Quantifying Calibration: Bridging Trust and Reliance in Automation Across Cultural Values and Dispositional Factors

A thesis submitted by
Evelyn S. Goroza
in partial fulfillment of the requirements for the degree of
Master of Science
in
Human Factors Engineering

Tufts University
August 2025

Adviser: Dave B Miller

Abstract

Collaborative automation systems, where the system and operator work together on a task, such as in partially automated driving (PAD) and AI-driven Clinical Decision Support (CDS) can offer operators control over the extent of automation reliance. But these rapidly developing systems— which often outpace regulation— persist the challenge in engineering systems for appropriate reliance. While decades of research on trust in automation focus on the influence of system factors and situational factors, the empirical landscape has yet to further investigate the role of operator dispositional factors— which tell the story of *who* the user is— and to establish measurement standards of calibrated use— to determine *if* reliance was indeed appropriate. To address these gaps, I investigated the impact of two dispositional traits—culture and personality—and explored how to systematically quantify instances of calibrated use, as well as misuse (overuse) and disuse (underuse).

I combined survey measures of cultural values, propensity to trust, faith in technology, and baseline trust with metrics of continuous reliance and repeated measures of in-task trust and self-rated performance using a simulation game (*Calibratio*) which modeled a collaborative, shared-goal sorting task with an adaptable system. Adult participants (N = 189) completed a solo baseline to fix operator performance at 80%. The shared task repeated in three blocks, or stages, which differed by a fixed level of automation capability (40%, 60%, or 80%), with block order counterbalanced across participants. Robust structural path modeling showed that baseline trust was shaped by Collectivism, Uncertainty Avoidance, and Faith in General Technology, while in-task trust was driven by automation capability and inversely related to self-rated performance. Reliance was predicted by Power Distance, capability, and self-rated performance. An

exploratory temporal-based analysis demonstrated reliance patterns classified by calibrated use, misuse, and disuse. These findings explore a quantitative framework for measuring automation-use calibration and provide empirical evidence of how system interactions may manifest across dispositional factors.

Acknowledgements

First, a massive thank you to everyone who stood by me during this wildly ambitious thesis adventure. Otto and I couldn't have done it without you!

This work is dedicated to:

My parents and friends – for being my constants through all the chaos (and yes, for tolerating the crazy). Dave Miller, my advisor – for not just encouraging the crazy, but making it feel like brilliance. Dan Hannon and Holly Taylor, my additional committee members – for patiently navigating the crazy and keeping me on track. Gavin McCarthy-Bui and Anne Zhao – for bringing the Unity game to life and fueling the crazy in the best way possible. Ryan Veiga – for guiding me through the statistical jungle and R coding, and for calming the crazy when needed. Elsa Ostenson, Parisa Arastu, Malu Sajith, Jacob Ratzliff, & Harlan Knightly – for your help with the usability side of Calibratio and for following me down the rabbit hole of crazy, and every brilliant author and researcher cited in my References – for being the kind of crazy I aspire to someday reach.

Table of Contents

Background

Foreword

As automation evolves, we encounter new challenges in designing human-machine systems. Over time, these systems have moved beyond simple, mechanical functions to perform more advanced roles—sometimes involving cognitive tasks like decision-making. As a result, some emergent human-machine systems are not merely technical tools but are designed to resemble social agency. This evolution raises interesting questions: What individual factors influence operators' decisions to rely on automation? And how closely does that reliance align with the system's actual capabilities?

To explore these ideas, the sub-sections of this Background section are framed around three themes:

- Yesterday: What historical steps led to the socio-technological world we inhabit today?
- Today: What do we currently know about the factors shaping trust in and reliance on automation?
- Tomorrow: Where are automation technologies headed, and what does that mean for the future of human-machine interaction?

Yesterday: A Brief History of Automation and Trust

> What historical steps led to the socio-technological world we inhabit today? This section briefly traces the history of humanity's inextricable relationships with technology, linked by cognitive capability and culture. This is followed by a discussion of a type of Sophisticated Tool known as automation, and its position within the historical Industrial advancement of society. Finally, I discuss how trust plays a crucial role in human-automation interaction, including the first seminal studies inside the Trust in Automation research domain.

From Tools to Technology

Life, it seems, always finds a way, and certain living organisms tend to seize more of these opportunities than others. For millennia, humans have exceled at manipulating our environment in the creation of tools, as Aristotle described:

> "The hand can become a claw, a fist, a horn or spear or sword or any other weapon or tool. It can be everything, because it can grasp anything or hold anything"
> Aristotle, The Animal Parts (IV, 10).

Though the scientific study of what precisely defines *tools* or *tool use* faces inherent challenges due to a lack of consensus on its definition (Mangalam et al., 2022), definitions typically involve the idea of manipulating external objects to modify the physical environment (Fragaszy & Mangalam, 2018; Mangalam et al., 2022; Osiurak et al., 2010). Approaches to understanding tool use broadly fall into three categories: The embodied approaches include the ecological perceptuomotor perspective (Gibson & Ingold, 1993)—emphasizing direct, perception-action interactions guided by environmental cues—and, an approach which stems from a neuropsychology (Goldenberg & Spatt, 2009; Johnson-Frey, 2004), which posits that conceptual knowledge is derived from prior sensorimotor experiences with familiar tools. This learned

information facilitates a more efficient acquisition of knowledge, while "avoiding reconstruction de novo of each step of the process" (Mangalam et al., 2022).

Essentially, the *perceptuomotor* approach focuses on actions guided by *perceived* information, while the *sensorimotor* approach emphasizes how *sensory* experiences shape motor behaviors. These both differ from the *disembodied* approach, which explains learned actions of tool use from *physical*, technical reasoning (Mangalam et al., 2022).

Humans, alongside several nonhuman species, engage in object manipulation; however, as Mangalam explains, humans are distinct from other tool-utilizing animals in that we have *technology*:

> "Many animals make and use tools, but humans are distinctive in the complexity, diversity, sophistication, omnipresence, and obligatory nature of our reliance on tools. Simply put, other animals use tools but only humans have technology. Indeed, humans inhabit a uniquely technological niche that we ourselves have constructed" (Mangalam et al., 2022, p. 13).

In doing so, we have constructed a so-called *technological niche* (Stout & Hecht, 2017). The Cognitive Niche Theory (Pinker, n.d.) was elaborated on by an "ontogenic niche" theory, (Stotz, 2010) which Stotz describes as a "cumulatively constructed cognitive–developmental niche: the set of epigenetic, social, ecological, epistemic and symbolic legacies inherited by the organism as necessary developmental resources (Stotz, 2010, p. 483).

Cognitive archaeologists and psychologists broadly classify modern implements like calculators and smartphones under tool use or technology (Osiurak et al., 2018; Stotz, 2010). Humanity's use of technology was thought to be inherited by cumulative cultural and cognitive development (Mangalam et al., 2022; Stout & Hecht, 2017). This can be described as Cumulative Technological Culture (CTC) (De Oliveira et al., 2019), a process of progressive diversification,

complexification, and enhancement of technological traits across generations. De Oliviera et al. (2019) found that technical reasoning (the ability to understand and reason about the physical properties of objects and how they interact) is crucial for cumulative performance, suggesting that domain-specific knowledge (i.e., technical-reasoning skills) remains critical for explaining CTC. In a sense: through social diffusion and incremental modifications (De Oliveira et al., 2019; Tomasello et al., 1993), the evolution of technology has been evoked by our cumulative practices—technology has been *handed down* via culture.

A framework which takes on the embodied approach by Osiurak et al. (2018) taxonomizes tools into three general categories graded by increasing complexity: Physical Tools, Sophisticated Tools, and Symbiotic Tools (the last of which is discussed in

Tomorrow: Research for Emerging Technologies). The most primitive of these types are Physical Tools, which emerge directly from the environment (Osiurak, 2018). For instance, stone tools used by our early ancestor species *Homo Erectus* were uncovered and identified as likely to have been used for used for carving wood or hunting in prehistoric times (Soressi et al., 2003).

According to Osiurak, these primitive tools enhanced sensorimotor abilities (Osiurak et al., 2018). As humans gained knowledge through individual discovery and learning, we developed Sophisticated Tools (Osiurak et al., 2018), which necessitate practical reasoning to solve complex problems. These been around for quite some time: By around 100 B.C., the ancient Greeks developed the Antikythera Mechanism for astronomical calculations (Freeth et al., 2006), and in 3rd century BCE, the Chinese invented the South-Pointing Chariot using differential gears (Needham, 1962).

Similar to Physical Tools, these Sophisticated tools are also intended to extend human sensorimotor capabilities. In 1978, Sheridan explained the benefits of undersea, remote-control teleoperators:

> "Teleoperators, i.e., submersibles having video and other sensors, actuators for mobility and manipulation, and remotely controlled by human operators, offer much promise for extending man's flexible, adaptable, perceiving and control capabilities[…] enabling him to extend his sensory-motor function to remote or hazardous environments"
> (Sheridan and Verplank, 1978, p. 10).

In contrast to naturalistic, Physical Tools, Sophisticated Tools create a greater distance between the creator and the user (Osiurak et al., 2018), potentially leaving the door open to a "black box" of design. This phenomenon is a driving force behind many researchers studying human-machine interactions, particularly those involving a type of Sophisticated Tool known as automation.

Automation

The *Springer Handbook of Automation* (Nof, 2023, p. 4) defines automation as "the operation of machines without human intervention, along with the enabling science and technology."

The etymological roots of the word *automation* can be traced back to Aristotle's *Physics* (Aristotle, *Physics*, Book II), where we first see the term *automatos*, which translates from Ancient Greek to "self-acting" or "occurring of itself" (Nof, 2023). Much later, in 1946, Delmar S. Harder— the Vice-President for Manufacturing at the Ford Motor Company— introduced the term "*automation*" as a nickname for the production process which linked together several automatic machines into one integrated process (Hayes, 2016) (however, some attest the first utterance of the term to a 1952 *Scientific American* article). Among many definitions and slightly different origin tales, one contemporary definition is offered by Sheridan:

> "…the term *automation* refers to (1) the mechanization and integration of the sensing of environmental variables (by aritificial sensors); (b) data processing and decision making (by computers); and (c) mechanical action (by motos or devices that apply forces on the environment)" (Sheridan, 2000, p. 9).

Parasuraman (2000) offers an even more siloed definition:

> "Automation does not simply refer to modernization or technological innovation. For example, updating a computer with a more powerful system does not necessarily constitute automation, nor does the replacement of electrical cables with fiber optics…[We] use a definition that emphasizes human-machine comparison and define automation as 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… We propose that automation can be applied to four broad classes of functions: 1) information acquisition; 2) information analysis; 3) decision and action selection; and 4) action implementation." (Parasuraman et al., 2000, p. 286).

In other words, the term automation referred to, in this context of Human Factors, focuses primarily on those which supplant an otherwise manually-performed task: It is less about the actual hardware or software, and more about the salient aspects of a machine which belongs to one or more of the four functional classes that we use to accomplish tasks and goals. In this way, automation serves as an *extension* of human performance.

Over the course of history, people have gradually integrated automation into society, marked by the transitional periods known as "Industrial Revolutions" (Figure 1). These revolutions, named after the periods of time when innovation profoundly transformed work and living, ushered in new eras of technological advancement, as Groumpos (2021) aptly describes:

> "Industrial revolutions are the transformation from old practices of powering and managing of "workplace" into new and sophisticated structures that meet the goals of modern development in order to serve better the needs of the society."

Notably, this aligns with the point in Parasuraman's definition in that there is a shared component of changing the role of the human; e.g., from manual laborer to machine operator. The first Industrial Revolution, known the *Mechanical* Industrial Revolution—or in a contemporary view, Industry 1.0—emerged in mid-18th century England, when mechanization in

the form of steam and water power, mining, and machine tools transitioned people from agrarian-based life to industrial settings (Groumpos, 2021; Mathur et al., 2022). This was followed by the *Electrical* Industrial Revolution. During the late-19thcentury, society began to harvest electrical power: an opportunity which Henry Ford seized to exponentiate mass production of Ford's automobiles by powering the assembly line system (Mathur et al., 2022).

Figure 1: Overview of the five proposed "Industrial Revolutions" throughout history.

Industrial Revolution	Time Period	Description of New Technologies
Industry 1.0: Mechanical Revolution	Mid-18th century	Steam + water power, mining, and machine tools (Groumpos, 2021; Mathur et al., 2022)
Industry 2.0: Electrical Revolution	Late-19th century	Electrical energy, mass production, the assembly line (Mathur et al., 2022; Hayes, 2016)
Industry 3.0: Automated Revolution	1950s	Introduction of personal computers, automated production, IT systems, microprocessors, wireless telephones, and the Internet (Groumpos, 2021; Mathur et al., 2022)

Industry 4.0: Digital/Information Revolution	2000s	Robotics + AI, Big Data, cloud computing, Internet of Things (IoT), AR, 3D printing for mass production, RFID, Cognitive Computing (Groumpos, 2021; Sharma et al., 2020; Taj & Zaman, 2022)
Industry 5.0: Personalization Revolution	2020s	Sustainable manufacturing, personalized services, sustainability, human-robot coordination, bionics, Smart Cities, interconnected bio-mechanical interfaces (Groumpos, 2021; Huang et al., 2022; Mathur et al., 2022; Sharma et al., 2020; Taj & Zaman, 2022)

The 1950s marked the onset of the *Automated* Industrial Revolution. It was during this period when society began to automate manual tasks, and began to transition technologies from analog to digital (Groumpos, 2021; Mathur et al., 2022). At the turn of the millennium, the *Digital* or *Information Technology (IT)* Revolution emerged, characterized by incorporation of novel technologies which, as the former revolutions did, built off of the previous revolution's existing technologies. To some scholars, a downside of Industry 4.0 is its focal emphasis on system and machine design, yet at the expense of neglecting the human factor (Alves et al., 2023; Groumpos, 2021). A better turn of events potentially awaits: Industry 5.0, the Personalization

revolution, is predicted to re-center human operators in design considerations and environmental concerns by pushing the boundaries of biological and technological innovation (see: Section Tomorrow: Research for Emerging Technologies).

In a rudimentary sense, our affinity for automation can be explained by the general principle that humans often gravitate towards machine affordances to tasks, by extending human capability: the introduction of capable systems such as those outlined (in Figure 1) can improve task consistency, accuracy, efficiency, and efficacy, which is is especially useful in repetitive, precise tasks such as industrial applications and hazardous, dangerous work such as in the operation of nuclear reactors or military applications (Kohn et al., 2021a; Takayama et al., 2009).

However, our relationship to these types of machines has multiple layers. Automation serves as an extension to human capability in the functional sense, but some automation technologies—particularly more advanced, digital systems—also play the roles of our agent counterparts in society. Reeves & Nass's Media Equation (Reeves & Nass, 1996) and the CASA paradigm (Nass & Moon, 2000) posit the view that people often treat machines "socially", as they would another human. Their studies were early demonstrations of how human-computer interaction paradigms parallel interactions in interpersonal interactions. These included studies demonstrating that people apply politeness norms and gender stereotypes to computers, and in a general sense, users responding to machines as independent entities rather than as a manifestation of their human creators (Sundar and Nass 2000). Rudimentarily, CASA posits that machines are not distinct from our social system of interactions—they are merely a part of the same society, embodying entities equivalent to non-human agents on the interaction level.

Despite its myriad benefits, automation has always held the promise of adverse consequences when design falters. In 1983, Bainbridge named what she called Ironies of Automation: “the more advanced a [system] is, the more crucial the contribution of the human operator” (Bainbridge, 1983a).

An assortment of transportation-sector systems are seemingly becoming classic examples of Bainbridge’s paradox: Tesla’s Autopilot, a Partially-Automated Driving (PAD) System has evoked numerous empirical accounts (Endsley, 2017; Lin et al., 2018; Morando et al., 2021) of operator misalignment. Endsley’s (2017) autoethnography of her experience with Tesla Autopilot noted that poor correspondence between system capability and her trust in that system was related to lack of driving training, mental model alignment, situational awareness (SA), mode confusion, and driver attention, and that future highly automated vehicle systems are predicted to exacerbate issues in SA, non-continuous control, and increased decision complexity. A survey study (Mueller et al., 2024) on regular users’ perceptions of partial automated systems (Level 2) including General Motors Super Cruise, Nissan/Infiniti ProPILOT Assist, and Tesla Autopilot found that drivers of all models are more likely to engage in non-driving-related activities while using their systems than while driving unassisted. Further, they discovered that Autopilot and Super Cruise users were also likely to engage in tasks with their eyes off the road and hands off the wheel (Mueller et al., 2024).

How do technologies like these emerge? Perhaps Icarus flew too close to the sun, and automation abuse—defined as “the automation of functions by designers and implementation by managers without due regard for the consequences for human performance” (Parasuraman & Riley, 1997)—played a role in these system failures. However, before diagnosing the root cause of these design atrophies, it may be beneficial to precisely define what constitutes a design failure.

A seminal, human-centric approach to this offered by Parasuraman and Riley (1997) distinguishes cases of operator behavior in relation to automation capability: *Misuse*, or overuse, can occur when operators place excessive trust in a system whose capability does not warrant that level of reliance. Conversely, *disuse*, or underuse, can happen if operators neglect the use of automation which forgoes any benefits or safety due to use.

