior. In other words, intentions indicate how much effort an individual is willing to exert to execute a behavior (i.e., the stronger the intention, the more likely the performance of that behavior is likely to occur; Ajzen, 1991). Studies have shown that intentions to perform an action, for example intention to use technology, are predictive of actual technology use (e.g., Davis, Bagozzi, & Warshaw, 1989; Venkatesh, Morris, Davis, & Davis, 2003).

Lee and See (2004) propose that trust is best conceptualized as an attitude, where beliefs about the characteristics of the automation help form the basis for adopting a particular level of trust (see Mental Model section). Depending on the level of trust, this leads a person to adopt an intention that leads to a behavior. Considering trust as a behavior or intention has the potential to confound its effects with other variables that likely affect behaviors (e.g., environmental and task constraints, workload, self-confidence). From this perspective, there is a clear distinction between trust as an attitude and behavioral responses.

Lee and See (2004) also highlight two important components of trust. First, the trustee is responsible for advancing the goal(s) of the trustor. Although most perspectives of trust do not explicitly include this component, most highlight the importance of allowing the trustee to perform a particular action on behalf of the trustor (cf. Mayer et al., 1995). This is particularly important for human-automation trust, as the trustee is often created to achieve the goal(s) of the trustor. Second, a common theme among most conceptualizations of trust is the notion of vulnerability and perceived risk, where the trustor willingly assumes risk by delegating responsibility to the trustee (Lyons & Stokes, 2012; Mayer et al., 1995). If a trustor does not perceive the risk associated with placing a trustee (e.g., human, automation) in charge of achieving their goal(s), then trust will not greatly affect intentions (Chancey, 2020) or behaviors (Chancey, Bliss, Yamani, & Handley, 2017). Risk is a characteristic of decisions and is defined as "the extent to which there is uncertainty about whether potentially significant and/or disappointing outcomes of decisions will be realized" (Sitkin & Pablo, 1992, p. 10). Reflecting these perspectives, trust is "an attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p. 51), where the perceived risk of being vulnerable to that agent determines if trust is translated into a behavior (Chancey et al., 2017; Mayer et al., 1995).

4.2. Trust Calibration

Trust is often described in terms of calibration, or degree of appropriateness. *Calibration* describes the relationship between trust in the system and the actual trustworthiness (or capabilities) of the system (i.e., the diagonal line in Figure 2). From the interpersonal trust literature, Mayer et al. (1995) define trustworthiness as the extent to which a trustee has the *ability* (skills, competencies in a specific domain) to achieve the goals of the trustor, is *benevolent* (positively oriented toward the trustor), and has *integrity* (adheres to a set of principles acceptable to the trustee). Much work has been conducted in operationalizing and explaining trust calibration (e.g., de Visser et al., 2019; Lee & See, 2004; McBride & Morgan, 2010; Okamura & Yamada, 2020; Seong & Bisantz, 2000; Seong, Bisantz, & Gattie, 2006; Wang, Pynadath, & Hill, 2016). Operators demonstrate poor trust calibration by over trusting the system (i.e., trusting it above its capabilities, generating misuse), or under trusting the system (i.e., trusting the system below its capabilities, generating disuse) (cf. Parasuraman & Riley, 1997; Parasuraman, Sheridan, & Wickens, 2008). The importance of trust appropriateness cannot be overstated, as the design goals for automation should not necessarily be to instill excessive trust (regardless of the rigor undertaken in the verification and validation process).

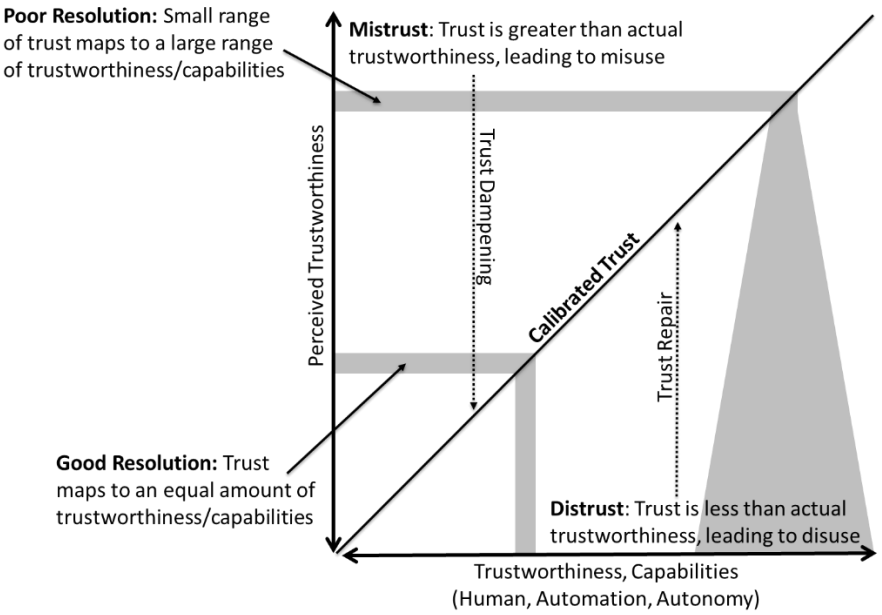


Figure 2. Model of trust calibration and resolution between trust and trustworthiness – Adaptive from Lee and See (2004), de Visser et al. (2019), and Gempler (1999).

Users often assume that "expert" automated systems will work properly and engage in other activities without worrying about the system making an error that will go undetected. Over-trust is frequently cited as one of the key contributing factors in misuse of automation (Lee & Seppelt, 2012). Unfortunately, over-trust can lead to disastrous outcomes when rare automation failures leave the human out-of-the-loop and unprepared to intervene or takeover (Parasuraman & Manzey, 2010). There are many examples within the aviation domain that point to over-trust in highly automated tasks that have led to near misses, incidents, and accidents (Bliss, 2003a). More recently, misuse of automation is increasingly appearing in the personal vehicle domain. As an example, several crashes have resulted from disengaged drivers not taking over for highly

automated vehicles failing to detect and avoid parked firetrucks (Stewart, 2018; Figure 3). Yet, similar to aviation examples, some drivers seemingly trust these technologies to the point of literally sleeping behind the wheel[2,3,4,5]. Here we remind the reader of the maxim introduced in the previous section: "*If you build vehicles where drivers are rarely required to respond, then they will rarely respond when required*" (Hancock, 2019; p. 485).

Figure 3. Car on "Autopilot" collided with parked firetruck. Image from Culver City Firefighters [CC_Firefighters]. (2018, January 22). While working a freeway accident this morning, Engine 42 was struck by a #Tesla traveling at 65 mph. The driver reports the vehicle was on autopilot. Amazingly there were no injuries! Please stay alert while driving! #abc7eyewitness #ktla #CulverCity #distracteddriving [Tweet]. Retrieved from https://twitter.com/CC_Firefighters/status/955529991319560192

Alternatively, the full potential and benefits of automation will not be realized if the human does not trust the technology. Sorkin (1988) noted that pilots ignore or disable unreliable, yet critical, alarm systems (e.g., "Some military pilots I know admit to witnessing the removal of circuit breakers so as to disable flight warning systems" p. 1107). Research has shown that even the perception of automation performance also affects trust and responses, regardless of actual performance characteristics (Bliss, Dunn, & Fuller, 1995). In a recent study on trust in the Automatic Ground Collision Avoidance System (Auto-GCAS), pilots conveyed accounts from

[2] ABC News (2019). Driver asleep at the wheel of his Tesla on busy freeway in Los Angeles. [YouTube]. Retrieved February 28, 2020, from https://www.youtube.com/watch?v=ZhObsMnipS8
[3] NBC News (2019). Tesla Driver Caught on Camera Apparently Asleep At The Wheel | NBC Nightly News. [YouTube]. Retrieved February 28, 2020, from https://www.youtube.com/watch?v=NHUZxeSUFUk
[4] CBS 17 (2019). Video appears Tesla driver asleep at the wheel of freeway. [YouTube]. Retrieved February 28, 2020, from https://www.youtube.com/watch?v=BjHpBFPz2xI
[5]KPIX CBS SF Bay Area (2018). Police Say Tesla Driver Was Asleep at Wheel With Autopilot On. [YouTube]. Retrieved February 28, 2020, from https://www.youtube.com/watch?v=wpsPPbnZxq4

other pilots of system failures resulting in crashes, even though Auto-GCAS had not been installed on the crashed aircrafts in question (Ho, Sadler, Hoffmann, Lyons, & Johnson, 2017).

A system that is not trusted, and subsequently ignored or disabled, represents a waste of invested resources (time, money) and ultimately has minimal impact on solving the issue it was designed to address, regardless of actual capability. Alternatively, if a system is trusted beyond its capabilities, when automation fails or autonomy behaves unpredictably, the human will not likely be able to effectively takeover (i.e., out-of-the-loop unfamiliarity, skill decay, inadequate training or experience) or will simply not notice system problems. The goal of HAI and HAT efforts should be to establish design and training strategies for appropriately calibrated trust. The notion of operationalizing appropriate trust, however, deserves some consideration.