Often, inappropriate use—i.e., misuse or disuse—may be explained by analogous cases of over-trust (*mistrust*) and under-trust (*distrust*). Many cases of Tesla's PAD System have been explicitly labelled as misuse: drivers were found to engage in safety-critical behaviors, including complacency over time when engaging Autopilot. This failure to monitor the system increased driver stress and mental and physical workload as drivers had to be constantly prepared for unsafe system behavior (Nordhoff et al., 2023). These respondents were observed to have developed mistrust in the system's capabilities, leading to misuse—perhaps the consequence of an "unfinished technology" (Nordhoff et al., 2023).

In general, trust itself has been extensively linked to automation use—but before digging in to this relation, it is important to first explain how trust may be approached with a systematic perspective. Applying a model leveraging Systems Engineering frameworks propose framing a 3-entity model of the human-machine-environment, human-machine-task—e.g., Holistic Requirements Model (Burge, n.d.)—to account for the salient aspects in a full model of the operator, environment, and machine. Sheridan explained how modeling cognition is a distinct process from explaining underlying mechanisms of behavior:

"…these theoretical structures used are adapted for describing what happened and from this **predicting** what will be. They make no presupposition about underlying mechanism or presupposition of behavior" (Sheridan et al., 1978).

Beyond the scope of the operator, systems integrating technical components with social, cultural, and organizational factors knowingly facilitates the design of "socio-technical systems" (STS) (Roque et al., 2025), which consider technical excellence, human value-centricity, and harmony with social, cultural, and ethical issues:

"This dynamic is becoming increasingly crucial, as STS becomes deeply integrated into society, bringing trust to the fore" (Roque et al., 2025, p. 2).

Evidently, regardless of whether a single operator or factors involving society at large are integrated into the system model, trust is a salient component of these systematic models, and the focus of the following section.

Trust

Interpersonal Trust and Trust in Automation

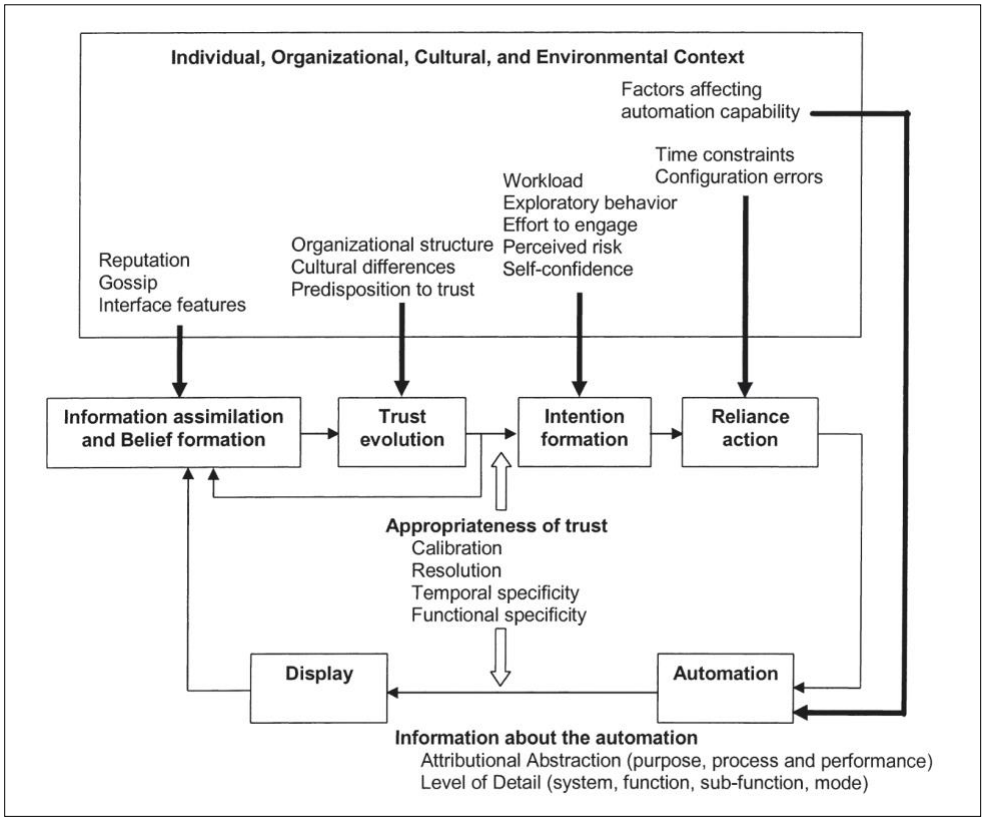
Mayer et al. (1995) define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Mayer et al., 1995). According to *The Oxford Handbook of Automation* (Meyer & Lee, 2013), the concept of Trust in Automation first surfaced in the 1970s (Sheridan & Ferrell, 1974). Sheridan (1975) and Sheridan and Hennessy (1984) argued that just as trust mediates relationships between people, it may also mediate the relationship between people and automation.

This was followed by Muir's (1987) pioneering experimental work, which provided a foundation for the Trust in Automation domain, encompassing various aspects such as: (1) assessing the extent to which models of trust between humans (interpersonal) can be generalized to human-machine trust, (2) emphasizing the need for empirical studies to predict the influence of factors on trust, and (3) offering a starting point for predicting the source of inappropriate trust to aid in system design. Sheridan (1975) and Sheridan and Hennessy (1984) argued that just as trust mediates relationships between people, it may also mediate the relationship between people and automation.

Following this, Lee & Moray's (1992) theory influenced the seminal review of systematized trust factors by Lee & See (2004). This seminal framework (**Figure 2**) systematized a model of factors driving appropriate automation use, with a focus on trust. Technically, the Lee model draws on the Theory of Reasoned Action (TRA) (Ajzen & Fishbein, 1980), a chain of constructs observed under instances of rational decision-making: belief->attitude->intention->behavior. According to Lee & See's work, trust is defined with an attitudinal approach:

> "Trust can be defined as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability. In this definition, an agent can be automation or another person that actively interacts with the environment on behalf of the person" (Lee and See, 2004, p. 51).

Figure 2: Lee and See's closed-loop cognitive model of trust in automation (Lee & See, 2004).



Trust is not the sole predictor of use, but it critically influences whether, when, and how operators engage automated systems (Hoff & Bashir, 2015a; Parasuraman & Riley, 1997). Sheridan (2002) distinguishes between trust as an outcome (effect of automation characteristics like reliability) and trust as a cause (influencing operator behavior). Trust significantly affects automation usage (Madhavan & Wiegmann, 2007, p. 280). Furthermore, system factors (e.g., transparency, perceived reliability, similarity; Razin & Feigh, 2024; Verberne et al., 2015) and operator factors can shape use jointly and separately (see:

Sources of Operator Variability).

Rempel et al. (1985) described trust as developing along three coherent dimensions: predictability (early, behavior-focused expectations based on observed consistency), dependability (confidence in the partner's underlying qualities and integrity, not just specific acts), and faith (a forward-looking belief that persists despite limited or ambiguous evidence). These perceptions arise through an attribution process in which the trustor interprets the trustee's motives and intentions—distinguishing, for example, between intrinsic and extrinsic drivers of behavior.

Lee and See (J. D. Lee & See, 2004a) extend this interpersonal account to human–automation trust, proposing three qualitative information-integration modes—analytical, analogical, and affective—that people use to assimilate evidence about an automated agent. Coupled with Rasmussen's skill–rule–knowledge (SRK) taxonomy of human performance (Rasmussen, 1983), they map trust onto three attributional bases: performance ("*is it reliably accurate and safe?"*), process ("*do I understand how it works and where it fails?"*), and purpose ("*are its goals and incentives aligned with mine?"*). A summary of these concepts are depicted in Figure 3 and Figure 4.

Figure 3. Mapping Of Trust Constructs: Interpersonal Trust Stages (Rempel, 1985) Vs. Automation Trust Bases (Lee & See, 2004)

Stage of Interpersonal Trust Development (Rempel)	**Description**	**Trust in Automation Construct** (Lee & See)	**Description**

Predictability	*Will it behave consistently?*	Performance	Accuracy, reliability
Dependability	*Can I count on its underlying characteristics?*	Process	How it achieves task
Faith	*Are its goals aligned with mine despite uncertainty?*	Purpose	Alignment of intent, incentives, and values

Figure 4. Mapping of Human Performance Behaviors (Rasmussen) vs. Information Assimilation Modes (Lee & See, 2004)

Behavior (Rasmussen)	**Description**	**Mode of Information Assimilation (Lee & See)**
Skill-based	Automatic, habitual	Affective
Rule-based	Applying if–then procedures	Analogical
Knowledge-based	Deliberate problem-solving under novelty/uncertainty	Analytical

As operators move from skill- to rule- to knowledge-based control, appropriate use of automation increasingly depends on shifting from surface cues of performance to deeper insight into process and purpose. Ideally, the objective of ethically sound automation design is to foster calibrated use (J. D. Lee & See, 2004a)—neither misuse (overtrust) nor disuse (undertrust) (Parasuraman & Riley, 1997). In regard to Lee and See's performance basis of trust, calibrated use would correspond to a level of operator use that matches the system's true capability.

A notable takeaway in this discussion of trust bases and related information-assimilation modes is that the significance lies not in their specific labels, but in the concept that trust is built upon diverse foundations, reflecting various information sources and experiences (Meyer & Lee, 2013). Despite variability in the construct, trust researchers have long-studied established methods for measurement (see Trust Measurement), in order to quantify calibrated use, which I turn to again next.

Calibrated Trust and Reliance

Lee and See (2004) define calibrated use as reliance that matches automation capability, and calibrated trust as the degree of trust that aligns with that same capability. Following Parasuraman and Riley (1997), automation use can be understood as the operator's voluntary decision to engage or disengage a system, meaning that calibration is not an abstract construct but a matter of real-time choices. Capability, in turn, reflects the system's performance on its intended goals—accuracy, efficiency, or speed—and can range from broad functions down to individual tasks. To measure calibration at a meaningful level, capability must often be decomposed into its component

subfunctions (Miller & Parasuraman, 2000), for example by clustering tasks by their information-processing stage as in the Levels × Stages model (Parasuraman et al., 2000).

Calibration is dynamic. It can vary in functional specificity (whether trust is aligned with a particular function), temporal specificity (whether alignment persists over time), and resolution (how finely changes in capability are reflected in changes in trust) (Lee & See, 2004). For instance, when calibration resolution is low, large shifts in performance yield only small changes in reliance (Figure 5).

Figure 5: Graph of Trust vs. Capability: resolution of trust calibration (Lee & See, 2004).

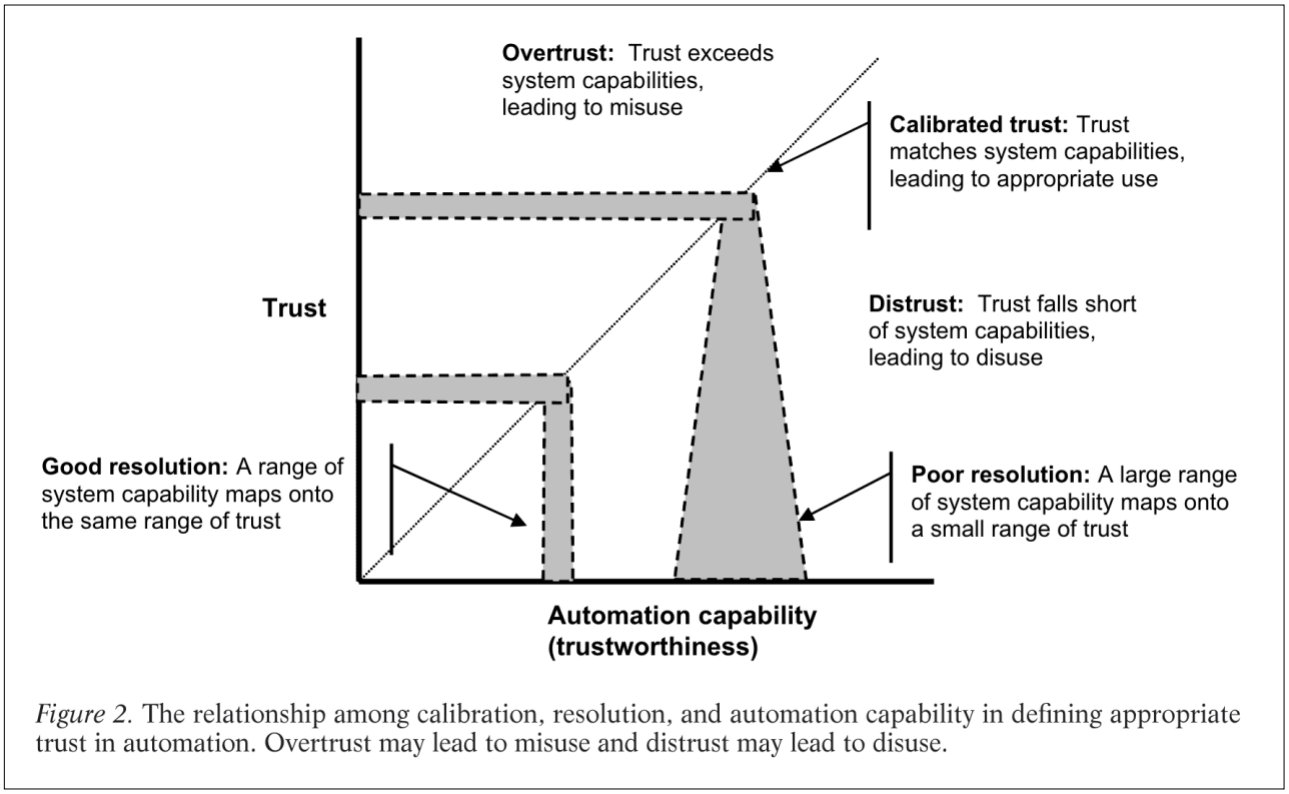
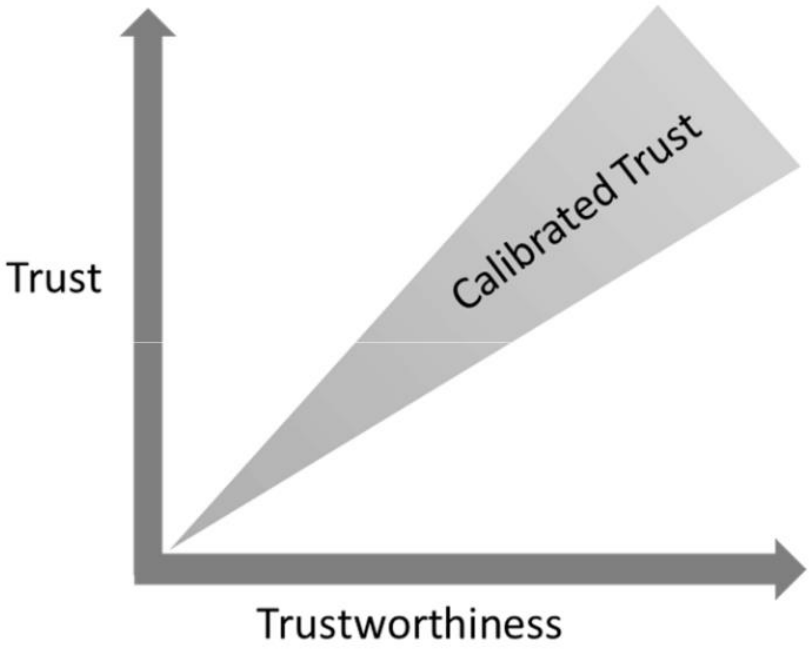


Figure 2. The relationship among calibration, resolution, and automation capability in defining appropriate trust in automation. Overtrust may lead to misuse and distrust may lead to disuse.

McDermott and Brink (2019) proposed "Calibration Points"—moments when a system excels or fails—as markers to evaluate whether trust adapts appropriately to changing performance. They

emphasize that, in practice, automation often performs inconsistently, highlighting an example canonical trend (Figure 6): as system capability increases, calibration resolution decreases.

Figure 6: As system trustworthniess or capability increases, so might the range of calibrated trust (McDermott and Brink, 2019, p. 362).



While trust helps to overcome the cognitive complexity people face in managing increasingly sophisticated automation (Lee & See, p. 54), it may also create vulnerabilities for some situations, contexts, or operators. Parasuraman and Manzey (2010) describe two cases of attention-driven, inappropriate use: (1) automation bias (AB), which reflects errors of commission (following incorrect advice) and omission (failing to act without a prompt), and (2) complacency, which reflects reduced monitoring under multitasking or high perceived reliability (Parasuraman & Manzey, 2010; Goddard et al., 2012).

In the domain of medicine, research on AI-Driven, Clinical Decision Support (CDS), some authors (Bittencourt, 2025; Khera et al., 2023) have argued that evidence of AB largely incidental and post

hoc, limiting systematic insights. Recent reporting guidance from The Journal of the American Medical Association (JAMA) was released in August 2025 regarding Chatbot Health-Advice (CHA) studies, which are "studies assessing one or more generative AI-driven chatbots for clinical evidence or health advice" (The CHART Collaborative, 2025, p. 6).

According to Huo et al. (2025), The Chatbot Assessment Reporting Tool (CHART) was "developed in accordance with the highest methodological standards through a comprehensive systematic review of CHA studies, a modified asynchronous Delphi process conducted by an international, multidisciplinary advisory committee" (The CHART Collaborative, 2025, p. 8).

At time of this writing, the CHART tool is a living clinical practice guideline, existing as a draft checklist with planned monitoring and updates on AI for the next two years (through end of 2026). David Rhew, MD, Microsoft's Global Chief Medical Officer, noted the rapid evolution of AI in clinical contexts: he suggested that opinions from only six months ago may need reassessment due to significant improvements, and the importance of AI literacy among clinicians (Perlis, 2025, p. 1).

The current discourse in AI-CDS highlights an example of a current, real-world pressing need for a systematic, quantitative approaches to reliance calibration. Currently, the CHART guidance suggests that CHA study Results (items 10a, 10b, 10c) require the following:

> "10a) Report the performance evaluation undertaken, including the alignment between generative AI–driven chatbot output and ground truth or reference standard using quantitative or mixed methods approaches as applicable.
>
> 10b) For responses deviating from the ground truth or reference standard, state the nature of the difference(s).

10c) Report the evaluation for potentially harmful, biased, or misleading responses"

(The CHART Collaborative, 2025, p. 6).

In effect, this tool asks clinicians to report the alignment between *system capability* and *ground truth*; but this still hinges on both clinicians' biases of interpreting data, and potential biases towards patients: "these [LLM dataset] biases may pertain to many factors, including but not limited to race or ethnicity, sex or gender, language, and culture". In effort to mitigate this, they implemented an open science framework, requiring Disclosures (item 12a) to include the following:

"12a) Report any relevant conflicts of interest for all authors"

(The CHART Collaborative, 2025, p. 9).

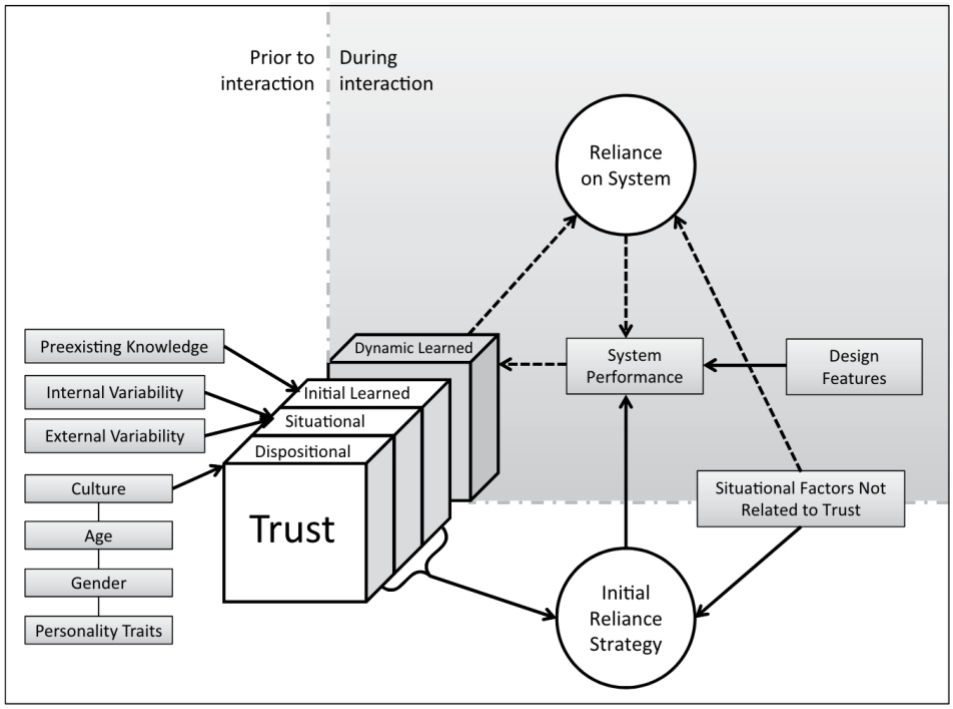
Clearly, whether calibrated trust and use occur is dependent on many factors. A model by Hoff and Bashir (2015) proposed a three-layer model that extended the Lee and See (2004) model, identifying dispositional, situational, and learned trust as interdependent layers shaped by operator, environment, and system factors (Figure 7). These attitudinal dimensions influence reliance behavior and help explain variation in calibration across contexts.

Having outlined the foundations, dynamics, breakdowns, and influencing factors of calibration, the next section examines trust layers in greater depth, situating the challenge of calibrated trust in increasingly autonomous systems within broader cognitive and socio-technical perspectives.

In practice, it is highly context-dependent who and what will lead to the incidence of calibrated trust and appropriate use. A framework by Hoff & Bashir (**Figure 7**) models trust as three inter-dependent layers—dispositional, situational, and learned trust. This model builds off of Lee

& See's work, devising an organization of factors stemming from the operator, the environment, and the system.

Figure 7: The three-layer model of trust, including: Dispositional, Situational, and Learned Trust factors (Hoff & Bashir, 2015). These attitudinal factors determine reliance behavior.



In the next section, I review some of these factors in-depth, focusing on Trust layer. Having framed how cognition, technology, and trust co-evolved through successive industrial revolutions, we now confront today's central challenge: ensuring calibrated trust in ever-more autonomous systems.

Today: Current Knowledge & Gaps in Factors Impacting Trust and Reliance

> What do we currently know about the factors shaping trust in and reliance on automation? This section discusses the current state of research on human factors impacting trust in automation. I then discuss the selected categories of Dispositional Factors—culture and personality—that I have chosen to investigate due to gaps in the literature. I then discuss recent empirical evidence investigating culture's impact on trust. Finally, I examine empirical methods of trust measurement used in the literature.

Sources of Operator Variability

Meyer & Lee (2013) proposed predicting trust based on operator factors. They emphasized that operators' trust in systems is likely influenced by the interface with the system, such as the information presented and its format, as well as operator training.

Schaefer et al. (2016) performed a meta-analysis of factors across Human-Robot Interaction (HRI) and HAI, identifying: (1) a number of research gaps in operator characteristics, including: age, trait factors, state factors, and emotive factors—such as confidence, other attitudes (e.g., respect), commitment, and comfort, as well as (2) the salience of automation performance on operator trust (Schaefer, 2016).

Correspondingly, much attention in this vein of work has focused on studying the impact of system-related factors on automation trust and use: including perceived reliability (Muir & Moray, 1996), credibility and error salience (P. Madhavan & Wiegmann, 2007), transparency (Chien et al., 2020a; van de Merwe et al., 2024), perceived attributes and anthropomorphism

(Bartneck et al., 2009; de Visser et al., 2016), and level of automation (LoA) (Bernabei & Costantino, 2024; Endsley, 2018; McWilliams & Ward, 2021), among others.

eAccording to Hoff and Bashir's human-centric model, system features are shown outside of the main focus of the diagram, with the main focus on operator factors. In this model, the layer *Dispositional Trust* is defined as a relatively stable trait layer, with four major sources of variability: *culture*, *age*, *gender*, and *personality*. Review articles (Hancock et al., 2023; Hoff & Bashir, 2015b; Kohn et al., 2021a; J. D. Lee & See, 2004a; Razin & Feigh, 2024) spanning the research landscape have been some of the only papers which highlight the lack of empirical evidence contributing to operator traits such as personality and culture. Perhaps, this may be because "implementation of a model that considers all interpersonal differences that moderate the calibration of trust is too complex to be considered for use in any practical setting" (Davis, 2019). Despite this, the review by the Australian Department of Defence (Davis, 2019) still argues there is clear evidence that individual differences strongly moderate the relationship between system trustworthiness and trust placed in it; i.e., drastic changes in system capability recalibrates trust in some operators, while leaving others unaffected (which is consistent with findings from Lee & Morray, 1994).

To consolidate this, the strategy in this review was to quantify trust with a single general construct that accounts for the variance of the relevant between-individual differences, known as propensity to trust. This trait is a measurable, valid construct with developed scales (Frazier et al., 2013; Jian et al., 2000; McKnight et al., 2011; Merritt & Ilgen, 2008a) for empirical trust ratings (see: Trust Measurement).

This raises the question: have we included all users in our discussions of trust? Have we truly been Human Centered designers?

Dispositional Factors and Culture

During the time of its conception, the CASA paradigm (Nass & Moon, 2000) was not received without critical examination: Madhavan and Wiegmann (2007) countered the view of interpersonal equivalence in human-machine trust when compared to interpersonal trust. Paired with a review of proponent cases (e.g., Dijkstra, 1999; Dijkstra et al., 1998), they constructed a framework juxtaposing the differences in human-human and human-automation in the category of Decision Support Systems (DSS), the type of cognitive automation which can relieve the operator of decision-making tasks. However, due to the seemingly exponential rate of innovation, the Madhavan framework likely does not apply to more advanced systems. As iterated by Sheridan in 2019:

> "In the future, with increasing computer "intelligence," sociological considerations of culture and morality will also become significant factors of trust in automation" (Sheridan, 2019, p. 2).

Sheridan's view, paired with a growing body of evidence (Atchley et al., 2023; Chien et al., 2020; Ge et al., 2024; Hoff & Bashir, 2015; Lee & See, 2004) indicates the view that certain operator characteristics remain under-represented in empirical models that seek to quantify trust calibration. Addressing this gap is essential, because miscalibrated trust continues to undermine system safety and acceptance across domains ranging from automated driving (Cui & Kraus, 2021) to health care and cybersecurity (Frazier et al., 2013; McKnight et al., 2011). For instance, Kraus et al. (2021) found that extraversion, neuroticism, self-esteem, general interpersonal trust, and technology affinity account for substantial variance in users' trust toward automated

vehicles, supporting a hierarchical model that links broad traits to domain-specific trust judgments.

Recent conceptual refinements challenge the assumption that dispositional trust is homogeneous. Razin and Feigh (2024) argue that dispositional trust comprises generic, pre-interaction attitudes encompassing faith in persons, institutions, and technology. They emphasize *faith in technology*—a slowly updated prior that exerts its strongest influence on shared mental models and initial trust formation, yet diminishes as operators accrue direct feedback (Razin & Feigh, 2024). Measures such as the Perfect Automation Schema (Merritt et al., 2015) capture individual differences in idealized expectations and provide one avenue for operationalizing these priors.

Culture

Culture, as defined by Hofstede and Bond, is "a state of the surrounding social system" (Hofstede & Bond, 1984, p. 1), or more elaborately:

> "Culture is the collective programming of the mind that distinguishes the members of one group or category of people from others. It can be applied to various levels, including societal, national, gender, occupational, and organizational, each with distinct characteristics" (Hofstede, 2011).

While human agents aggregate into societies, machine agents are aggregated into technological landscapes: as illustrated with respect to the theory of Cumulative Technological Culture (CTC) (De Oliveira et al., 2019), demonstrated by the history of technological changes during the Industrial Revolutions (Groumpos, 2021), and by rationale that Compters Are Social Actors (Nass & Moon, 2000). Because culture is a property emerging from systems of interacting agents, each of Hofstede's cultural dimensions reflect measures which emerge from group-level

interactions. Especially due to the ubiquity of automation technologies today, in studying human-machine interaction, we are studying a newer, *inter-agent* type of culture: that which include both machines and humans. Thus, it makes sense to understand how existing measures of interpersonal culture from groups of people fit into this new definition of culture, which encompasses our interactions with machines.

Though there are a variety of approaches to operationalizing culture, some of the more popular frameworks include Triandis' *Cultural Syndromes* comprised of *cultural complexity*, *cultural tightness*, *individualism*, and *collectivism* (Triandis, 1996), Hofstede and Bond's *Cultural Values* (Hofstede & Bond, 1984), and Leung and Cohen's *Culture x Person x Situation (CuPS)* approach (Leung & Cohen, 2024), proposing the existence of three Cultural logics —Honor, Dignity, and Face (Leung & Cohen, 2024). Each of the cultural logics groups largely clustered by how individuals value self-worth: internally (Dignity), externally (Face), or both (Honor), and by each of Hofstede's dimensions.

Importantly, Hofstede (2011) warned against stereotyping individuals based on national culture scores, emphasizing the statistical link between culture and personality and the wide variety of individual personalities within each culture. Due to difficulties with "psychometrically disappointing" results of reliability at the individual level in prior studies of dimensionalizing Hofstede's Cultural Values, Yoo et al. (2011) developed and validated the Individual Cultural Values Scale (CVScale), which translates the five dimensions into an adequately reliable, valid, across-sample and across-national generalizable scale at the individual level (Figure 8).

Figure 8Figure 8).

Figure 8: CVScale Dimensions: Hofstede's Cultural Values and corresponding definitions (Hofstede & Bond, 1984; Yoo et al., 2011).

Dimension	Definition
Power Distance	The extent to which the less powerful members of institutions and organizations accept that power is distributed unequally (Hofstede & Bond, 1984, pp. 3–4)
Uncertainty Avoidance	The extent to which people feel threatened by ambiguous situations and have created beliefs and institutions that try to avoid these. (Hofstede & Bond, 1984, pp. 3–4)
Individualism vs. Collectivism	**Individualism** = a situation in which people are supposed to look after themselves and their immediate family only **Collectivism** = a situation in which people belong to in-groups or collectives which are supposed to look after them in exchange for loyalty (Hofstede & Bond, 1984, pp. 3–4)
Masculinity vs. Femininity	**Masculinity** = a situation in which the dominant values in society are success, money, and things

	Femininity = a situation in which the dominant values in society are caring for others and the quality of life (Hofstede & Bond, 1984, pp. 3–4)
Long-Term Orientation	Long- versus short-term orientation toward planning for the future (Yoo et al., 2011, p. 194)

Empirical Studies

In the research domain of Trust in Automation, Chien et al. (Chien et al., 2014; Chien, Sycara, et al., 2016; Chien et al., 2018, 2020b) conducted a series of studies examining how culture influences automation use. Among the few empirical studies evaluating culture as a dispositional factor of automation trust, the researchers measured culture using Leung and Cohen's three cultural logics. Among these papers, they developed a culturally-sensitive trust instrument (Chien et al.,2014) and examined how cultural logics impact trust during an Unmanned Aerial Vehicle (UAV) task (Chien et al., 2016), finding significant links between these cultural logics and trust. For example, Westerners in the Dignity group exhibited the "swift trust" hypothesis, aligning with the logic of internalized valuation of self-worth prevalent in individualist cultures.

Chien et al.'s study limitations leave much for researchers to explore in the relationship between trust in automation and cultural dimensions: (1) cultural cohorts of Dignity, Face, and Honor are qualified by geographic location (US, Taiwan, and Turkey, respectively), and (2) the contextual

application. While applied practice is often useful, this task represents only one of many contexts in which the factor of operator experience interacts with performance.

Another study relevant to human-automation interaction, using the Moral Machine online survey study platform (Awad et al., 2020), investigated cross-cultural moral decision-making in hypothetical situations involving autonomous vehicles. From the large survey sample (N=70,000) arose three clusters of cultures arising from moral decision-making when faced with difficult situations of autonomous vehicle routes (which, interestingly, appear similar to Leung and Cohen's Dignity, Face, and Honor logics). Although not directly within the same human-machine trust domain, this study demonstrates findings which further underscore variability in cultural dispositions towards human-machine interactions.

Trust Measurement

Traditional measures of trust include self-report scales (such as Jian et al., 2000; J. D. Lee & Moray, 1994; Merritt & Ilgen, 2008b; and Merritt, 2011). Treating trust as an antecedent of reliance (Dzindolet et al., 2003) allows behavioral reliance to serve as a metric, as demonstrated by Miller et al. (D. Miller et al., 2016) and Fu et al. (Fu et al., 2019) in partially automated driving tasks, where the driver of a partially automated vehicle needed to demonstrate appropriate trust: either allowing the vehicle full control, or taking control of the vehicle when necessary.

Kohn et al. (Kohn et al., 2021a) list nine trust behaviors: combined team performance, outcomes, compliance/agreement rate, decision time, delegation, stakes invested, intervention, reliance, response time, and verification.

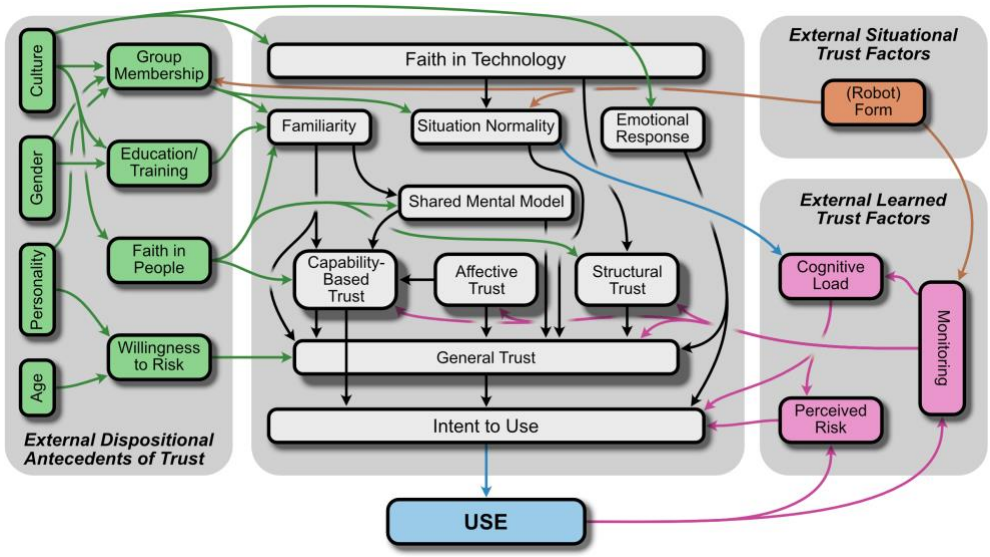
Measurement of dispositional traits has matured considerably in terms of granularity and specificity to human-automation interactions. For instance, personality traits specific to automation trust propensity—Faith in General Technology (McKnight et al., 2011), Propensity to Trust in Machines (McKnight et al., 2011), Propensity to Trust (Interpersonal; Frazier et al., 2013), Trust in a Specific Technology (McKnight et al., 2011)—measure an operator's dispositional indices of baseline automation trust.