4.2.1. Operationalizing Trust Calibration

Although we have characterized trust as an attitude, trust calibration has often been used to explain automation response behaviors (e.g., compliance, reliance, agreement rate). Two common behaviors associated with trust calibration are probability matching (e.g., Bliss, Gilson, & Deaton, 1995; Wiegmann, Rich, & Zhang, 2001) and system monitoring (e.g., Bailey & Scerbo, 2007). Although both of these response behaviors have been the focus of a considerable number of studies, neither method offers a perfect approximation of trust calibration. The emerging concept of Signal Detection Theory for Social Trust Calibration is a promising alternative to these methods (de Visser et al., 2019).

Probability Matching. Probability Matching describes a behavioral response pattern in which humans tend to match their agreement rates with the expected reliability of automated decision aids and alarm systems (Bliss, Gilson, et al., 1995; Manzey, Gerard, & Wiczorek, 2014; Wiegmann et al., 2001). From a functional perspective, reliability is defined by the number of errors (false alarms and misses) that an automated aid produces during a given time period (e.g., an alarm system that produces 1 false alarm every 10 times the alarm sounds is 90% reliable; Sullivan, Tsimhoni, & Bogard, 2008). Theoretically, probability matching occurs as a result of the human calibrating their trust with the perceived reliability of the system (i.e., perceived trustworthiness). To illustrate, Bliss et al. (1995) reported that the number of agreements with the output of an alarm system tended to approximate its reliability (e.g., in the 75% reliability group, participants agreed with 75% of the alarms). In a literature review, Bliss (2003b) noted that participants tended to probability match when provided with information that could be used to crosscheck automation output (i.e., there is a degree of transparent design). Without the ability to crosscheck the automation, however, participants tend to maximize their agreements with the automation (see also Manzey et al., 2014 for review). Explicitly disclosing the reliability of the automation can also affect the calibration of agreement rates. Wang, Jamieson, and Hollands (2009) showed that when participants were explicitly informed of the reliability of an automated decision aid, they tended to vary their agreements more accurately than those not provided with reliability information. Regardless of actual system reliability, perceived reliability can also significantly affect response strategies. Bliss et al. (1995) reported a study in which participants interacted with a 50% reliable signaling system across two sessions. Before beginning the second session, an experimental confederate falsely informed participants that the signaling system was 75% reliable, which resulted in an increased number of agreements during the subsequent session (also see Chancey & Bliss, 2012).

Potential Issues. Although probability matching has been used as an approximation of trust calibration, it can result in poor performance (not the desired effect of well-calibrated trust). To illustrate, if a system is 80% reliable, then across 100 signals or decisions issued by the automation, the human will agree with the automation 80% of the time. This will result in approximately 64 correct agreements (i.e., $0.8 \times 80 = 64$). For the remaining 20 responses, the number of correct disagreements would be approximately 4 (i.e., $0.2 \times 20 = 4$), resulting in a total of 68 correct responses. Instead, if the operator had agreed with the system every time, the human-automation pairing would have arrived at 80 correct responses (i.e., $0.8 \times 100 = 80$ correct; Wiegmann et al., 2001). From this example, if depending entirely on operationalizing trust as a response behavior, it would appear that the design recommendation would be to encourage *over trust* in the automation.

System Monitoring. Following the notion of crosschecking automation to evaluate its trustworthiness, monitoring strategies may also provide insight into operationalizing trust calibration (e.g., eye tracking). System monitors are described as complacent (i.e., monitoring the automation less than an "optimal observer"), skeptical (i.e., monitoring the automation more than an "optimal observer"), or eutactic (i.e., "well calibrated"/optimal)(Moray & Inagaki, 1999). Complacency due to over trust has received a great deal of attention in the research community (see Parasuraman & Manzey, 2010, for review). Research has shown that in system monitoring tasks, a "complacent" human monitor often fails to intervene or detect a rare failure by a highly automated system (e.g., Politowicz, Chancey, & Glaab, 2021; Prinzel, DeVries, Freeman, & Mikulka, 2001). It does make practical sense that if a human does not take the time to crosscheck the automation to ensure it is correct, then that may plausibly indicate a high level of trust.

Potential Issues. To determine if a monitor is complacent, an "optimal" monitoring strategy needs to be operationally defined to contextually establish what is above and below eutactic/well calibrated monitoring (Parasuraman & Manzey, 2010). Bahner, Huper, and Manzey (2008) were able to establish an optimal sampling rate by cleverly incorporating a monitoring option into a micro-world experiment. Moray and Inagaki (2000) proposed using the Nyquist sampling theorem to mathematically establish an optimal sampling rule (i.e., "sample a variable whose bandwidth is *W*Hz at 2*W*Hz," Moray, 2003, p. 176). Yet, both approaches may be difficult to test outside of tightly controlled laboratory studies (cf. Parasuraman & Manzey, 2010). Parasuraman and Manzey (2010) suggest that models of visual attention may be useful in explaining monitoring strategies (e.g., the Salience, Effort, Expectancy, Value [SEEV] Model, Wickens & McCarely, 2008, pp. 41-61), and some researchers are beginning to use eye tracking as an indirect measure of trust (e.g., Hergeth, Lorenz, Vilimek, & Krems, 2016; Karpinsky, Chancey, Palmer, & Yamani, 2018; Karpinsky, Chancey, & Yamani, 2016a, 2016b;).

Signal Detection Model for Social Trust Calibration. de Visser et al. (2019) recently proposed a signal detection model for social trust calibration, which is both novel and highly applicable to the HAT paradigm. de Visser et al.'s (2019) model focuses on maintaining relationship equity, an emotional resource that predicts the degree of goodwill between humans and automation. The authors introduce methods for *trust repair* and *trust dampening*, which serve to maintain calibrated trust between a human and an autonomous agent (see Figures 2 and 4). The autonomous agent maintains trust calibration through anticipated and unanticipated trust violations. Maintaining well-calibrated trust is both important, and potentially difficult to operationalize (see **Probability**

Matching and **System Monitoring** sections above). This method, however, provides a compelling mechanism to regulate trust calibration for HAT, which also reaches beyond somewhat contrived HAI research paradigms. Moreover, because this approach is grounded in signal detection theory, it is amenable to establishing the sensitivity (d') and, importantly, the response bias (β, c) of the autonomous agent. Response bias determines whether a system is more likely to make a false alarm or miss. Recent research suggests that trust is more likely to affect behavioral responses of humans if a system is false alarm prone, rather than miss prone (Chancey et al., 2017; Chancey, Bliss, Liechty, & Proaps, 2015).

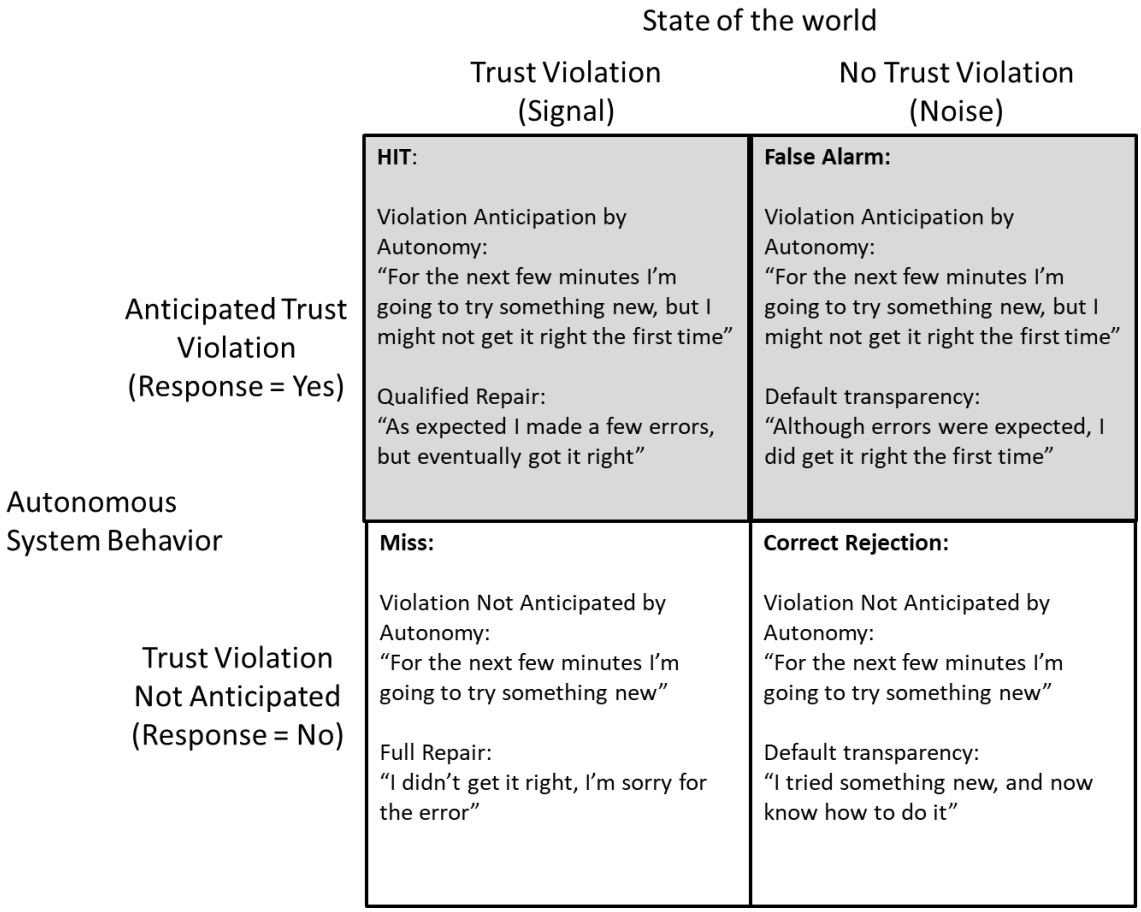


Figure 4. Illustration of de Visser et al.'s (2019) signal detection model for social trust calibration. Note: examples from Figure 7 of de Visser et al., 2019.