Measuring trust calibration by alignment of trust, reliance, and system capability has attempted by some (Lucas et al., 2024; McDermott & Brink, 2019; Pop et al., 2015, Chancey, Bliss, Yamani, & Handley, 2017; Mercado, Rupp, Chen, Barber, Procci & Barnes, 2015; Merritt, Lee, Unnerstall, & Huber, 2015; Wiegmann, Rich, & Zhang, 2001).

Razin and Feigh's (2024) meta-analytic framework integrates these and more dispositional measures into Hoff and Bashir's three-layer trust taxonomy, including links between culture, personality, and dispositional faith in technology.

By foregrounding this measurement gap between highly differentiated trait measures and dynamic behavioral indices, the next section turns to the future agenda of automation systems, examining how emerging regulatory frameworks and interaction paradigms must draw on person-centered trust models to guide the safe, equitable, and effectively calibrated deployment of next-generation autonomous technologies.

Figure 9: A model of trust constructs arising from a meta-analysis of trust measures (Razin & Feigh, 2024).



Tomorrow: Research for Emerging Technologies

> Where are automation technologies headed, and what does that mean for the future of human-machine interaction? This section briefly discusses examples of practical regulatory practices in place for emergent automation technologies. I discuss the prevalence of collaborative interaction designs foreseen in these systems, and how this might foster effective human-automation interactions. In the technical context, this section supports reasoning about selecting a delegation or tasking interface, which is a crucial aspect of the system design I have chosen to explore.

Designing Novel Systems

> "Engineering relies on standards often at the expense of diversity, and civic life thrives on diversity with frequently no standards for coordination."
> - Guru Madhavan, Wicked Problems: How to Engineer a Better World (G. Madhavan, 2024)

A fundamental tension exists between the imperatives of engineering precision and the vitality of human variation. Nowhere is this opposition more salient than in the design of emerging

autonomous technologies. Ideally, human–automation teams circumvent human errors. Realistically, human–machine systems are complex, and effective design requires understanding variability in human cognition (Madhavan & Wiegmann, 2007, p. 280).

Osiurak et al. (2018)'s work follows Physical Tools and Sophisticated Tools by referring to their next generation artifacts as Symbiotic Tools: highly individualized technologies that adapt to—and are, in turn, shaped by—the user. Research on brain–computer interfaces illustrate this category, but more pressing future technologies await for researchers to impose safer and more ethical solutions. These include the assessment of cognitive health (Dergaa et al., 2024), educational practice (Triberti et al., 2024), design of decision-support systems (Marocco et al., 2024), and digital agriculture (Dara et al., 2022).

A recent article defines Artificial-Intelligence-Chatbot (AIC) Induced Cognitive Atrophy (AICICA), which refers to the potential deterioration of essential cognitive abilities resulting from an overreliance on AICs (Dergaa et al., 2024). The authors call for research to investigate the effect of AICs across individual differences, as the human-like conversational nature, and immediacy of active and/or personalized information (as compared to static information, such as search engine results) might foster a deep sense of trust and reliance in some users, which can induce changes in brain circuitry—such as decision-making processes, learning, and emotional responses. They call for studies meticulously controlling for diverse populations and contexts to gain insights into engagement with AICs to assess over-reliance and implications on cognitive functioning.

This need aligns with the general, emerging vision of Industry 5.0, which seeks to move beyond the machine-centric ethos of Industry 4.0, and towards a more human-centric paradigm.

Shneiderman (Shneiderman, 2020) describes this shift (originally used in the scope of AI) as a "Second Copernican Revolution" which places humans at the center of the design process. Rather than replacing the worker, Industry 5.0 aspires to blend human creativity and flexibility with the precision and endurance of smart machines, thereby personalizing services, enhancing sustainability, and restoring agency to end-users (Groumpos, 2021; Huang et al., 2022; Mathur et al., 2022; Taj & Zaman, 2022). Realizing this vision, however, requires adept implementation.

Regulation of Novel Automation Systems

The U.S. National Institute of Standards and Technology's voluntary AI Risk Management Framework (Tabassi, 2023) articulates broad principles but offers little specificity on how to accommodate individual differences in designing emergent systems.

Likewise, the European Commission's proposed AI Act (European Commission, 2021) defines "high-risk" AI systems according to intended use and potential harm, mandating transparency and post-market monitoring.

Dynamic and Collaborative Systems

> "Over a longer period of years, as computer control and artificial intelligence become more sophisticated, certain human functions in teleoperation may be replaced, but greater need and demand will be placed upon other human functions, and in these respects the need for improved man-computer interaction will increase, not [diminish]" (Sheridan et al., 1978).

Flexible or relational interaction paradigms are typically discussed as alternatives to human-monitored, supervisory control in terms of the type of Levels of Automation (LoA) (Parasuraman

et al., 2000). Flexible systems are generally characterized by the proportion of work being automated for a given task.

Recent discourse on the relational approach to trust posits benefits to future interaction designs. Essentially, in this type of paradigm agents are arranged into interdependent, lateral structures of social teaming, which resemble relationships similar to co-workers.

Chiou & Lee (2023) propose that the design of modern automation systems has the potential to enhance human-automation interaction viz-a-vis a "relational" approach to human-machine trust. Building on Schilbach et al.'s (2013) interactive view of social cognition, Chiou and Lee (2023) argue that relational teaming within "interdependent, lateral structures that replace Tayloristic hierarchies" (Adler, 2001; Chiou & Lee, 2023). These frameworks reframe static trust into a real-time construct of "trusting" between two-agent "dyads".

This dynamic relational framework to trust resonates with forecasts by Frey and Osborne (2013, 2017), who catalogue an expanding roster of AI agents—conversational assistants (Allen, 1999), affect-sensitive robots (Breazeal, 1999), cognitive tutors (Castelfranchi, 1998), safety-critical "co-bots" (Knight, 2003), ethically adaptive autonomous vehicles (Wagner & Arkin, 2008), and alliance-management agents (Blomqvist & Ståhle, 2004).

Flexible function-allocation and mixed-initiative interfaces have been shown to enhance situation awareness and workload balance when humans and automation collaborate on a hierarchical task model. In particular, so-called tasking or delegation interfaces (C. A. Miller & Parasuraman, 2007)--wherein the operator delegates a level of reliance on the automated system continuously throughout interaction-- are useful adaptable system designs which may improve task efficacy by mitigating operator fatigue to a level fine-tuned to their own capabilities, improving situational

awareness, and enhancing trust dynamics compared to static, supervisory control structures (C. A. Miller & Parasuraman, 2007).

This type of adaptable automation both allows the operator to control the amount of the task, and which tasks, each agent performs. Parasuraman et al 2007 propose that tasking interfaces "allow the operator to "finish the design" opportunistically in the context of use" (Miller and Parasuraman, 2007, p. 8). More recent examples of dynamic, flexible relationships includes a multilevel model of trust in human–agent teams proposed by Wildman et al. (Wildman et al., 2024), in which trust is "*multireferent*"—i.e., not only should we study users' trust in AI, *but also AI's trust in users*—and "*multilevel*," emphasizing the importance of team-level trust in relevant, emerging applications involving a team of more members than a dyad.

As discussed in Calibrated Trust and Reliance, assessing calibrated trust or use necessitates task or functional decomposition first, down to some layer of primitive actions that are executable by the event-handling and control algorithms of the to-be-controlled system in order to assess capability level of either agent with respect to a given task. Interestingly, task decomposition seems to reflect an innate way to coordinate work between supervisors & subordinates (Klein, 1998)—which serves as a useful vocabulary for issuing supervisory control instructions, reasoning about work to be performed, and organizing reports on the progress of such work (Miller and Parasuraman, 2007, p. 8).

Yet the empirical knowledge base for understanding the factors driving choices in use of adaptable systems remains thin, particularly with regard to how dispositional traits shape delegation strategies. Without this knowledge, both regulators and designers risk prescribing one-size-fits-all solutions.

However, the number of observable and latent factors converging to ecologically-valid attitudinal attributions towards an automated entity tend towards infinity as systems become more complex. This review of the literature underscores a convergence of quantitative metrics to evaluate trade-offs in system design. The exploratory research study design that follows is positioned at this intersection, in which I attempt to quantify calibrated use of automation, as well as examine the influence of cultural values and dispositional traits on this calibration—i.e., to understand how different people interact with automation.

Method

Research Questions

1. How do dispositional traits and cultural values influence trust and reliance on an adaptable, automated system?
2. Can we quantify the alignment of an operator's reliance to a system's actual capability?

Study Design

I designed a behavioral study that includes (1) survey measures to collect cultural values and dispositional trust traits, and (2) a *tasking* or *delegation interface* (C. A. Miller & Parasuraman, 2007), which I named "Calibratio" to collect real-time metrics of reliance on an adaptable system across varying levels of capability and repeated trust measurements. Details on both the survey measures and the delegation interface are detailed below.

Survey Measures

I chose to focus the survey measure battery on investigating the largely understudied dispositional factors of culture and personality as related to trust in automation. The full table of measures collected are in the **Appendix**, but the most relevant dimensions which are used in the SEM model and exploratory analyses are found in **Figure 10**.

I operationalize the factor culture using Yoo et al.'s (2011) CVscale, which includes dimensions Power Distance, Uncertainty Avoidance, Collectivism (reverse-score for Individualism), Masculinity, and Long-Term Orientation. I chose this set of traits to measure Hofstede's

dimensions of cultural values at the individual level, and to further granularize the limitation of measuring culture purely from geographic location, as done by prior studies (Awad et al., 2020; Chien, Lewis, et al., 2016).

The personality dimensions are largely based on traits which describe an operator's dispositional trust in technology. We initiated this process by investigating the most relevant trust constructs to our study, drawing on information from recent reviews of trust measurement (Kohn et al., 2021b; Razin & Feigh, 2024). The model derived by Razin and Feigh in their meta-review (Razin & Feigh, 2024) ultimately guided my selection of trust constructs and scales.

For general trust, I use Frazier et al.'s (Frazier et al., 2013) Propensity to Trust scale (coded as PTT). We also selected McKnight's (McKnight et al., 2011) Trusting Stance—General Technology scale, and renamed this to General Propensity to Trust in Machines scale (coded as GPTM) for the purpose of highlighting a the comparison with Frazier's interpersonally-oriented PTT scale. For Faith in Technology, we use McKnight's (McKnight et al., 2011) Faith in General Technology (coded as FIGT).

To assess trust, I use McKnight's (McKnight et al., 2011) Trusting Belief-Specific Technology (combining subscales for Reliability and Functionality), adapted for the automated agent in our study. I operationalize this adaptation by presenting the scale alongside a vignette briefly describing the task. Accordingly, I refer to this scale as Trust in Otto (coded as TIO). Finally, I also collected a self-assessment of performance after each round by asking participants, "If I sorted 100 shapes, how many would I sort correctly?" Because this is not a validated measure of performance, I employ this as "Self-Assessment of Performance" (coded as SA). The self-assessment of performance (SA) construct (i.e., perceived ability on a simple shape-sorting task)

was created for high context relevance to the task. Prior research (Wanous et al., 1997) shows single items can be valid for highly concrete and unidimensional constructs. Further, the item's face validity was assessed following pilot trails.

Figure 10: Measures used for SEM analysis with symbolic representation, description, and point of collection in reference to the Methodology

Scale	Measure	Symbol	Details	Collection Point
CVscale (Yoo et al., 2011)	Power Distance	PD	Composite score of each serve as predictors of baseline trust	Pre-Task Surveys
	Uncertainty Avoidance	UA		
	Collectivism	CO		
	Masculinity	MA		
	Long-Term Orientation	LTO		
Converged Reliance	Reliance	R_{conv}	Median value of reliance for last 30s of each Stage	During each Stage for each Capability Level (40, 60, 80)

Trust in a Specific Technology—Functionality and Reliability subscales (McKnight et al., 2011)	Baseline and In-Task Trust Score	TIO, T_{comp}	Composite Score of Reliability and Functionality subscales	After each Stage for each Capability Level (40, 60, 80)
Propensity to Trust (Interpersonal) (Frazier et al., 2013)	Perceived Self-Competence Rating	SA	Considering both your and Otto's performance in the last round of the game: If I sorted 100 shapes, how many would I sort correctly? If I sorted 100 shapes, I think I would sort __ shapes correctly out of 100 shapes.	After each Stage for each Capability Level (40, 60, 80)
Faith in General Technology (McKnight et al., 2011)	Faith in General Technology	FIGT	Composite score of each serve as predictors of baseline trust	Pre-Task Surveys

General Propensity to Trust Machines (McKnight et al., 2011)	Propensity to Trust Machines	GPTM		
Propensity to Trust (Frazier et al., 2013)	Propensity to Trust - Interpersonal	PTT		

Delegation Interface

I conducted a simulation-based task to study operators in a human–machine system, drawing this technique from past empirical studies (including: Alarcon et al., 2021; Capiola et al., 2024; Chien, Lewis, et al., 2016; Hussein et al., 2020; J. D. Lee & Moray, 1994b). As Lee and Moray (1994) put it: although such microworlds are orders of magnitude less complex than real-world processes, they capture the essential elements of supervisory control while enabling systematic manipulation and replication. Lee and Moray's orange-juice pasteurization microworld was "designed to capture some of the complexities of the actual work domain, but in a way that allowed a greater degree of experimental control" (J. D. Lee & Moray, 1994b, p. 156). Essentially, this trade-off balances ecological validity with the flexibility researchers need to vary conditions and repeat the task across participants.

The task, which I named "Calibratio", is a collaborative shape-sorting task built with Unity that quantifies participants' interactions with the sorting assistant, an adaptable system I named "Otto", by tracking delegation of the task load between the participant and the system. The collected data from Calibratio includes repeated measures of: (1) real-time quantification of

reliance on a continuous ratio scale, (2) post-round quantification of different dimensions of trust in Otto using a survey measure, and (2) post-round quantification of self-assessment of performance.

Game Detail

The game includes the 1 Baseline Stage, and 3 Game Stages, which are detailed below:

(1) Baseline Stage

The game begins with a Baseline Stage where the participants are given instructions to play the sorting game on their own. The goal of the Baseline Stage is two-fold: (1) for participants to get acquainted to the task and (2) to calibrate a spawn rate of the puzzle pieces for the participant's individualized performance. The spawn rate adjusts on a rolling basis to the corresponding individual performance level, setting the actual gameplay spawn rate to yield participant performance of 80% ability to sort shapes in time, thus requiring at least some delegation of the task to Otto. The same individualized spawn rate for each participant is carried on throughout the remaining Stages of the game.

The goal of Calibratio is to sort puzzle piece shapes as they travel down a conveyor belt (**Figure 11**), in order to accumulate as many points as possible. The participants sorts the shapes by pressing A, S, or D with their left hand (

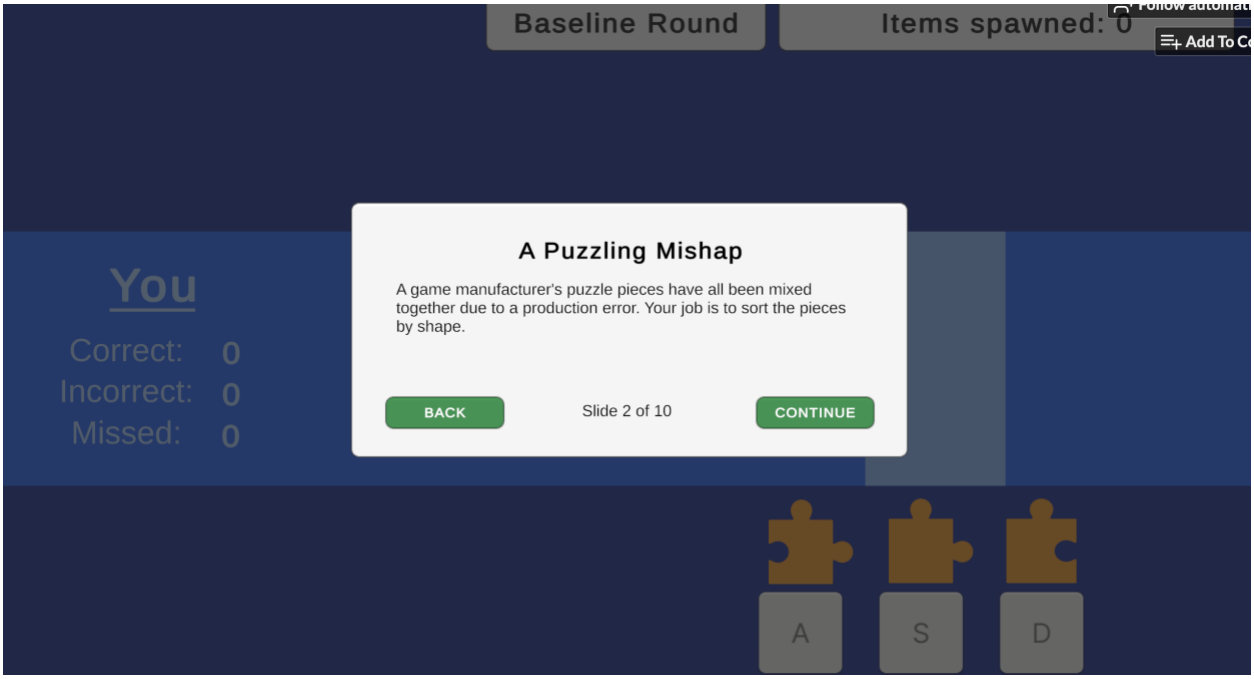


Figure 12) to sort a shape when it is in the dark blue "sorting zone". This accumulates a score displayed to the participant on the interface in the form of "Correct", "Incorrect" and "Missed". One point is given to the participant's Total for each "Correct" sorted shape, 0 for each "Incorrect" or "Missed" shape, and the accumulated tallies are updated respectively.

Figure 11: An introduction to the delegation interface poses a factory setting to induce participant engagement with a realistic, goal-oriented setting.

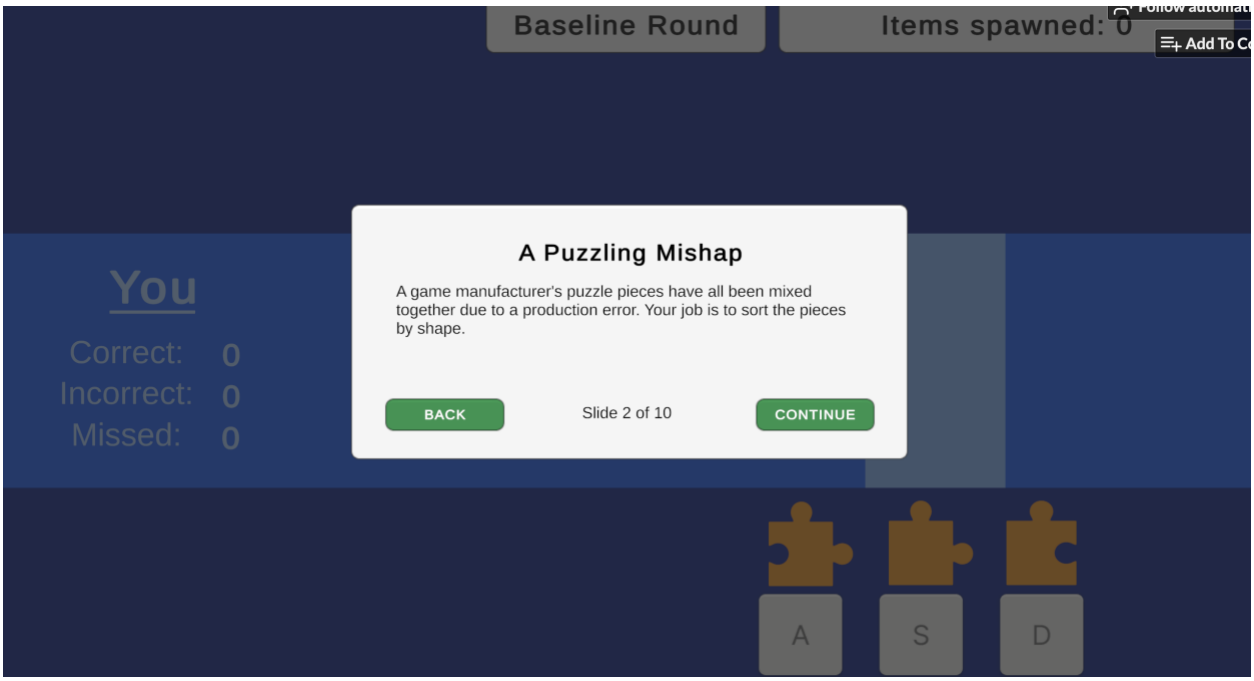
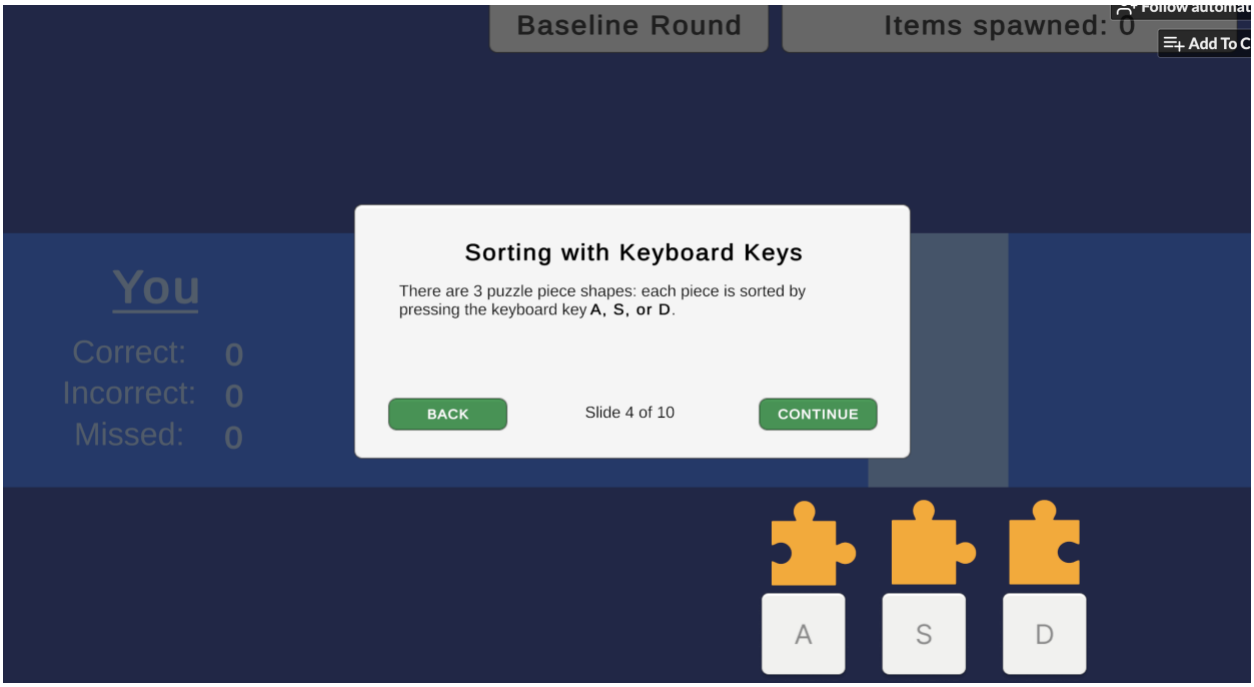


Figure 12: Instructions given to participants before the Baseline Round to sort with the A, S, and D keys.



(2) Gameplay Stages: Stage 1, Stage 2, and Stage 3

There are three Gameplay Stages in Calibratio. A sample screenshot of the game interface during the Stage 1 is shown in **Figure 13**. Gameplay round 1 starts with an introduction of Otto, the robot employed to help participants sort the shapes on a separate track. The participant is given instructions on how to delegate the proportion of "Workload" (0-100%) to Otto throughout the game, including during intermittent breaks during gameplay. The breaks are inspired by the general idea of situational awareness (SA) orientation in task pauses from Endsley's Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1988). Although we do not actually collect SA as a metric, we employ the breaks so that participants are able to better employ SA of the score board.

The tasking interface is similar to the interface in the Baseline Round, but additionally includes: (1) Otto's track above the participant's track, (2) a separate scoreboard for Otto, (3) a display of the Joint Score to depict the sum total of the participant's score and Otto's score, (4) real-time feedback in the form of X-marks and check-marks on Otto's track above the sorting zone to depict correctly sorted pieces (check marks) and missed pieces (x marks), and (5) the reliance or workload "knob", including numerical representation of their real-time delegation to either agent. During round start, the knob begins at an initial position of 100% delegated to the player, and 0 to Otto, which enables the player to calibrate their reliance based on explicit knowledge and on trust.

During gameplay, the participant can change allocation of the workload— i.e., reliance on the system— by using the up and down arrow keys with their right hand (**Figure 14**)—which provides a near-continuous ability to adjust their reliance on Otto. Otto cannot make an error in

sorting but becomes overloaded if allocated more of the task than Otto's capability level for the round. Otto's capability level varies throughout gameplay: in randomized order across participants, Otto is able to sort pieces on time with respect to three levels of capability: 40% (low), 60% (medium), or 80% (high) of the task load.

Following each round, participants are given the same scales to measure trust, using McKnight's (2011) Reliability and Functionality subscales.

Figure 13: A screenshot sample of Calibratio during Round 1, where 80% of the workload has been delegated to Otto.

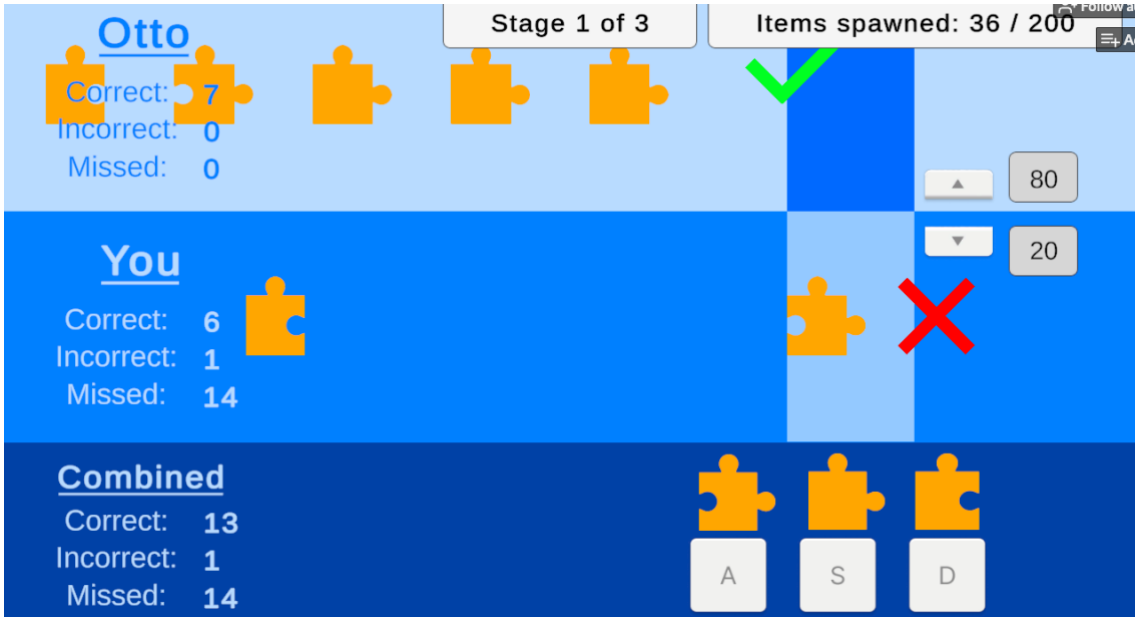


Figure 14: Instructions given to participants to change their workload/reliance on Otto using the up and down keys.

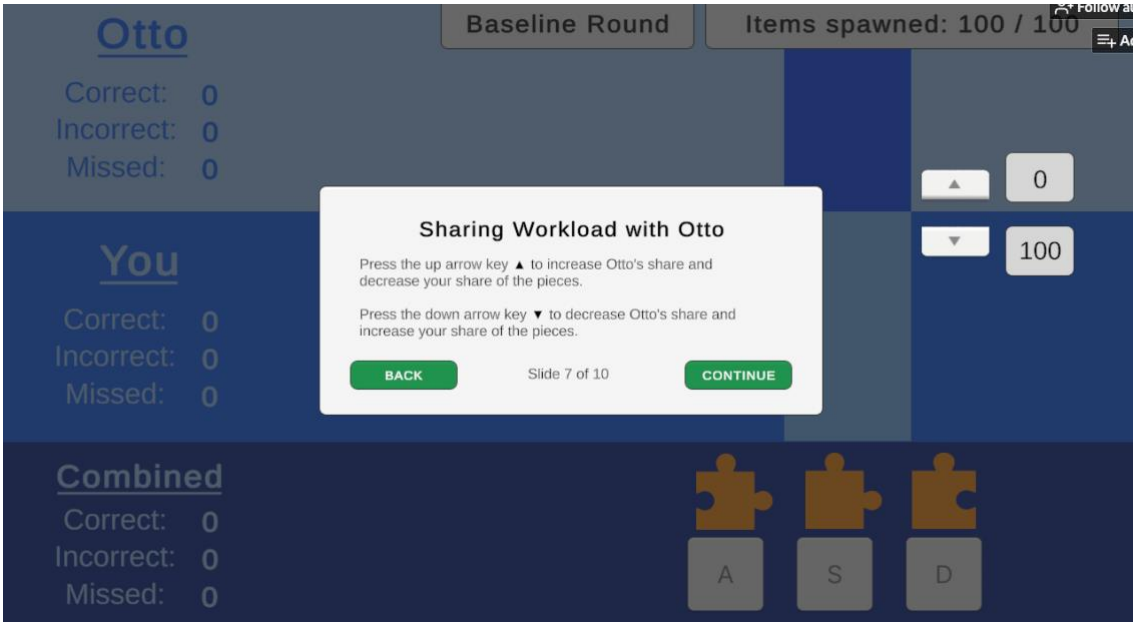
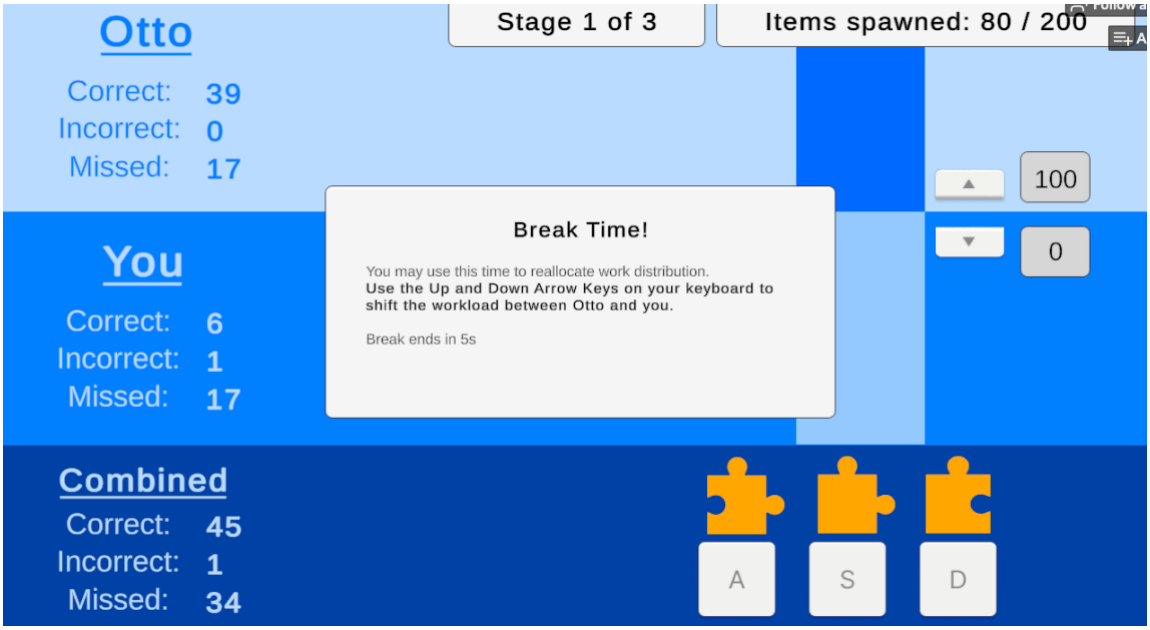


Figure 15: Example of a break during Gameplay



Task Considerations

The flexible delegation interface also addresses the pitfalls as advised by Miller and Parasuraman (2007):

> "It is critical that such an approach avoid two potential problems. First, it must make the task of instructing automation to behave as desired achievable without excessive workload. Second, it must ensure safe and effective overall behavior" (Miller and Parasuraman, 2007, p. 8).

To address these two points, we implemented the following:

(1) To mitigate excessive workload:

a. As described above, the participant's workload is fixed at 80% performance during the Baseline Stage, therefore controlling for independent skill levels which may very across individuals.

b. As described above, the delegation interface lets participants voluntarily adjust their desired reliance on Otto.

c. As described above, intermittent breaks are given to participants each round to enable better SA of the scoreboard.

(2) To ensure effective overall behavior (safety is likely less of a concern in this simulation context), Otto's reliability (performance) is designed to be highly transparent to the participant in the following ways:

(3) During each pre-round briefing, Otto's level of capability to the player prior to the start of the task (

(4) Figure 17).

(5) The score board accurately reflects Otto's real-time score during the gameplay, as well as after each round.

(6) Feedback in the form of red X's and green checkmarks are shown on Otto's track as each piece is sorted.

Figure 16: Participants are told that their the allocation of workload is voluntary.