4.2.2. Trust Resolution and Specificity

In addition to calibration, trust resolution and specificity are useful descriptors of trust appropriateness (Lee & See, 2004). *Resolution* indicates the sensitivity of trust to differentiate among a range of automation capability levels (Figure 2). To illustrate, an operator who trusts a 60% reliable system the same as a 90% reliable system illustrates poor resolution. Presumably, if the operator trusts the 60% and 90% reliable system equally, trust should not cause the operator's agreement rate to be markedly different between these two systems. An operator who trusts a 90% reliable system slightly more than an 89% reliable system, however, illustrates good resolution and should demonstrate behavior that approximates the reliability levels accordingly (i.e.,

probability matching, though see ***Potential Issues*** section above). *Specificity* denotes the level of trust associated with a particular function (*functional specificity*) at a particular time and situation (*temporal specificity*). Functional specificity is similar to the concept of system-wide trust (Keller & Rice 2009; Rice & Geels, 2010). Keller and Rice (2009) showed that participants' responses to an individual perfectly reliable aid depended upon the presence of unrelated unreliable aids. The authors concluded that participants based their responses on "system-wide trust" rather than trust in a specific component.

4.3. The Developmental Process of Trust

Lee and See (2004) described trust as an affective evaluation of the characteristics of the trustee. Moreover, that evaluation helps determine if the trustee can achieve the goals of the trustor. This premise implies two components that form the basis of trust: the focus (i.e., what is to be trusted) and the type of goal-oriented information supporting the trust. The *focus* of trust is described according to the degree of detail (e.g., trust in an organization versus an individual). This concept is often related to general versus specific trust, which corresponds to trust specificity outlined above. From this perspective, trust might correspond to an attitude toward beliefs about the overall system of automations or beliefs about a particular mode of an automated aid (Lee & See, 2004, p. 58).

Researchers often describe *goal-oriented information* that supports trust in terms of attributional abstraction. From this perspective, trust is initially based on observable behaviors of the trustee and progresses to being based on more abstract concepts in reference to the trustee. Based on relationships among close partnerships (i.e., couples), Rempel et al. (1985) theorized that interpersonal trust is initially based on direct "coding" of partner behaviors and then, once trust becomes more established, trust is based more on the trustor's belief about the trustee's motivations (p. 98). Rempel et al. (1985) denote this evolution of trust as progressing from *predictability*, which is influenced by the predictability of a partner's behaviors, to *dependability*, which is influenced by the perception of the characteristics of the trustee, to *faith*, which is not "securely rooted" in past behaviors, but is instead based on a belief that the trustee can be depended upon irrespective of the available evidence. Another well cited article among organizational psychology is that of Mayer et al. (1995), which proposed similar bases of trust, describing ability, integrity, and benevolence (each corresponding to predictability, dependability, and faith, respectively). Based on Rempel et al. (1985), and originally proposed by Lee and Moray (1992), Lee and See (2004) proposed similar bases for trust in automation: *Performance, Process, and Purpose*.

> Performance describes what the automation does and corresponds to the current and historical operation of the automation to include reliability, predictability, and ability. This closely resembles Rempel et al.'s (1985) concept of predictability, where trust is based on observable behavior or performance. For this component, automation that readily achieves the operator's goals will lead to greater trust. To illustrate, a remote operator's trust in highly automated autopilot systems will increase in proportion to the successful observed, experienced, and reported flights that are safely completed with little or no operator intervention.
>
> Process describes how the automation operates and corresponds to the appropriateness of the automation's algorithms in achieving the operator's goals. This closely resembles

Rempel et al.'s (1985) concept of dependability, where the focus shifts from observable behaviors of the automation to the characteristics attributed to the automation. For this component, automation that appears capable of achieving the operator's goals and is understandable will lead to greater trust. To illustrate, a remote operator that has a conceptual understanding for how the autopilot systems work will be less likely to distrust it because of a rare mid-flight course correction that was executed to deconflict intersecting flight paths or re-route to an alternate landing site (i.e., the algorithm is not simply plotting a direct route between two points but considering other factors).

Purpose describes why the automation was developed and corresponds to how well the designer's intent has been communicated to the operator. This closely resembles Rempel et al.'s (1985) concept of faith, where trust is based on the belief that the automation can be depended upon in the absence of observing past behaviors. For this component, automation that achieves the goals it was designed to achieve (i.e., the operator's goals) will lead to greater trust. To illustrate, a remote operator is more likely to trust alternative auto-generated route options if they understand the reason for each (e.g., when traveling from one landing site to another, each route may represent a tradeoff between the quickest, most energy efficient, or greatest opportunity to accomplish multiple mission objectives).

In contrast to interpersonal theories (e.g., Mayer, et al., 1995; Rempel et al., 1985), where trust is hypothesized to evolve sequentially through stages of attributional abstraction (i.e., performance then process then purpose), Lee and See (2004) conceptualized trust as being based on different levels of attribution that do not necessarily follow a pre-defined sequence. Early in the human-automation relationship the operator may not have had the opportunity to observe the automation's behaviors (i.e., performance), yet may have a clear understanding of the purpose of the automation. Therefore, trust may initially be faith-based or based on purpose, rather than on the coding of observed behavioral performance.

Lee and See (2004) proposed that although trust is largely influenced by affective processes, analytical and analogical processes also determine the assimilation of goal-oriented information. From an analytical perspective, trust reflects accumulated knowledge from previous interactions with the trustee. These interactions are used to rationally and probabilistically determine the behavior of the trustee (cf. APT model by Cohen et al., 1998). To illustrate, when given the opportunity to reroute a flightpath for optimization, a remote operator may create a rational argument to analyze the expected outcome or probability of reaching their destination quickly when using the route provided by a flight path optimization tool verses a route recommended by an affiliated remote operator (e.g., optimization tool improved arrival time 24/33 times during previous flights, weighted against the affiliated remote operator being correct 7/12 times during previous flights). Lee and See (2004) argued, however, that this perspective overemphasizes the cognitive capability of the human decision maker to effectively engage in conscious calculations or to make exhaustive comparisons among alternatives (p. 62). Analytical processes, therefore, are likely complemented by other processes such as analogical judgments that rely on category membership. From this perspective, trust develops through direct observations, intermediaries who convey their own observations, and assumptions based on existing standards, category memberships, and procedures (Lee & See, 2004, p. 62). For example, the remote operator may

have learned in training that the aircraft GPS position data is less reliable in one specific region, so he or she decides not to comply with the directive because the vehicle is currently operating in that location (cf. hearsay technique used by Bliss, Dunn, et al., 1995). This process is similar to the concept of rule-based behaviors (Rasmussen, 1983), where behavior is determined by condition-action pairings. Yet, Lee and See (2004) proposed that affective processes largely influence the effect of trust on behavior, because trust is not only thought about but also felt (Fine & Holyfield, 1996, p. 25). When expectations about the trustee's performance do not conform to predictions, trust is betrayed, and emotions signal the need to change the behavior of the operator. With increasingly sophisticated automation, operators often lack the cognitive resources to rationally predict its behavior. Compounding this effect, the nondeterministic algorithms leveraged by increasingly autonomous systems may not even allow the designer insight into system behaviors, much less the human interacting or collaborating with the system. Lee and See (2004) suggest, therefore, that emotions guide behaviors when rules do not apply or when cognitive resources are not available to make a rational choice.

Importantly, the robustness and stability of trust depends on how the human mentally represents the goal-oriented informational bases referencing the automation and determines the appropriateness of intentions to use the automation and behavioral interactions with the automation (Lee & See, 2004). One approach to conceptualizing the belief structures of Performance, Process, and Purpose, are to think of them as the user's mental models of the automation.

5. A Mental Model Approach to Support Appropriate Trust

Researchers and practitioners interested in the ways in which humans interact with computers and complex systems (e.g., automation) have frequently used the term "mental model" to describe the mental representations of system processes humans use to reason, infer, and make predictions about the technologies they interact with (Allen, 1997; Moray, 1998, 1999; Norman, 1983). Although not directly observable, different categories of evidence have been used to infer characteristics of mental models: e.g., operators can predict system processes and how those processes may affect or interact with other subsystems; operators can explain the cause of system events or reasons for errors; operators that are trained with models perform better than those without training (Allen, 1997). These models are approximations of the target system, which maintain essential aspects of the original, but are neither entirely accurate or provide an exhaustive account of system processes (i.e., these models are not isomorphic representations of the real-world system, but are instead homomorphic; Allen, 1997; Moray, 1999).

Mental models dynamically evolve through interactions with a system and are affected by prior operator experience and knowledge about that system. Importantly, these models need not be technically accurate (and often are not), only functional, and are continuously modified as they relate to achieving the goals of the operator (Norman, 1983, p. 3; cf. dynamic nature of trust). If an operator possesses an inappropriate mental model for what the automation is doing (Performance), how it is doing it (Process), or why the automation was developed (Purpose), then these models will dictate to the operator that the system is unable to regularly achieve their goals. In this case, the goal-oriented informational bases that support trust are impoverished, and for the operator to obtain a "workable result" it makes sense to reject the automation. Supporting this point, Beggiato and Krems (2013) showed that participants with a more complete mental model of an adaptive cruise control system tended to show an increase in trust over time, compared to a

group with incomplete mental models where trust decreased (see also Beggiato, Pereira, Petzoldt, & Krems, 2015). Similarly, Fogg and Tseng (1999) proposed that an interface is less likely to be perceived as credible when it does not match the user's mental model (a component of which is trustworthiness).

Although researchers and practitioners tend to focus on the mental model of the system user, or *User Model* (e.g., pilot, passenger, remote operator), the mental model of the designer, or *Design Model*, needs to be considered as well (Norman, 1986). The Design Model allows the designer to conceptualize the automation and is based on the task and technical limitations and capabilities of the automation. Ideally, the Design Model should also be based on the capabilities, limitations, attitudes, intentions, and behaviors of the (active or passive) human involved in the completion of the (partially or fully) automated task. Importantly, the User Model is not formed directly from the Design Model, but is instead mediated by the *System Image*, or the physical system itself, which includes the training and instructional documentation (Norman, 1986). For trust to be appropriately calibrated to system capabilities, the Design Model and User Model should be aligned. To accomplish this alignment, the System Image needs to support the belief structures of trust (i.e., Performance, Process, and Purpose). One method to accomplish this is to design and train for automation transparency.

5.1. The System Image and Transparency

Automation transparency is "…the communication of system-centered factors and human-centered factors that promote shared awareness and shared intent within a human-machine team" (Lyons, Clark, Wagner, & Schuelke, 2017, p. 41). Lyons (2013) outlines a design framework of transparency, which partially focuses on the information that the automation needs to share with the human about its own perspective on the task it is completing, how it is completing it, and awareness of its intentions and limitations given a particular environmental context (p. 51). This information is termed Automation-to-Human[6] Transparency and is captured in three System Image models: Task Model, Analytical Model, and Intentional Model (Lyons, 2013). These models closely resemble the informational bases, or beliefs, that support trust (i.e., Performance, Process, and Purpose).

The *Task Model* describes the information that allows the human to analyze the actions of the automation (Lyons, 2013). In the proposed framework, the closer the Task Model corresponds to the Performance-based User Model (i.e., mental model), the more likely trust will be calibrated to match system capabilities. As suggested by Lyons (2013), the automation should indicate an understanding of the task structure and its current goals in that task structure. For example, a highly automated UAM flight might be divided into separate phases: passenger loading, takeoff, cruise/navigation, landing, and passenger unloading. Phases need to be communicated to both the passenger and pilot to support a shared understanding (i.e., transparency) of what the automation is doing (e.g., display the current phase) and what it intends to do (e.g., estimated time until next phase; cf. Lyons, 2013, example of the Task Model). In this example, trust can be calibrated to how well the display matches the behaviors of the UAM aircraft: trust will increase or be maintained if the aircraft is descending while the display indicates it is in the landing phase, whereas trust will decrease if the aircraft is descending while the display indicates it is in the

[6] Note: Lyons uses the term "robot-to-human," yet automation was chosen for the purposes of this work and does not differ in meaning or intent

passenger-unloading phase. If no phase indication is provided (i.e., the system is opaque), then trust will fluctuate erratically as the passenger and/or pilot are unable to predict phase transitions. Moreover, Lyons (2013) specifies that there should be a shared awareness of the capabilities of the automation in specific contexts. For example, an auto-land feature may not be capable of safely executing landings on vertiports (i.e., elevated landing pads for UAM vehicles) of certain diameters and altitudes when crosswinds exceed a specified threshold. Here, there should be a design feature that removes these vertiports as destination options or indicates that a manual landing is required (if that is indeed a safer option). Training may be necessary to reinforce these design factors, otherwise pilots may think there is a system error when a desired vertiport is not available, or when the automation unexpectedly requests a handoff to manual control during the landing phase.

The *Analytical Model* describes the information that allows the user to analyze how the automation is making decisions, such as the calculations and algorithms it is relying upon, and reasons for system errors (Lyons et al., 2017). In the proposed framework, the closer the Analytical Model corresponds with the Process-based User Model (i.e., mental model), the more likely trust will be calibrated to match system capabilities. Returning to the previous auto-land feature example, displaying wind speeds and indicating breaches in preset thresholds would help the pilot (or operator) calibrate their trust to match the capabilities of the automation in those environmental conditions. Indicating this information to passengers would also help calibrate their trust to match UAM transportation capabilities, to show that there are not system failures, but instead excessive winds are impacting operations (particularly if ground level winds experienced by a waiting passenger are drastically different than at a vertiport multiple stories high at the desired destination). Designers should be cautious, however, not to display too much analytical information to UAM pilots, who may have limited information processing resources to re-allocate during even highly automated flights (see Chancey, 2021, and Chancey & Politowicz, 2020, for discussions on minimally trained UAM pilots/operators). The minimal training that pilots do receive, should focus on providing a general understanding for how the automated tools they rely upon work, and how that relates to automation limitations in certain situations.

Finally, the *Intentional Model* describes information that helps the user analyze why the automation was created and allows the user to place the actions (e.g., Task Model) of the automation in the appropriate strategic context (Lyons, 2013; Lyons et al., 2017). In the proposed framework, the closer the Intentional Model corresponds with the Purpose-based User Model (i.e., mental model), the more likely trust will be calibrated to match system capabilities. Design techniques that provide links between the physical appearance of the system or display features and the functional goal of the system should be used to communicate the automation's purpose in a given context (e.g., metaphors that relate familiar functions to new ones). To illustrate, because the display and controls of UAM aircraft will need to be drastically simplified (i.e., SVO), designers could leverage the existing User Models for common ground-based GPS navigation aids (e.g., Google Maps, Apple Maps application). A modified UAM-version of these common GPS designs would likely communicate their purpose and functionality to most users (pilots/operators and passengers), given the ubiquity of these devices and the context of operations. Yet pilots or passengers that have not interacted with these aids, and therefore possess an incomplete mental model, may disuse this type of automation due to a lack of trust. Alternatively, users that have extensively developed mental models with specific ground-based GPS devices/applications may

apply inappropriate User Models to a UAM-GPS and fail to detect system malfunctions or inappropriately attribute system errors to normal functionality (e.g., ground-based GPS devices do not display altitude or relative positions of other vehicles). Communicating the Intentional Model of automated functions should be an important System Image consideration to establish appropriately calibrated trust from the outset, to ensure existing User Models are not interacting with the Design Model to produce unintended effects.

6. Facilitating Appropriate Trust in HAI and HAT

From the perspective outlined in the current work, a transparent System Image is a design and training technique that translates the Design Model into the User's Model, with the particular purpose of calibrating human-automation trust to match automation capabilities (i.e., trustworthiness). This section outlines initial concepts for achieving calibrated trust through various "design paths." The intention here is to provide only an introduction to these concepts and is not intended to be a final test-plan. The following sub-sections provide a framework to begin operationalizing key aspects of the proposed framework, and to outline a path for future research and development efforts to facilitate HAI and HAT in AAM operations.

6.1. Fixed Design

Although we do not expect the AAM community to pursue this path (nor do we offer a specific validation plan), the Fixed Design model is intended to serve as a caution against pursuing technologically difficult operations too aggressively at early stages in development (e.g., proceeding directly to highly automated operations). This path offers little, if any, opportunity to actively realign the Design Model with the User Model via the System Image (Figure 5). Ideally, this path should be the culmination of a methodically well thought out and researched System Image that has undergone numerous re-designs, as the human-automation interaction feedback loop will increasingly engrain and fortify the belief structures with each encounter between the human and the target automation. If the System Image effectively aligns the Design and User models, then trust will be both robust and match the capabilities of the automation (i.e., well calibrated). If the Design and User Models are misaligned, however, then a re-design is warranted (see next section). Unfortunately, from a user perspective, many AAM concepts will not initially be robust to even mild misalignments.