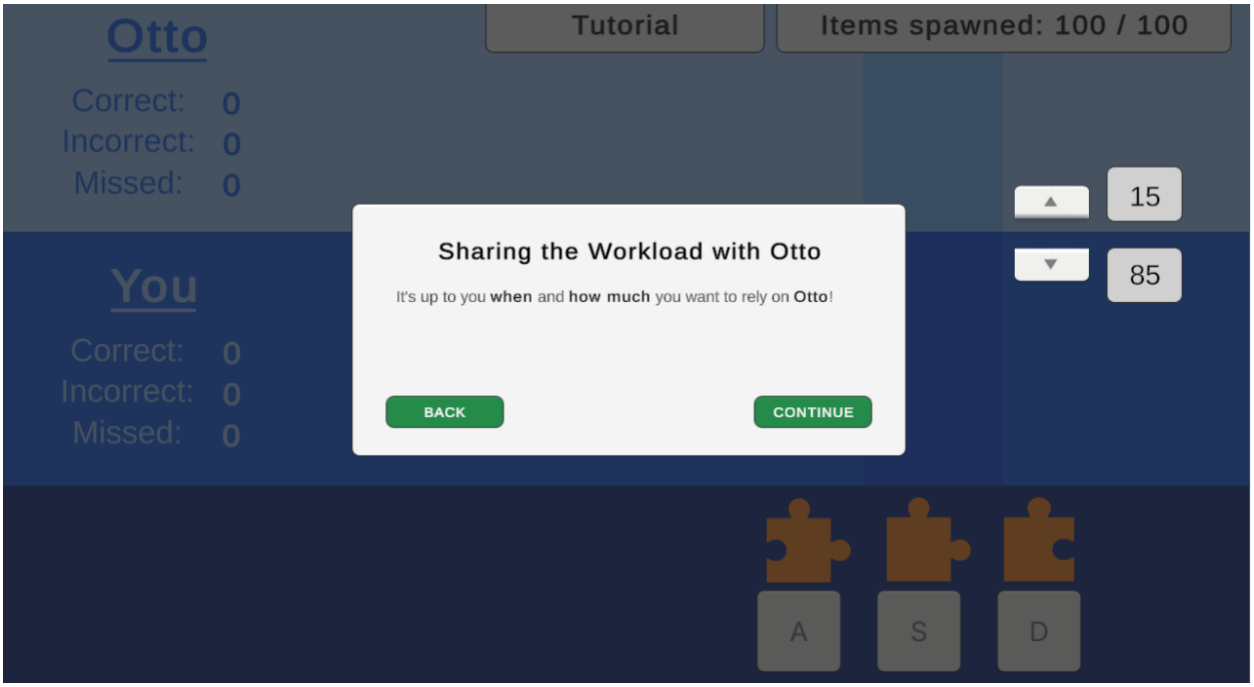
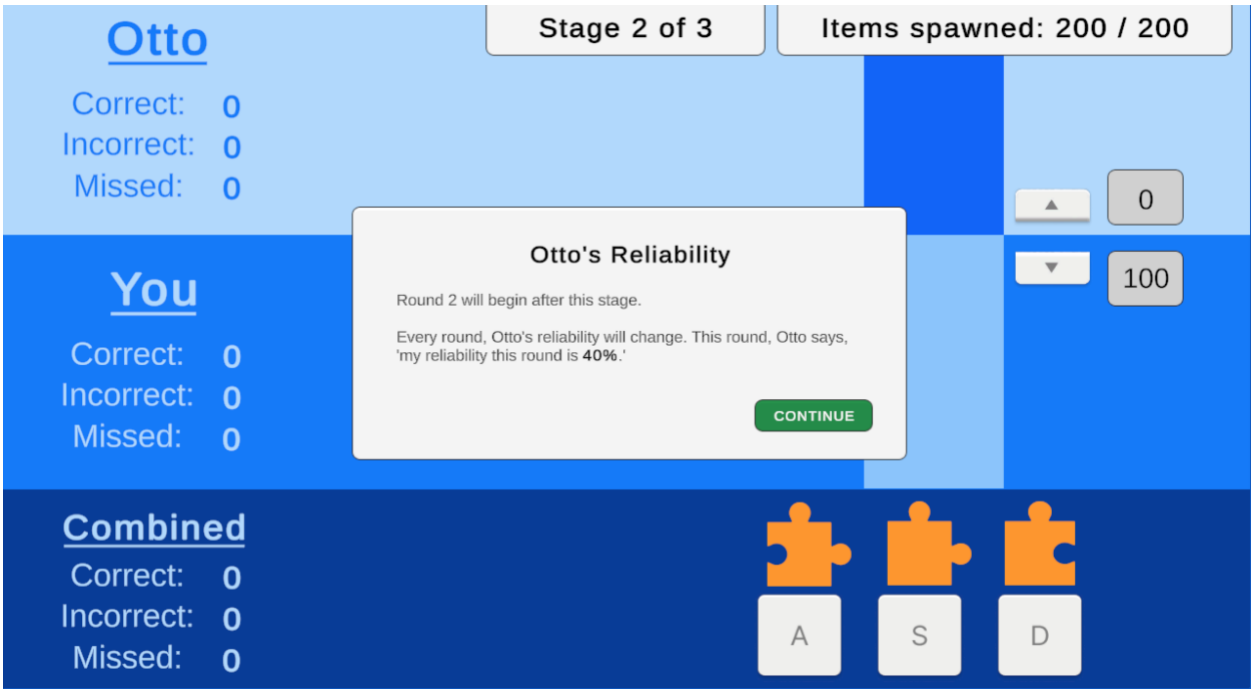


Figure 17: An example of Otto's stated reliability prior to Stage 2, which was randomly assigned as 40% for this particular instance.



Sample

Inclusion and Exclusion Criteria

The inclusion criteria for participants were as follows: they must be at least 18 years of age, located within the United States, proficient in English, and active users of Prolific.

Planned Sample and Power Analysis

I planned to recruit a representative U.S. sample of 100 participants through the Prolific platform, resulting in approximately N = 300 observations across all task stages. The target sample size was based on an *a priori* power analysis to detect a small-to-medium effect size while accounting for potential exclusions and missing data. Because GPower does not directly support power analysis for structural equation modeling (SEM), I approximated the required sample size using a linear multiple SEM model with the largest number of predictors in any SEM equation (k = 11). Using G*Power, Linear multiple SEM: Fixed model, R^2 deviation from zero, I conducted an *a priori* power analysis using G*Power 3.1.9.7 ($\alpha = .05$, $1 - \beta = .80$, $f^2 = 0.06$) indicated that a minimum of 290 participants would be needed. The effect size ($f^2 = 0.06$) was based on Cohen's guidelines (Cohen, 2009), where f^2 values of 0.02, 0.15, and 0.35 represent small, medium, and large effects, respectively. I selected 0.06 to reflect a small-to-medium effect, which is typical in trust in automation research (Hoff & Bashir, 2015b; J. D. Lee & See, 2004a). I targeted 200 participants to ensure sufficient power for the most complex part of the SEM and to allow for attrition.

Sampling Procedures

I developed a quantitative, repeated-measures experimental design conducted across participants. The study was administered using Prolific to a representative sample of US participants, stratified by ethnicity.

Analysis

Modeling Approach

SEM was chosen because it allows simultaneous estimation of direct and indirect effects (Kline, 2016), providing a holistic understanding of relationships among dispositional factors, trust, and reliance.

The SEM included three linked components:

(a) baseline trust formation modeled as a function of cultural dimensions (PD, UA, CO, MA, LTO) and dispositional trust measures (PTT, FIGT, GPTM);

(b) Overall trust after experience modeled as a function of baseline trust, system capability (cap), and updated self-assessment of performance (SA); and

(c) Reliance calibration modeled as a function of overall trust, system capability, self-assessment of performance, cultural dimensions, and dispositional trust measures. Converged reliance was quantified using the median value of reliance during the last 30 s of Stages 1–3, identified via preliminary time-series plots of a pilot study.

Structural Equation Model Specification

Guided by Lee & See's (2004) trust framework, the cultural-factor findings in Chien et al. (2016), and Hoff & Bashir's (2015) three-layer model, I specified an observed-variable path model with three layers, accounting for both direct and indirect effects:

Equation 1: Baseline Trust Formation

$$T_{comp_0} = \beta^1 \cdot PD + \beta^2 \cdot UA + \beta^3 \cdot CO + \beta^4 \cdot MA + \beta^5 \cdot LTO + \beta^6 \cdot PTT + \beta^7 \cdot FIGT + \beta^8 \cdot GPTM + \varepsilon^1$$

Equation 2: State Trust After Stage

$$T_{comp} = \gamma^1 \cdot T_{Comp0} + \gamma^2 \cdot cap + \gamma^3 \cdot SA + \varepsilon^2$$

Equation 3: Converged Reliance Calibration

$$R_{cnv} = \delta^1 \cdot T_{Comp} + \delta^2 \cdot cap + \delta^3 \cdot SA + \delta^4 \cdot PD + \delta^5 \cdot UA + \delta^6 \cdot CO + \delta^7 \cdot MA + \delta^8 \cdot LTO + \delta^9 \cdot PTT + \delta^{10} \cdot FIGT + \delta^{11} \cdot GPTM + \varepsilon^3$$

Legend

PD = Power Distance

UA = Uncertainty Avoidance

CO = Collectivism

MA = Masculinity

LTO = Long-Term Orientation

PTT = Propensity to Trust

FIGT = Faith in General Technology

GPTM = General Propensity to Trust Machines

cap = Otto's Capability

SA = Self-Assessment of Performance

T_{Comp0} = Composite Trust Score at Baseline

T_{Comp} = Composite Trust Score After Stage

$\varepsilon^1,\varepsilon^2,\varepsilon^3$ = Error Terms

Exploratory Model Specification

To quantify calibrated use, I adapted the Time-in-Range (TIR) technique from endocrinology (Wright et al., 2020). Often used as Percent-TIR (%TIR), this metric evaluates glucose levels in real-time for people wearing on-body continuous glucose monitoring devices for diabetes. With this technique, the %TIR is calculated by taking the percentage of the total time range in which the dependent variable is within a specified, desirable range of a target. Accordingly, the percent of time spent above the target band would be considered Misuse, and the converse being the percent of time spent below the target would be considered Disuse.

%TIR was defined as the percentage of time reliance remained within ±10% of Otto's capability during each stage:

Equation 4: Use-in-Range (UIR)

$$U_{ij} = \frac{C_{ij} \pm 10}{T_{ij}}$$

Legend

$U_{ij} = proportion\ of\ pieces\ allocatedd\ to\ Otto\ within \pm 10\%\ of\ C_{ij}$

$T_{ij} =$ total pieces of stage j for user i

$C_{ij} =$ Otto's capability for user i at stage j

Results

The results section includes the statistical examination of the following: sample data, survey metrics, trust and reliance metrics, correlational and covariance matrices of SEM model variables, SEM model outcomes, and exploratory model findings.

Sample Data

The final analytical sample included 189 participants, which totaled to N = 567 observations, after removing 29 cases for incomplete surveys (on cultural and dispositional trust traits) and 45 cases for incomplete game metrics (of trust and reliance outcomes). This was more than a sufficient N to detect the expected effect size, considering the *a priori* power analysis.

Survey Metrics

Due to the large number of collected variables, the distributions of only the variables used in the SEM model and the exploratory model are analyzed. Comprehensive tables of all correlation and variance matrices are found in the **Appendix**.

The survey metric summary is depicted in **Figure 18**. Comparing the means and trimmed means, the data demonstrates how all dispositional factor measures with the exception of Masculinity (MA) are robust to outliers. The difference in the mean and trimmed mean for MA (~0.2) suggests some skew or outliers in these responses, which aligns with its skew (1.351) and kurtosis (1.436), which means the Masculinity dimension is not as robust as the other dimensions to outliers, which may have slightly impacted a bias in this trait on the SEM outcomes (discussed in appropriate section below).

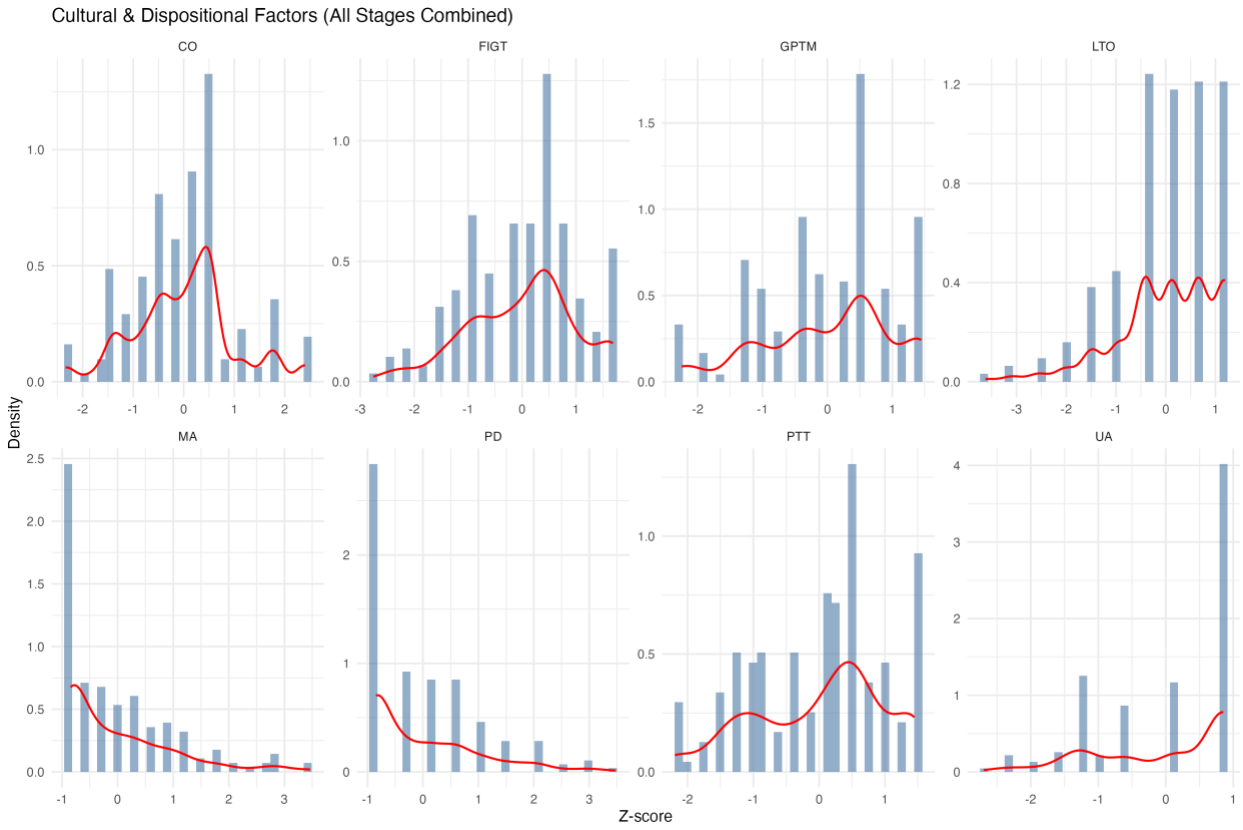
Figure 18: Descriptive Statistics

Variable	N	Mean	SD	Min	Max
T_Comp0	567	77.86	23.75	0.00	100.00
T_Comp	567	83.24	21.76	0.00	100.00
converged_reliance	567	52.18	20.06	0.00	100.00
ottocap	567	60.00	16.34	40.00	80.00
SA	567	11.95	16.61	0.00	100.00
PD	567	1.59	0.69	1.00	3.67
UA	567	4.56	0.54	2.67	5.00
CO	567	3.36	0.89	1.00	5.00
MA	567	1.95	1.05	1.00	5.00
LTO	567	4.26	0.62	2.00	5.00
PTT	567	3.41	1.09	1.00	5.00
FIGT	567	3.67	0.79	1.50	5.00
GPTM	567	3.42	1.08	1.00	5.00

The summary metrics of composite outcome variables are listed in **Table Figure** 21: Composite Trust and Reliance Summary . Across all levels of capability, participants reported high composite trust (T_{comp}) (mu = 83.24, sd = 21.76) and baseline trust (T_{comp0}) scores (mu = 77.26, sd = 23.83), with modest negative skew (-1.91 and -1.14, respectively), meaning most people clustered towards the upper end. Converged reliance across all levels of capability sits closer to the midpoint (mu = 52.18, sd = 20.06), indicating trust may not inherently translate into full automation use.

For ease of understanding the skew and kurtosis, the distribution of all dispositional factors are displayed in **Figure 19**. Notably, I likely did not capture a full range of cultural values in the sample: individuals were more likely to rate high on Long-Term Orientation and Uncertainty Avoidance, and more likely to rate very low on both Masculinity and Power Distance. The sample also captured a characteristic distribution of the Collectivism ratings, which starts to increase moving from the lower end of Collectivism scores until a steep drop observed right after the midpoint of these ratings. Furthermore, the ratings on trust propensity traits (Faith in General Technology, General Propensity to Trust Machines, and Propensity to Trust) each skewed towards the upper-middle range of the respective trait, meaning that people in this sample were more likely to have a higher tendency to trust, in both people and technology.

Figure 19: Standardized Distribution of Dispositional Factors



Trust and Reliance Metrics

Summary statistics of the Trust subscales (reliability and functionality) are shown in **Figure** **Figure** 20: Trust Subscale Summary These subscales appeared to rate very similarly (reliability: mu = 81.85, sd = 22.65; functionality: mu = 85.10, sd = 21.77) across all stages. Because the findings reflect an overlap of these system attributes (mu = 90.00, sd = 22.65, skew = -1.80, kurtosis = 3.01; and mu = 93.33, sd = 21.77, skew = -2.02, kurtosis = 3.95, respectively), it confirms that analysis of the collapsed subscale ratings (T_{comp0} and T_{comp}) is appropriate.