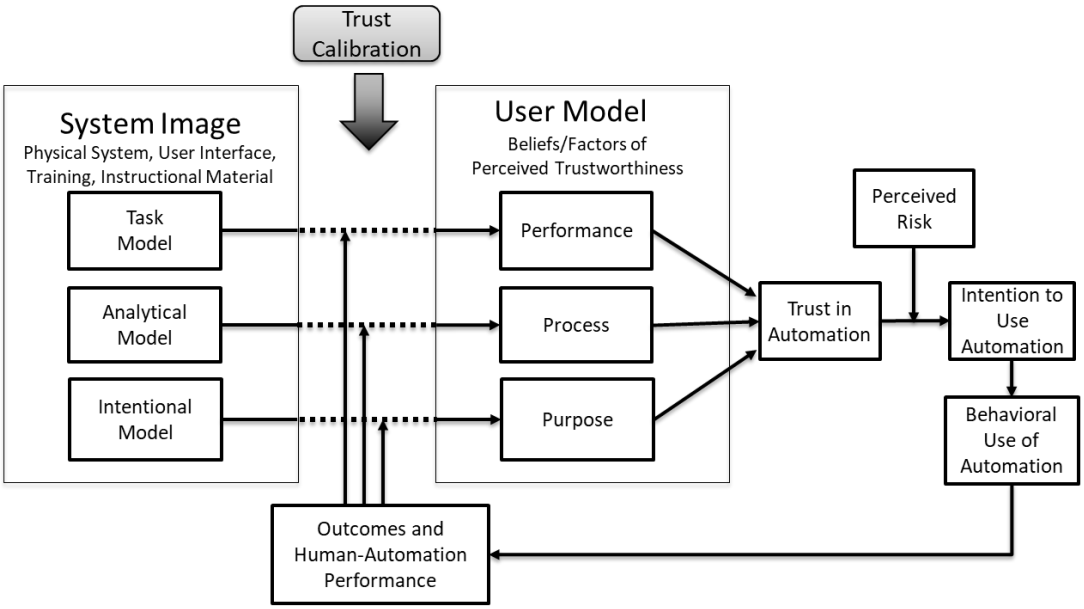


Figure 5. Model for Fixed Design in HAI and HAT paradigms.

To illustrate, trust tends to be quite stable or robust when based on several belief structures (i.e., Performance, Process, Purpose; Lee & See, 2004). Yet during early UAM operations, for example, many potential passengers will not have had the opportunity to personally experience UAM transportation or witness many UAM flights above their community. These types of interactions are critical for developing the Performance aspect of trust. Moreover, even current commercial airline passengers prefer human-piloted aircraft to highly automated aircraft, indicating a lack of understanding for how these operations are actually executed (Rice et al., 2014). With emerging UAM operations, it is difficult to envision passengers having a basic understanding for how these operations are executed either (i.e., a lack of the Process basis of trust). Instead, passenger trust will likely be rooted in the Purpose base of trust, as marketing and news coverage will provide a basis to understand the reason for UAM. Compounding this issue, interpersonal trust research also shows that high perceived risk is a significant predictor of fragile trust in new relationships (i.e., trust level is likely to undergo large changes in a short timeframe; McKnight, Cummings, & Chervany, 1998). Clearly, the perceived risk of UAM transportation accidents will likely be an early concern for most passengers.

Given these conditions, a real or perceived incident could lead to an abrupt and drastic drop in public trust (and subsequent demand). To mitigate the effects of initially fragile trust, the Task and Analytical Models should be given particular attention to ensure the System Image supports the Performance and Process bases of trust early on. Clearly, the Fixed Design Path (Figure 5) presents a potentially fraught strategy if it is the only available option. If pursued early, this path places a great deal of the burden on the designer to "get it right" the first time.

6.2. Iterative Design and Experimentation

Many of the AAM concepts discussed in this work have been geared toward the types of automation that will emerge in the coming years and decades. Fortunately, at the time of this writing, these approximate periods offer a buffer to allow the research and development community to test and evaluate System Image concepts that align the Design and User Models to

support appropriate trust in AAM operations. The Iterative Design Path is similar to the concept of bridging the gulf of evaluation (i.e., the mismatch between a system's representation and what an operator expects, which is joined by "moving" the system closer to the user; Norman, 1986). The model in Figure 6 lends itself to iterative experimentation that attempts to establish the effects of the match (or mismatch) between the Design Model and User Model on trust and behavioral responses toward the automation, which would begin to validate the theoretical framework proposed in the current work. Yet, as stated earlier, mental models are not directly observable, and instead are generally inferred from user performance metrics or verbal think aloud protocol (see Rowe & Cook, 1995, for comparisons of techniques).

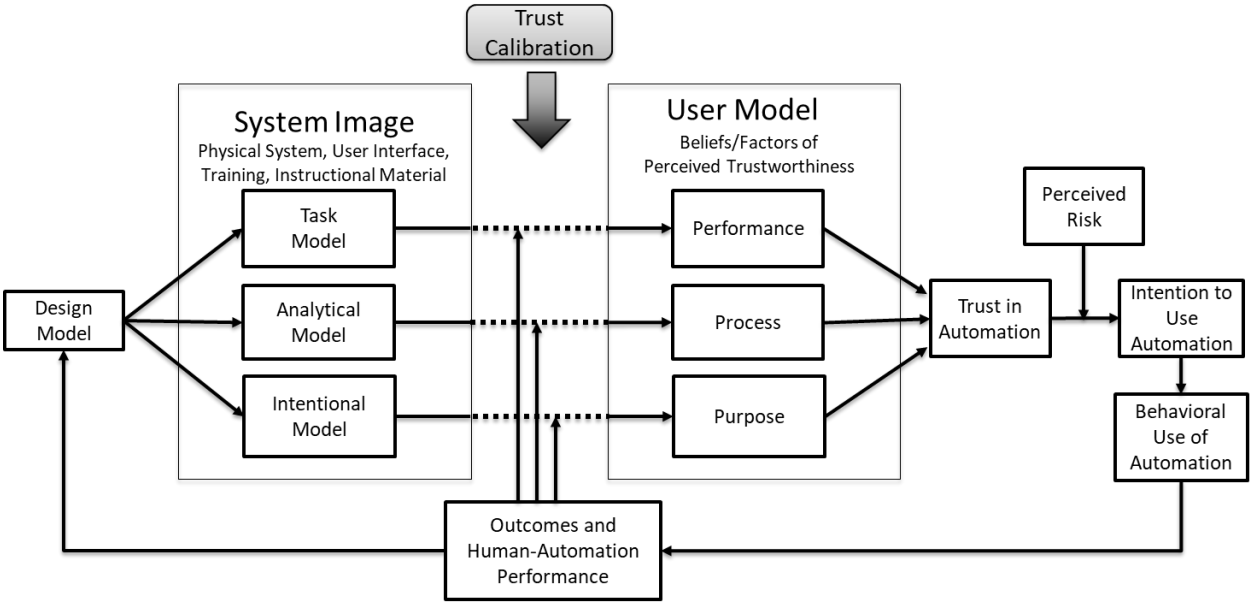


Figure 6. Model for Iterative Design in HAI and HAT paradigms. *Note*: Added feedback loop through Design Model.

One option for operationalizing mental models is with Pathfinder Network Analysis. Based on graph theory, Pathfinder is a statistical technique that represents knowledge structures in graphical form (Schvaneveldt, Durso, & Dearholt, 1989), and has been used extensively in human-computer interaction studies to represent and quantify mental models (see Cooke, Neville, & Rowe, 1996). Moreover, quantitative comparisons between individual networks can be made using the C statistic (ranging from 0 [not related] to 1 [strongly related]), which is a measure of shared links for matching nodes. Specifically, the Pathfinder method provides the ability to quantify the degree to which a representative Design Model matches a User Model. A parallel multiple-mediation analysis could be used to analyze specific pathways (see Figure 7). Chancey and Politowicz (2020) used a similar statistical technique to establish the relationship between UAM concept of operation factors on public acceptance through individual factors of trust (i.e., Performance, Process, Purpose; see also Chancey et al., 2017). This approach could be used as a framework to pursue iterative experimental studies that begin validating the concepts outlined in this document.

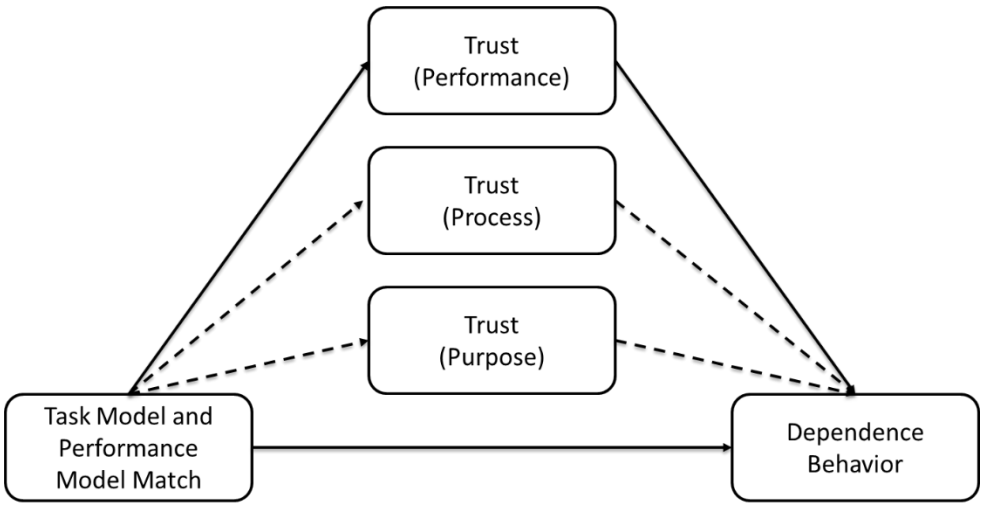


Figure 7. Parallel multiple mediation model for the effects of Design and User mental model match on dependence behaviors through factors of trust. Note: The example model indicates that the Performance basis of trust should provide the strongest mediating effect on dependence behaviors because it is based on the degree of similarity between the Task Model and Performance Model.

6.3. Adaptive Trust Calibration

The concept of HAT implies that a human and automated system interact, dynamically, interdependently, and adaptively toward a common goal (cf. Salas et al., 1992, p. 7). Clearly, increasingly autonomous systems may be better "equipped" to enter into this type of collaboration with a human partner than systems at lower levels of automation (e.g., FMS). To this point, increasingly autonomous systems were described earlier as possessing the ability to be "generative and learn, evolve and permanently change their functional capacities as a result of the input of operational and contextual information" (Hancock, 2017, p. 284). Trust will play an important role in mediating the relationship between the human and increasingly autonomous system, and true teaming may be difficult to achieve if the system is unable to dynamically adapt to facilitate collaboration. To accomplish this, we introduce the concept of *adaptive trust calibration*, and propose an initial descriptive model that may be used to operationalize this concept (Figure 8).

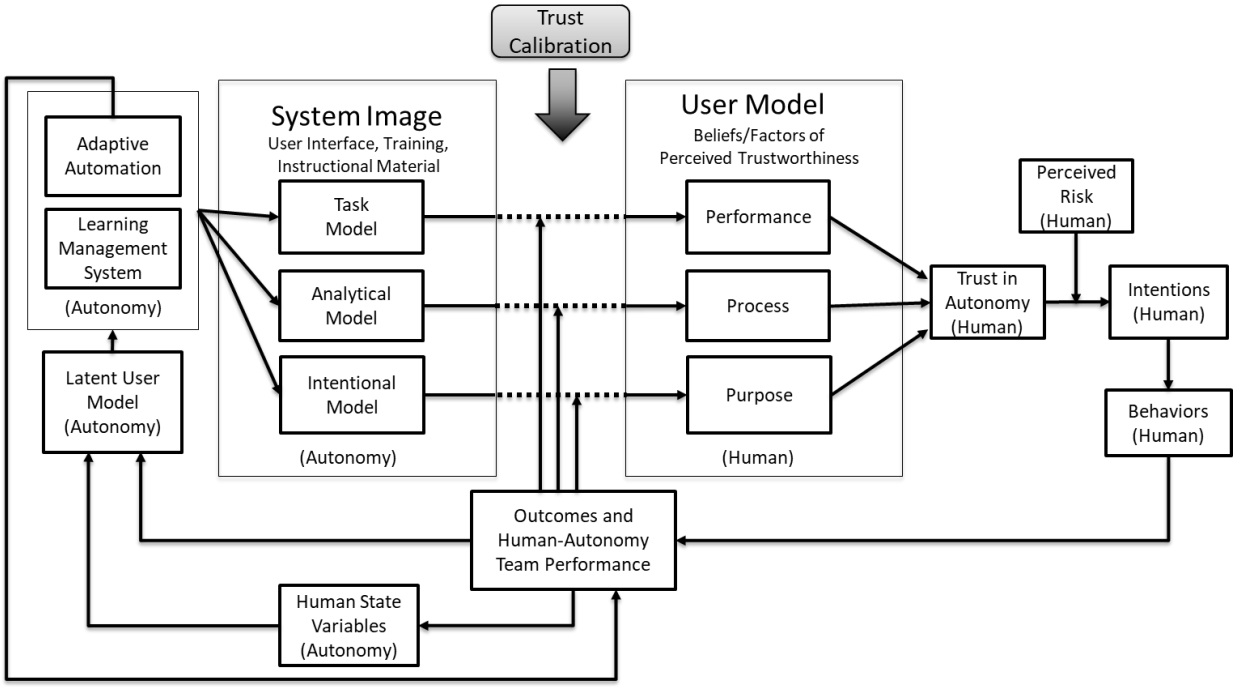


Figure 8. Model for Adaptive Trust Calibration in HAT Paradigm. Note: Added feedback loop through Human State Variables, Latent User Model, Adaptive Automation, and Learning Management System.

Although the model in Figure 8 possesses the same underlying theoretical principles outlined throughout this document, and in Figure 6, there are important differences that allow for the "adaptive" aspect of the model to function as a closed-loop process (i.e., closing the loop does not require iterative experiments to align the User Model and Design Model). Specifically, the Adaptive Trust Calibration model assumes that the increasingly autonomous system is consistently sampling, storing, and analyzing information about the human operator and the overall performance of the human-automation team. *Human State Variables* provide a baseline to establish the current state of the human (e.g., fatigue, workload, attentional tunneling), given the cognitive and physiological metrics available to the system (see Stephens et al., 2018, for overview of biocybernetic adaptation strategies in a closed-loop system). Both the Human State Variables and overall HAT performance metrics inform the *Latent User Model*, which is formed by the system as a dynamic analog to the Design Model discussed in the previous section. Here, the system constructs a hypothesized User Model in an attempt to anticipate potential trust miscalibration between the System Image and the actual User Model. Lattice Theory may offer a method to formalize the Latent User Model. Moray (1998; 1999) proposed the use of lattice notation as a mathematical modeling technique to represent homomorphic models that share similar qualities to the theoretical descriptions of mental models. Graphically, lattices form interconnected nodes (e.g., knowledge about a system) that are partially ordered sets to show how elements relate to each other (Moray, 1999; compare to Pathfinder method discussed in section 6.2). Moreover, adapting causal classifications originally introduced by Aristotle, Moray (1999) proposes that the ordering of nodes can be considered as causal links. Those classifications align well with the informational bases of trust outlined in the current work:

- *Efficient Causes* (related to *Performance*) refer to actions that bring about change. For example, clicking on a displayed drone causes the interface to give me additional control options for that drone. Selecting a destination causes that selected drone to go to that location.
- *Material Causes* (related to *Process*) refer to the underlying processes. For example, if fog reduces visibility to less than 1 mile at the peninsula, then that will cause the vertiports to be out of service in that location.
- *Final Causes* (related to *Purpose*) refer to the end purpose for which the event happens. For example, the package arrived at my doorstep by drone because I wanted it within the hour.

To construct the lattice, however, the system requires a method to sample pertinent information from the user and organize it into a coherent model. The Conant Method of Extended Dependency Analysis may provide a means to construct user mental models of increasingly autonomous systems via operator control strategies (see Conant, 1976, Conant, 1996, Jamieson, 1996, and Moray 1999 for descriptions of this method). Beyond interactions and control strategies, eye tracking techniques may offer additional information to create robust intentional strategy models to complement this method.

Once the autonomous system has created the Latent User Model, if the system hypothesizes a misaligned mental model that would lead to inappropriate (miscalibrated) trust, then it has two methods to alter the system image. First, *adaptive automation* strategies could be employed to dynamically reconfigure or add/remove informational elements in the displays (see Kaber, Riley, Tan, & Endsley, 2001, Rouse, 1988, and Scerbo, 1996, for reviews). Additionally, the system could also attempt to reorient or alert the user to important environmental or display elements. Second, a *Learning Management System* (LMS; e.g., Blackboard®) could prepare and tailor training material that explicitly attempts to realign the User Model, or Artificial Intelligence scheduling algorithms that choose training courses and even schedule learning events for students, as used in the United States Air Force (Carlin, Ward, & Freeman, 2016). Both approaches could be used to update the System Image and support Adaptive Trust Calibration.

7. Conclusion

To enable effective HAI and HAT in the context of AAM, the current paper has outlined a theoretical framework to design and train for appropriate trust in automation. The main contributions of this work reside in connecting the construct of trust to mental models. Using the outlined mental model approach, novel HAT strategies such as Adaptive Trust Calibration could be researched for use in increasingly autonomous systems within AAM operations and beyond. This work, however, represents only an initial proposal for future studies and more research is required to validate the ideas presented in this paper.

8. References

Adams, B. D., Bruyn, L. E., Houde, S., & Angelopoulos, P. (2003). *Trust in automated systems literature review.* Defense Research and Development Canada, Toronto.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211.

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84* (5), 888-918.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Upper Saddle River, NJ: Prentice Hall.

Allen, R. B. (1997). Mental models and user models. In M. Helander, T.K. Landauer, & P. Prabhu (Eds.), *Handbook of human-computer interaction* (2nd ed.). Amsterdam: Elsevier.

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

Bailey, N. R., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science, 8* (4), 321-348.

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

Beggiato, M., & Krems, J. F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation Research part F, 18*, 47-57.

Beggiato, M., Pereira, M., Petzoldt, T., & Krems, J. (2015). Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transportation Research Part F, 35*, 75-84.

Bliss, J. P. (2003a). Investigation of alarm-related accidents and incidents in aviation. *The International Journal of Aviation Psychology, 13*(3), 249-268.

Bliss, J. P. (2003b). An investigation of extreme alarm response patterns in laboratory experiments. *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting*, (pp. 1683-1687). Denver, CO: The Human Factors and Ergonomics Society.