Figure 20: Trust Subscale Summary Figure – All Stages

var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Reliability	567	81.85	22.65	90.00	86.44	14.83	0	100	100	-1.80	3.01	0.95
Functionality	567	85.10	21.77	93.33	89.89	9.88	0	100	100	-2.02	3.95	0.91

The composite baseline (T_{comp0}) and in-task trust (T_{comp}) ratings, as well as the converged reliance ratings, across all stages are shown in **Figure Figure** 21: Composite Trust and Reliance Summary Figure, with its distribution visualized in **Figure 22**. Moving from the baseline to the in-task trust ratings in this collapsed view, more people who rated lower baseline trust in Otto tended to increase their trust after experiencing the gameplay. A further breakdown by stage for in-task trust and converged reliance is visualized in **Figure 23**, which demonstrates general stability (with a very slight, visible downwards trend) in both in-task trust and converged reliance as the stages progress from 1 to 2 to 3— however, the order of stages is agnostic to Otto's capability, due to randomization of order, so this breakdown by capability is visualized in **Figure 24**: Standardized Distribution of Converged Reliance and In-Task Trust Ratings - By Otto's Capability. A clear trend in the distribution densities are demonstrated here, in which people tended to rate higher trust in Otto as the capability increased from 40 to 60 to 80. Furthermore, the "flattening" of the converged reliance distribution is observed from 40 to 60 to

80, demonstrating a case in which there is a wider variability in automation use as the capability of the automation increases.

Figure 21: Composite Trust and Reliance Summary Figure – All Stages

var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
T_Comp	567	83.24	21.76	90.43	87.78	13.34	0	100	100	-1.91	3.50	0.91
T_Comp0	567	77.66	23.83	87.14	81.31	19.06	0	100	100	-1.14	0.44	1.00
Converged_reliance	567	52.18	20.06	55.00	53.33	14.83	0	100	100	-0.54	0.67	0.84

Figure 22: Standardized Distribution of Converged Reliance, Composite Trust, and Baseline Trust - All Stages

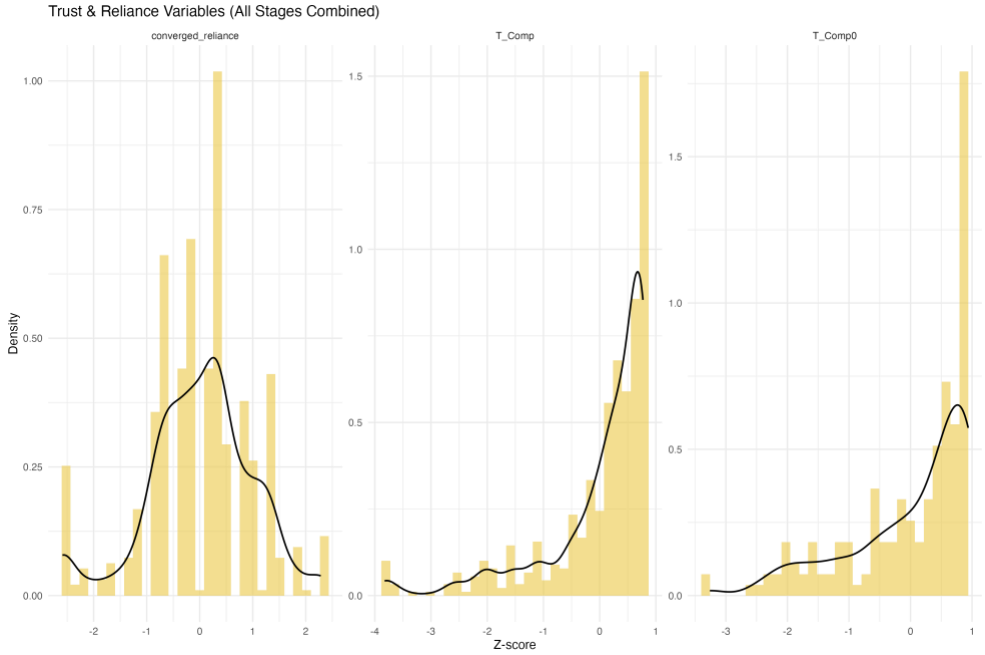


Figure 23: Standardized Distribution of Converged Reliance and In-Task Trust - By Stage

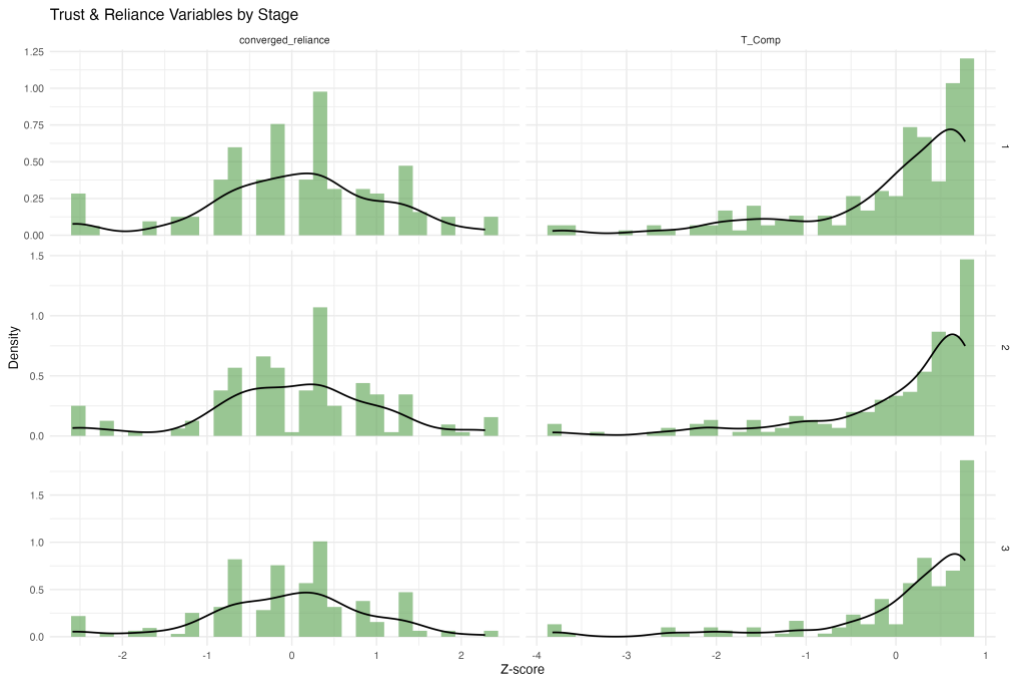
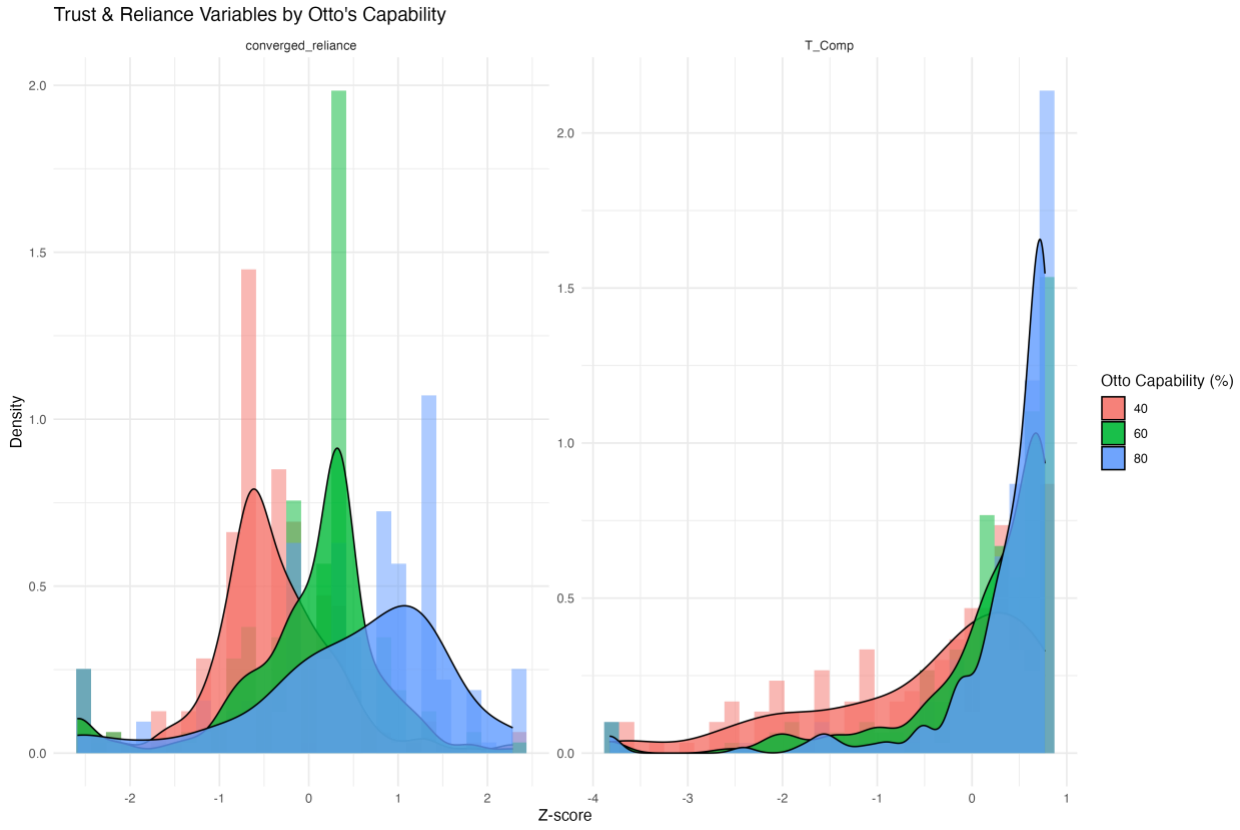
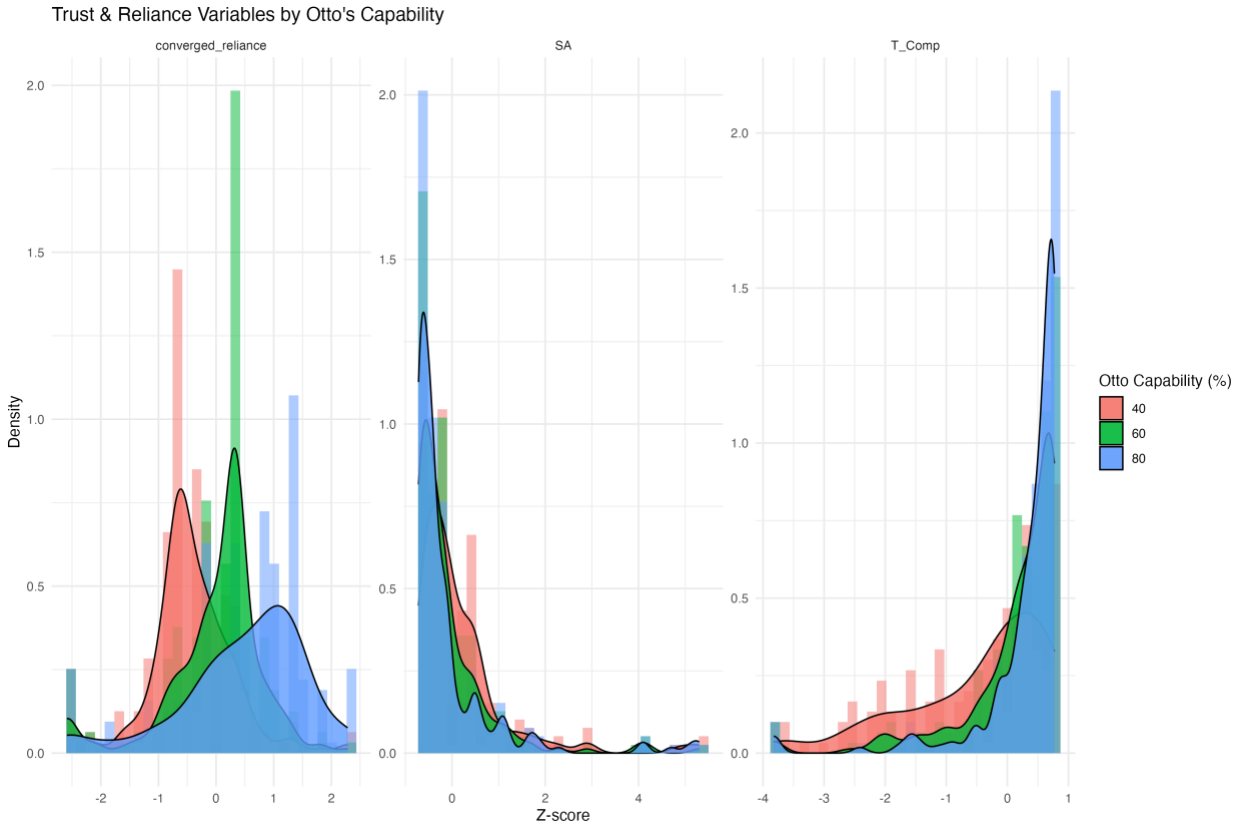


Figure 24: Standardized Distribution of Converged Reliance and In-Task Trust Ratings - By Otto's Capability



An additional visualization which includes the dimension Self-Assessment of Performance (SA) is visualized between the prior graphs in **Figure 25**. This highlights an almost opposite relationship between in-task trust and self-assessment, and an opposite trend with converged reliance. Further, this shows the same trend in less variability in self-assessment of performance as the automation capability increases. People almost always rated themselves and Otto orthogonally, and tended to rate themselves lower as Otto's capability increased. Albeit a weaker relationship, this trend of self-assessment of performance is also reflected in their behavioral reliance on Otto.

Figure 25: Standardized Distribution of Converged Reliance, Self-Assessment, and In-Task Trust Ratings - By Otto's Capability



SEM Model Outcomes

I estimated a structural equation model (SEM) in R (lavaan 0.6-19) with robust maximum likelihood (MLR; Yuan & Bentler, 2000; Rosseel, 2012) on N = 567 observations with Full Information Maximum Likelihhood (FIML) (Enders, 2010). Endogenous variables were baseline trust (T_Comp0), in-task trust (T_Comp), and converged reliance. Exogenous predictors included dispositional trait and cultural factors (PD, UA, CO, MA, LTO, PTT, FIGT, GPTM) and task/context variables (ottocap = Otto/system capability; SA = self-assessment of performance). Missing data were handled using FIML (Enders, 2010; Little & Rubin, 2019). I report robust fit indices (scaled χ^2, CFI, TLI, RMSEA). For inference on path and indirect effects, I obtained bootstrap 95% confidence intervals (1,000 resamples) from a parallel ML bootstrap run (Preacher & Hayes, 2008). Bootstrap replaces analytic/robust SEs in that run, so fit indices are taken from the MLR run and CIs from the bootstrap run (Efron & Tibshirani, 1993; Preacher & Hayes, 2004, 2008).

Model Fit Indices

All indices (**Figure 26**) indicate good fit against conventional cut-offs, indicating that the model fits the data well.

Figure 26: Global Fit Indices (Robust, MLR)

Model	χ^2 (p)	df	RMSEA	RMSEA 90% CI	SRMR	CFI	TLI	Overall
SEM	16.19 (p = .134)	11	0.029	0.000–0.057	0.017	0.985	0.954	Good
Interpretation			Good		Good	Good	Good	Good

Explained Variance

Variance explained was modest but meaningful: R^2(T_Comp0) = .208; R^2(T_Comp) = .228; R^2(converged reliance) = .222 (**Figure**).

Standardized Structural Path Estimates

The standardized path estimates with p-values are depicted in **Figure 28.** A path model diagram is depicted in **Figure 29.**

Baseline Trust (T_Comp0)

I found higher baseline trust among participants with stronger:

- Collectivism (CO): β=.231, p<.001
- Faith in General Technology (FIGT): β=.185, p<.001

- General Propensity to Trust Machines (GPTM): β=.136, p=.016
- Uncertainty Avoidance (UA): β=.135, p=.003

Other dispositional paths to T_Comp0 (PD, MA, LTO, PTT) were not significant.

In-Task Trust (T_Comp)

In-task trust (T_Comp) increased with baseline trust (T_Comp0) (β=.217, p<.001) and capability (ottocap) (β=.304, p<.001) and decreased with self-assessment of performance (SA) (β=−.272, p<.001).

Converged Reliance (converged_reliance)

Converged reliance increased with in-task trust (β=.151, p=.022) and capability (β=.291, p<.001) and decreased with SA (β=−.188, p<.001). Two dispositional predictors had smaller but significant direct links to reliance: PD positive (β=.088, p=.023) and MA negative (β=−.107, p=.025). All other dispositional paths to reliance were non-significant.