Bliss, J. P., Dunn, M., & Fuller, B. S. (1995). Reversal of the cry-wolf effect: An investigation of two methods to increase alarm response rates. *Perceptual and Motor Skills*, *38* (11), 1231-1242.

Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics, 38* (11), 2300-3212.

Brunswik, E. (1952). The conceptual framework of psychology. *International Encyclopedia of Unified Science.* (Vol. 1, No. 10). Chicago, IL: The University of Chicago Press.

Carlin, A., Ward, D., & Freeman, J. (2016). Representation, selection, and scheduling of training in a lifelong learning context. *MODSIM World 2016 (No. 49), pp 1-10.*

Chancey, E.T. & Bliss, J.P. (2012). Unreliable information in infantry situation awareness: Improvement through game-based training. *Simulation & Gaming, 43* (5), 581-599.

Chancey, E. T., Bliss, J. P, Liechty, M., & Proaps, A. B. (2015). False alarms vs. misses: Subjective trust as a mediator between reliability and reaction measures. *Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting*. (pp. 647-651). Los Angeles, CA.

Chancey, E.T., Bliss, J.P., Yamani, Y., & Handley, H.A.H. (2017). Trust and the compliance-reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors, 59* (3), 333-345.

Chancey, E. T. & Politowicz, M. S. (2020). Public trust and acceptance for concepts of remotely operated Urban Air Mobility transportation. *Proceedings of the Human Factors and Ergonomics Society.*

Chen, J. Y. C., K. Procci, M. Boyce, J. Wright, A. Garcia, and M. Barnes. 2014. Situation Awareness-Based Agent Transparency. Report No. ARL-TR-6905, Aberdeen Proving Ground, MD, U.S. Army Research Laboratory.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: A model and its training implications. *Proceedings of the Command and Control Research and Technology Symposium,* (pp. 1-37).

Conant, R. C. (1976). Laws of information which govern systems. *IEEE Transactions on Systems, Man, and Cybernetics, 6* (4), 240-255.

Conant, R. C. (1996). An information-theoretic method for revealing system structure. *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting*. (pp. 196-172).

Cooke, N. J., Neville, K. J., & Rowe, A. L. (1996). Procedural network representations of sequential data. *Human-Computer Interaction, 11*, 29-68.

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

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R. & Neerincx, M. A. (2019). Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics*.

de Winter, J. C. F. (2019). Pitfalls of automation: A faulty narrative. *Ergonomics, 62* (4), 505-508.

de Winter, J. C. F. & Dodou, D. (2014). Why the Fitts list has persisted throughout the history of function allocation. *Cognition, Technology & Work, 16,* 1-11.

Dekker, S. W. A. & Woods, D. D. (2002). MABA-MABA or abracadabra? Progress on human-automation co-ordination. *Cognition, Technology & Work, 4*, 240-244.

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human Relations, 13*, 123-139.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology, 13* (3), 147-164.

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.

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

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors, 59* (1), 5-27.

Feary, M. (2018). A first look at the evolution of flight crew requirements for emerging market aircraft. *AIAA Aviation.* Atlanta, GA: AIAA.

Fine, G. A., & Holyfield, L. (1996). Secrecy, trust, and dangerous leisure: Generating group cohesion in voluntary organizations. *Social Psychology Quarterly, 59* (1), 22-38.

Fishbein, M., & Ajzen, I. (1975). Belief, attitude, intention, and behavior: An introduction to theory and research. Reading, MA: Addison-Wesley.

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

Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. In *Proceedings of CHI* (pp. 80–87). Pittsburgh, PA: ACM.

General Aviation Manufacturers Association (2019). "A Rational Construct for Simplified Vehicle Operations (SVO)." GAMA EPIC SVO Subcommittee Whitepaper, Version 1.0, Washington, DC.

Gempler, K. S. (1999). Display of predictor reliability on a cockpit display of traffic information (Master's thesis). Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a366236.pdf.

Hancock, P. A. (2017). Imposing limits on autonomous systems. *Ergonomics, 60* (2), 284-291.

Hancock, P. A. (2019). Some pitfalls in the promises of automated and autonomous vehicles. *Ergonomics, 62* (4), 479-495.

Hergeth, S. Lorenz, L. Vilmek, R., & Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors, 58* (3), 509-519.

Ho, N. T., Sadler, G. G., Hoffman, L. C., Lyons, J. B., & Johnson, W. W. (2017). Trust of a military automated system in an operational context. *Military Psychology, 29* (6), 524-541.

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

Hoffman, R. R. & Hancock, P. A. (2017). Measuring resilience. *Human Factors, 59* (4), 564-581.

Holbrook, J. B., Stewart, M., Smith, B., Prinzel III, L. J., Matthews, B., Avrekh, I., Cardoza, C., Ammann, O. Adduru, V., Null, C. (2019). Human performance contributions to safety in commercial aviation. NASA-TM-2019-220417. Washington, DC: NASA, 2019.

Holbrook, J. B., Prinzel III, L. J., Chancey, E. T., Shively, M. S., Feary, M. S., Dao, Q. V., Ballin, M. G., & Teubert, C. (2020). Enabling urban air mobility: Human-autonomy teaming research challenges and recommendations. *AIAA AVIATION*, Reno, NV: AIAA.

Holden, J., & Goel, N. (2016). Uber Elevate: Fast-forwarding to a future of on-demand urban air transportation. *Retrieved from*: https://www.uber.com/elevate.pdf/

Hollnagel, E. (2015). RAG-The resilience analysis grid. *Resilience engineering in practice. A guidebook. Farnham, UK: Ashgate*.

Jamieson, G. A. (1996). Using the Conant method to discover and model human-machine structures. *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting*, (pp. 173-177).

Jordan, N. (1963). Allocation of functions between man and machines in automated systems. *Journal of Applied Psychology, 47*, 161-165.

Kaber, D., Riley, J. M., Tan, K., & Endsley, M. R. (2001). On the design of adaptive automation for complex systems. *International Journal of Cognitive Ergonomics, 5,* 37-57.

Karpinsky, N. D., Chancey, E.T., & Yamani, Y. (2016a). Modeling relationships among workload, trust, and visual scanning in an automated flight task. *Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting*. (pp. 1550-1554). Washington, D.C.

Karpinsky, N. D., Chancey, E. T., & Yamani, Y. (2016b). Trust and attention in flight simulation with imperfect signaling system. *Proceedings of the 2016 Industrial and Systems Engineering Research Conference.* Anaheim, CA.

Karpinsky, N. D., Chancey, E. T., Palmer, D. B., & Yamani, Y. (2018). Automation trust and attention allocation in multitasking workspace. *Applied Ergonomics, 70,* 194-201.

Keller, D. P., & Rice, S. (2009). System-wide versus component-specific trust using multiple aids. *The Journal of General Psychology, 137* (1), 114-128.

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review Psychology, 50,* 569-598.

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

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40,* 153-184.

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

Lee, J. D. & Seppelt, B. D. (2012). Human factors and ergonomics in automation design. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (4th ed., pp. 1651-1642). Hoboken, NJ: John Wiley & Sons.

Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. *Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium*, 48-53.

Lyons, J. B., Clark, M. A., Wagner, A. R., & Schuelke, M. J. (2017). Certifiable trust in autonomous systems: Making the intractable tangible. *AI Magazine, 38* (3), 37-49.

Lyons, J. B. & Stokes, C. K. (2012). Human-human reliance in the context of automation. *Human Factors, 54* (1), 112-121.

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

Manzey, D., Gerard, N., & Wiczorek, R. (2014). Decision-making and response strategies in interactions with alarms: The impact of alarm reliability, availability of alarm validity information and workload. *Ergonomics, 57* (12), 1833-1855.

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

McBride, M., & Morgan, S. (2010). Trust calibration for automated decision aids. *Institute for Homeland Security Solutions.*[Online]. Available: https://www.ihssnc. org/portals/0/Documents/VIMSDocuments/McBride_Research_Brief. pdf.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review, 23* (3), 473-490.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors, 43* (4), 563-572.

Moore, M., & Goodrich, K. (2015). *On-Demand Mobility: Goals, Technical Challenges, and Roadmaps*. Retrieved from https://ntrs.nasa.gov/citations/20160006950.

Moray, N. (1998). Identifying mental models of complex human-machine systems. *International Journal of Industrial Ergonomics, 22,* 293-297.

Moray, N. (1999). Mental models in theory and practice. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 223-258). Cambridge, MA: MIT Press.

Moray, N. (2003). Monitoring, complacency, skepticism and eutactic behavior. *International Journal of Industrial Ergonomics, 31*, 175-178.

Moray, N., & Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control, 21* (4/5), 203-211.

Moray, N. & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomic Science, 1* (4), 354-356.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*, 527-539.

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.

National Academies of Sciences, Engineering, and Medicine (2020). *Advanced Aerial Mobility: A National Blueprint.* Washington, DC: The National Academies Press.

Nickerson, R. S. (1999). Automation and human purpose: How do we decide what should be automated? In M. W. Scerbo, & M. Mouloua (Eds.), *Automation Technology and Human Performance* (pp. 11-19). Mahwah, NJ: Lawrence Erlbaum Associates.

Norman, D. (1983). Some observations on mental models. In D. Gentner & A. Stevens (Eds.), *Mental models* (pp. 7-14). Hillsdale NJ: Erlbaum.

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman and S. W. Draper (Eds.*), User centered system design* (pp. 31-61). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Okamura, K., & Yamanda, S. (2020). Adaptive trust calibration for human-AI collaboration. *PLoS One, 15*(2).

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors, 56* (3), 476-488.

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

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

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all of these years of automation. *Human Factors, 50* (3), 511-520.

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

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making, 2* (2), 140-160.

Politowicz, M. S., Chancey, E. T., & Glaab, L. J. (2021). Effects of autonomous sUAS separation methods on subjective workload, situation awareness, and trust. *2021 AIAA SciTech Forum.*

Prinzel III, L. J., DeVries, H., Freeman, F. G., & Mikulka, P. (2001). *Examination of automation-induced complacency and individual difference variates* (Technical Memorandum No. TM-2001-211413). Hampton, VA: National Aeronautics and Space Administration Langley Research Center.

Pritchett, A., Portman, M. & Nolan, T. (2018). Research & Technology development for human-autonomy teaming – Final report: Literature review and findings from stakeholder interviews, NASA Langley Research Center, Hampton, VA.

Rasmussen, J. (1983). Skills, rules, and knowledge: Signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13*, 257-266.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology, 49* (1), 95-112.

Rice, S., & Geels, K. (2010). Using system-wide trust theory to make predictions about dependence on four diagnostic aids. *The Journal of General Psychology, 137* (4), 362-375.

Rice, S., Kraemer, K., Winter, S. R., Mehta, R., Dunbar, V., Rosser, T. G., & Moore, J. C. (2014). Passengers from India and the United States have differential opinions about autonomous auto-pilots for commercial flights. *International Journal of Aviation, Aeronautics, and Aerospace, 1* (1), 1-12.

Riley, V. (1996). A theory of operator reliance on automation. In M. Mouloua & R. Parasuraman (Eds.), *Human Performance in Automated Systems: Recent Research and Trends* (pp. 8-14). Hillsdale, NJ: Erlbaum.

Rouse, W.B. (1988). Adaptive aiding for human/computer control. *Human Factors, 30*, 431-443.

Rowe, A. L., & Cooke, N. J. (1995). Measuring mental models: Choosing the right tools for the job. *Human Resource Development Quarterly, 6* (3), 243-255.

Salas, E., Dickinson, T. L., Converse, S. A., & Tannenbaum, S. I. (1992). *Toward an understanding of team performance and training.* In R. W. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (p. 3–29). Ablex Publishing.

Scerbo, M. W. (1996). Theoretical perspectives on adaptive automation. In R. Parasuraman, & M. Mouloua (Eds.), *Automation and human performance: Theory and applications. Human factors in transportation* (pp. 37-63). Hillsdale, NJ, England: Lawrence Erlbaum.

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. *Academy of Management Review, 32* (2), 344-354.

Schutte, P. C., Goodrich, K. H., Cox, D. E., Jackson, E. B., Palmer, M. T., Pope, A. T., Schlecht, R. W., Tedjojuwono, K. K., Trujillo, A. C., Williams, R. A., Kinney, J. B., & Barry, J. S. (2007). *The Naturalistic Flight Deck system: An integrated system concept for improved single-pilot operations* (Technical Memorandum No. TM-2007-215090). Hampton, VA: National Aeronautics and Space Administration Langley Research Center.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 24, pp. 249-284). New York: Academic Press.

Seong, Y., & Bisantz, A. M. (2000). Modeling human trust in complex, automated systems using a lens model approach. In M. W. Scerbo & M. Mouloua (Eds.), Automation Technology and Human Performance: Current Research and Trends (pp. 95-100). Mahwah, NJ: Lawrence Erlbaum Associates.

Seong, Y., Bisantz, A. M., & Gattie, G. J. (2006). Trust, automation, and feedback: An integrated approach. In A. Kirlik (Ed.), Adaptive Perspectives on Human-Technology Interaction: Methods and Models for Cognitive Engineering and Human-computer Interaction (pp. 105-113). New York, NY: Oxford University Press, Inc.

Sheridan, T. B. (1988). Trustworthiness of command and control systems. *Man-Machine Systems.* 427-431.

Sheridan, T. B. (2000). Function allocation: algorithm, alchemy or apostasy?. *International Journal of Human-Computer Studies, 52*, 203-216.

Sheridan, T. B. (2019a). Individual differences in attributes of trust in automation: Measurement and application to system design. *Frontiers in Psychology, 10* (1117).

Sheridan, T. B. (2019b). Extending three existing models to analysis of trust in automation: signal detection, statistical parameter estimation, and model-based control. *Human Factors, 61* (7), 1162-1170.

Sheridan, T. B., Fischhoff, B., Posner, M., & Pew, R. W. (1983). Supervisory Control Systems. *Research Needs for Human Factors* (pp. 49-77). Washington, D.C.: National Academy Press.

Sheridan, T. B., & Hennessy, R. T. (Eds.). (1984). Research and Modeling of Supervisory Control Behavior: Report of a Workshop. Washington, D.C.: National Academy Press.

Sheridan, T. B., & Verplank, W. L. (1978). Human and computer control of undersea teleoperators. MIT Man-Machine Systems Laboratory Report.

Sitkin, S. B., & Pablo, A. M. (1992). Reconceptualizing the determinants of risk behavior. *Academy of Management Review, 17,* 9–38.

Sorkin, R. D. (1988). Why are people turning off our alarms? *Journal of the Acoustical Society of America, 84* (3), 1107-1108.

Stephens, C., Dehais, F., Roy, R. N., Harrivel, A., Last, M. C., Kennedy, K., & Pope, A. (2018). Biocybernetic adaptation strategies: Machine awareness of human engagement for improved operational performance. In D. Schmorrow & C. Fidopiastis (Eds.) *Augmented Cognition: Intelligent Technologies. AC 2018. Lecture Notes in Computer Science, 10915*. Springer, Cham.

Stewart, J. (2018, January 24). *People Keep Confusing Their Teslas for Self-Driving Cars*. Retrieved from WIRED: https://www.wired.com/story/tesla-autopilot-crash-dui/

Sullivan, J. M., Tsimhoni, O., & Bogard, S. (2008). Warning reliability and driver performance in naturalistic driving. *Human Factors, 50* (5), 845-852.

Thipphavong, D. P., Apaza, R. D., Barmore, B. E., Battiste, V., Belcastro, C. M., Burian, B. K., Dao, Q. V., Feary, M. S., Go, S., Goodrich, K. H., Homola, J. R., Idris, H. R., Kopardekar, P. H., Lachter, J. B., Neogi, N. A., Ng, H. K., Oseguera-Lohr, R. M, Patterson, M. D., & Verma, S. A. (2018). Urban Air Mobility airspace integration concepts and considerations. *18th AIAA Aviation Technology, Integration, and Operations Conference.* Atlanta, GA: AIAA.

Venkatesh, V., Moris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27* (3), 425-478.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors, 51* (3), 281-291.

Wang, N., Pynadath, D. V., & Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 109-116). IEEE.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors, 50* (3), 433-441.

Wickens, C. D. (2018). Automation stages & levels, 20 years later. *Human Factors, 12* (1), 35-41.

Wickens, C.D. & McCarley, J.S. (2008). *Applied Attention Theory*. CRC Press.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effect of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science, 2* (4), 352-367.

Wing, D. J., Chancey, E. T., Politowicz, M. S., & Ballin, M. G. (2020). Achieving Resilient In-Flight Management Performance for UAM Simplified Vehicle Operations. *AIAA Aviation 2020 Forum.*

Zuboff, S. (1988). *In the age of smart machines: The future of work technology and power.* New York: Basic Books.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
12/02/2020	TECHNICAL MEMORANDUM	3/01/2019-12/2/2020

4. TITLE AND SUBTITLE

Designing and Training For Appropriate Trust in Increasingly Autonomous Advanced Air Mobility Operations: A Mental Model Approach
Version 1

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Eric T. Chancey
Michael S. Politowicz

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

109492.02.07.07.07

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NASA Langley Research Center
Hampton, VA 23681-2199

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, DC 20546-001

10. SPONSOR/MONITOR'S ACRONYM(S)

NASA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

NASA/TM-20205003378

12. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Category
Availability: NASA STI Program (757) 864-9658

13. SUPPLEMENTARY NOTES

14. ABSTRACT

To enable effective human-autonomy teaming (HAT) in Advanced Air Mobility (AAM) operations, the current paper presents a theoretical framework to design and train for appropriate trust in automation. The novel contribution of this work resides in connecting the construct of trust to mental models and showing how this method could be used to enable emerging HAT concepts such as Adaptive Trust Calibration.

15. SUBJECT TERMS

Trust, Mental Models, Transparency, Advanced Air Mobility (AAM), Human-Autonomy Teaming (HAT), Adaptive Trust Calibration

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			HQ - STI-infodesk@mail.nasa.gov
U	U	U	UU	43	19b. TELEPHONE NUMBER *(Include area code)* 757-864-9658