Figure 27: R-squared (Endogenous Variables)

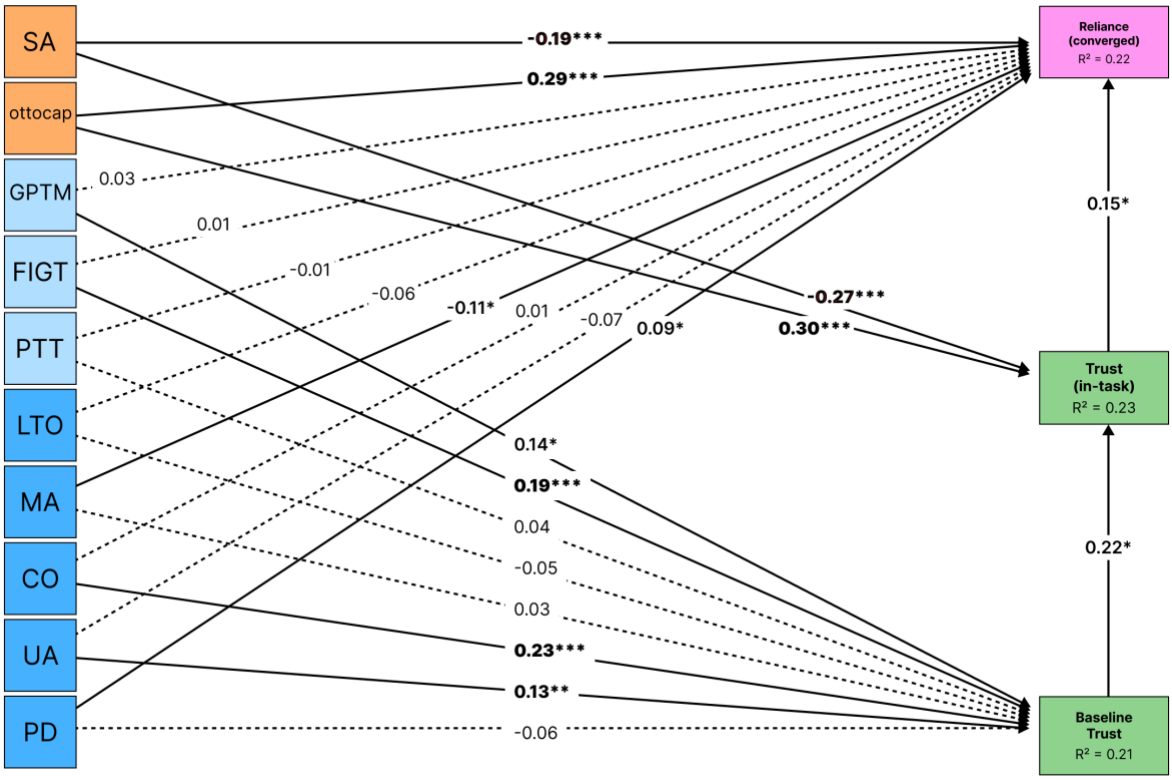
Variable	R2
Baseline Trust	0.208
In-Task Trust	0.228
Converged Reliance	0.222

Figure 28: Structural Paths (B, SE, 95% CI, β, p)

Outcome	Predictor	B	SE	95% CI	Beta	p	Sig
T_Comp0	PD	-1.913	1.460	[-4.831, 1.149]	-0.055	0.190	
T_Comp0	UA	5.904	1.972	[2.197, 9.666]	0.135	0.003	**
T_Comp0	CO	6.144	1.046	[4.011, 8.209]	0.231	<.001	***
T_Comp0	MA	0.621	1.195	[-1.521, 3.063]	0.027	0.603	
T_Comp0	LTO	-1.929	1.515	[-4.801, 1.186]	-0.050	0.203	
T_Comp0	PTT	0.929	0.897	[-0.757, 2.782]	0.043	0.300	
T_Comp0	FIGT	5.582	1.467	[2.805, 8.507]	0.185	<.001	***
T_Comp0	GPTM	2.972	1.234	[0.503, 5.268]	0.136	0.016	*
T_Comp	T_Comp0	0.199	0.046	[0.108, 0.285]	0.217	<.001	***
T_Comp	ottocap	0.405	0.054	[0.297, 0.506]	0.304	<.001	***
T_Comp	SA	-0.356	0.090	[-0.541, -0.192]	-0.272	<.001	***
converged_reliance	T_Comp	0.139	0.061	[0.017, 0.258]	0.151	0.022	*
converged_reliance	ottocap	0.357	0.056	[0.239, 0.461]	0.291	<.001	***
converged_reliance	SA	-0.227	0.056	[-0.337, -0.113]	-0.188	<.001	***
converged_reliance	PD	2.555	1.128	[0.453, 4.814]	0.088	0.023	*

Outcome	Predictor	B	SE	95% CI	Beta	p	Sig
converged_reliance	UA	-2.436	1.466	[-5.242, 0.538]	-0.066	0.097	
converged_reliance	CO	0.187	0.865	[-1.457, 1.953]	0.008	0.829	
converged_reliance	MA	-2.038	0.911	[-3.907, -0.316]	-0.107	0.025	*
converged_reliance	LTO	-2.042	1.409	[-4.721, 0.712]	-0.063	0.147	
converged_reliance	PTT	-0.249	0.836	[-1.892, 1.384]	-0.014	0.766	
converged_reliance	FIGT	0.276	1.288	[-2.326, 2.787]	0.011	0.830	
converged_reliance	GPTM	0.591	0.876	[-1.039, 2.282]	0.032	0.500	

Figure 29: Path Model Diagram



Indirect and Total Effects

The indirect ("ind_") and total ("tot_") effects are displayed in **Figure 29**.

I observed a small but reliable mediation from capability through trust to reliance (ottocap → T_Comp → reliance, indirect=0.056, p=.022). The SA → reliance path via trust trended negative (indirect=−0.050, p=.061). For total effects on reliance (unstandardized), capability was the largest positive driver (0.413, p<.001), SA the largest negative (−0.277, p<.001), with PD (2.502, p=.028) and MA (−2.021, p=.028) also significant.

Figure 30: Indirect and Total Effects (with 95% CI)

Effect	Estimate	SE	95% CI	p	Sig
ind_PD	-0.053	0.055	[-0.191, 0.033]	0.334	
ind_UA	0.163	0.115	[0.008, 0.427]	0.155	
ind_CO	0.170	0.100	[0.016, 0.408]	0.090	
ind_MA	0.017	0.038	[-0.048, 0.107]	0.653	
ind_LTO	-0.053	0.056	[-0.189, 0.033]	0.336	
ind_PTT	0.026	0.031	[-0.027, 0.101]	0.401	
ind_FIGT	0.155	0.097	[0.013, 0.394]	0.111	
ind_GPTM	0.082	0.064	[0.002, 0.245]	0.195	
ind_ottocap	0.056	0.025	[0.007, 0.108]	0.022	*
ind_SA	-0.050	0.027	[-0.107, -0.005]	0.061	
tot_PD	2.502	1.140	[0.352, 4.756]	0.028	*
tot_UA	-2.272	1.459	[-5.051, 0.608]	0.119	
tot_CO	0.357	0.875	[-1.320, 2.150]	0.683	
tot_MA	-2.021	0.921	[-3.910, -0.237]	0.028	*

Effect	Estimate	SE	95% CI	p	Sig
tot_LTO	-2.095	1.414	[-4.743, 0.648]	0.138	
tot_PTT	-0.223	0.840	[-1.919, 1.415]	0.790	
tot_FIGT	0.431	1.290	[-2.212, 2.968]	0.738	
tot_GPTM	0.673	0.876	[-0.932, 2.370]	0.442	
tot_ottocap	0.413	0.049	[0.313, 0.503]	<.001	***
tot_SA	-0.277	0.063	[-0.400, -0.154]	<.001	***

Correlation Matrix (lower triangle; * p<.05, ** p<.01, *** p<.001)

Variable	T_Comp0	T_Comp	converged_reliance	ottocap	SA	PD	UA	CO	MA	LTO	PTT	FIGT	GPTM
T_Comp0													
T_Comp	0.21***												
converged_reliance	0.05	0.31***											
ottocap	0.00	0.33***	0.36***										
SA	0.03	-0.30***	-0.27***	-0.10*									
PD	-0.04	-0.01	0.02	-0.00	0.06								
UA	0.26***	0.11*	-0.04	-0.00	-0.05	-0.02							
CO	0.28***	0.05	-0.01	0.00	0.09*	0.02	0.13**						
MA	-0.00	-0.02	-0.09*	0.00	0.12**	0.49***	-0.02	-0.04					
LTO	0.12**	-0.00	-0.09*	0.00	0.03	0.03	0.29***	0.15***	0.09*				
PTT	0.12**	0.09*	-0.00	0.00	0.05	-0.12**	0.01	0.04	-0.16***	0.15***			

FIGT	0.35***	0.17***	0.01	-0.00	-0.01	0.04	0.37***	0.16***	0.08	0.33***	0.16***	
GPTM	0.31***	0.17***	0.02	0.00	0.03	0.01	0.30***	0.05	0.05	0.18***	0.32***	0.65***

Exploratory Analysis Outcomes

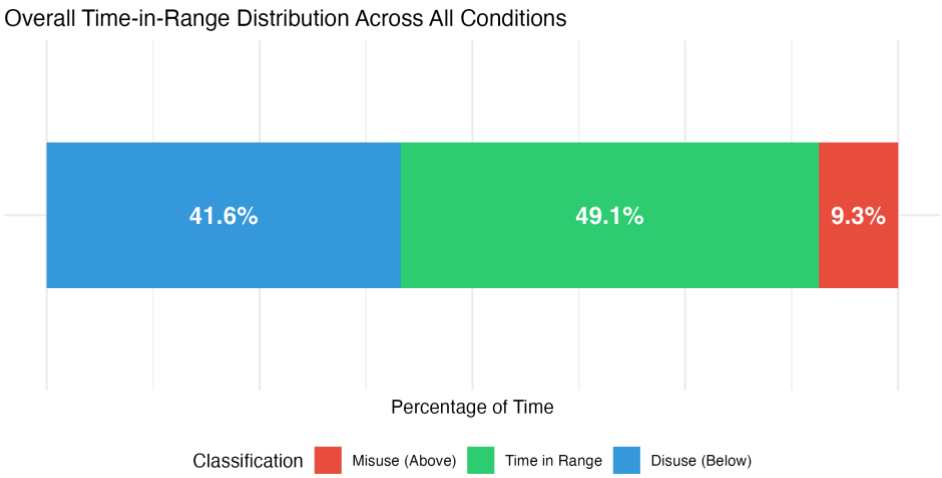
The exploratory analysis to investigate %TIR quantifies the percent of time people fell into the defined ranges of calibrated use, misuse, and disuse across all conditions (across stage and across Otto's capability level). The summary stats for these metrics are depicted in **Figure 31**. The corresponding visualization is depicted in a distribution bar chart (**Figure 32**).

The mean percent time of calibrated use across all stages and capability levels was 49.09% (SD = 39.00%). The median percent time of calibrated use was 52.43%. The mean percent time of misuse (reliance exceeding 10% of Otto's capability) was 9.31% (SD = 21.00%); and the mean percent time of disuse (reliance below 10% of Otto's capability) was 41.60% (SD = 39.85%). Regardless of the stage or Otto's capability, operators tended to calibrate use to Otto's capability about half of the time. About two-fifths of the time, operators would disuse Otto; and rarely, less than one-tenth of the time, operators would misuse Otto, or rely on Otto more than his capabilities warranted.

Figure 31: Summary Stats of Percent Time Calibrated Use, Misuse, Disuse - All Conditions

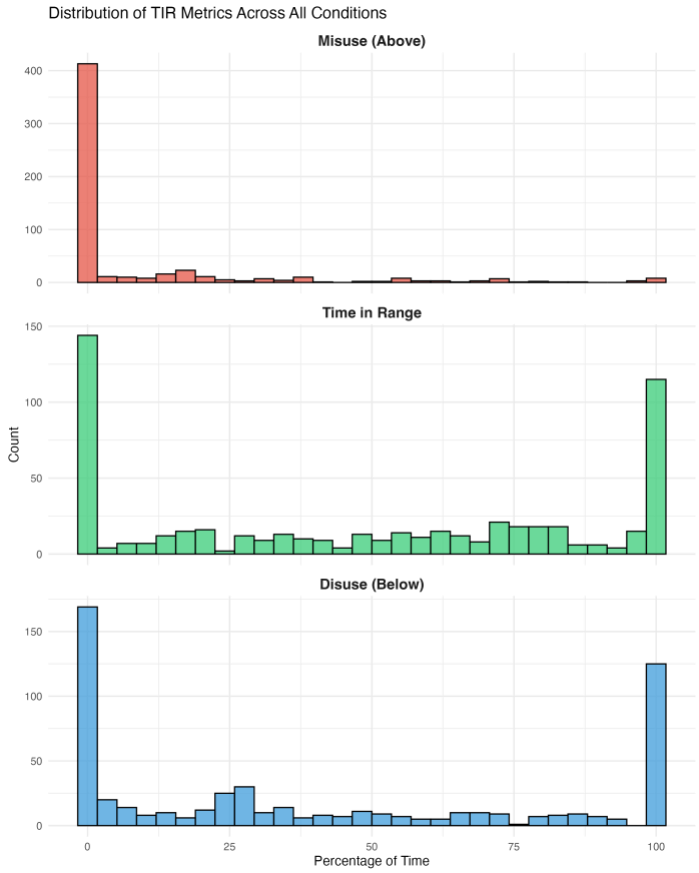
Mean Calibrated Use (SD)	**Median Calibrated Use**	**Mean Misuse (SD)**	**Mean Disuse (SD)**
49.09 (39.00)	52.43	9.31 (21.00)	41.60 (39.85)

Figure 32: Distribution Bar Chart of Calibrated Use, Misuse, and Disuse - All Conditions



The distribution plots of calibrated use, misuse, and disuse for all conditions, segmented by each use case, are depicted in **Figure 33**. There appears to be a similar trend in the percent of time spent in calibrated use as there is in disuse: where most operators either spent most of the time calibrating reliance on Otto to Otto's actual capability, or—to a greater extent—spent most of the time using Otto less than Otto's capabilities warranted. Furthermore, operators spent hardly any time misusing Otto.

Figure 33: Distribution of Calibrated Use, Misuse, and Disuse by Use Case - All Conditions



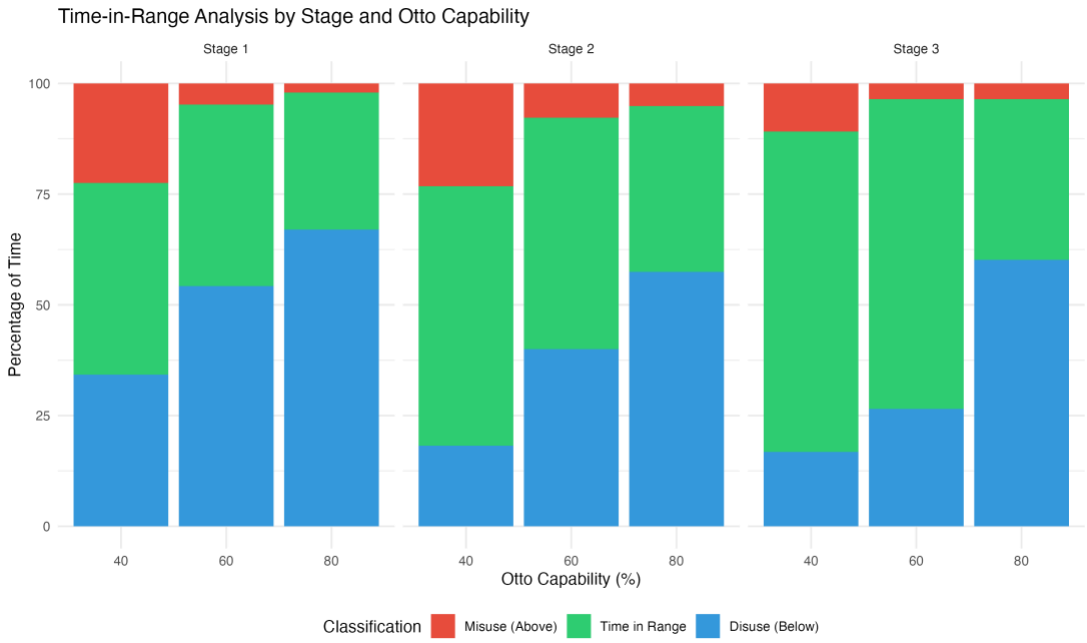
The summary stats segmented by stage and capability of calibrated use, misuse, and disuse are depicted in **Figure 34**. A corresponding visualization is depicted in a distribution bar chart (**Figure 35**). Even with the segmentation by Otto's capability and stage, similar patterns emerge as the collapsed metrics: most of the time, operators spent calibrating reliance on Otto to Otto's capabilities, followed by time spent disusing Otto, and least frequent of all did operators spend time misusing Otto. However, for capability levels 40 and 60, as operators progressed from stage 1 to 2 to 3, they tended to calibrate use to Otto's capabilities better, as shown by the enlarging proportion of time spent in calibrated use. And, across all capability levels as operators moved

from stage 1 to 2 to 3, there is a trend towards increasing the amount of time calibrating use to Otto's capabilities.

Figure 34: Calibrated Use, Misuse, and Disuse - By Stage and Capability

Stage	Capability	N	Mean % TIR (SD)	Mean % TAR (SD)	Mean % TBR (SD)
			Calibrated Use	Misuse	Disuse
1	40	67	43.3 (30.3)	22.5 (27)	34.2 (27.7)
	60	60	41 (34.1)	4.8 (9.5)	54.2 (35.2)
	80	62	30.9 (33.2)	2.1 (9.1)	67 (34.2)
2	40	61	58.6 (35)	23.2 (30.2)	18.2 (30.1)
	60	64	52.3 (41.8)	7.7 (21.3)	40 (40.8)
	80	64	37.4 (37.7)	5.1 (17.6)	57.5 (40.3)
3	40	61	72.3 (36.3)	10.8 (22.2)	16.8 (32.3)
	60	65	69.9 (36.9)	3.6 (14.7)	26.5 (36.1)
	80	63	36.3 (42.4)	3.6 (13.7)	60.1 (43.8)

Figure 35: Calibrated Use, Misuse, and Disuse - By Stage, Capability



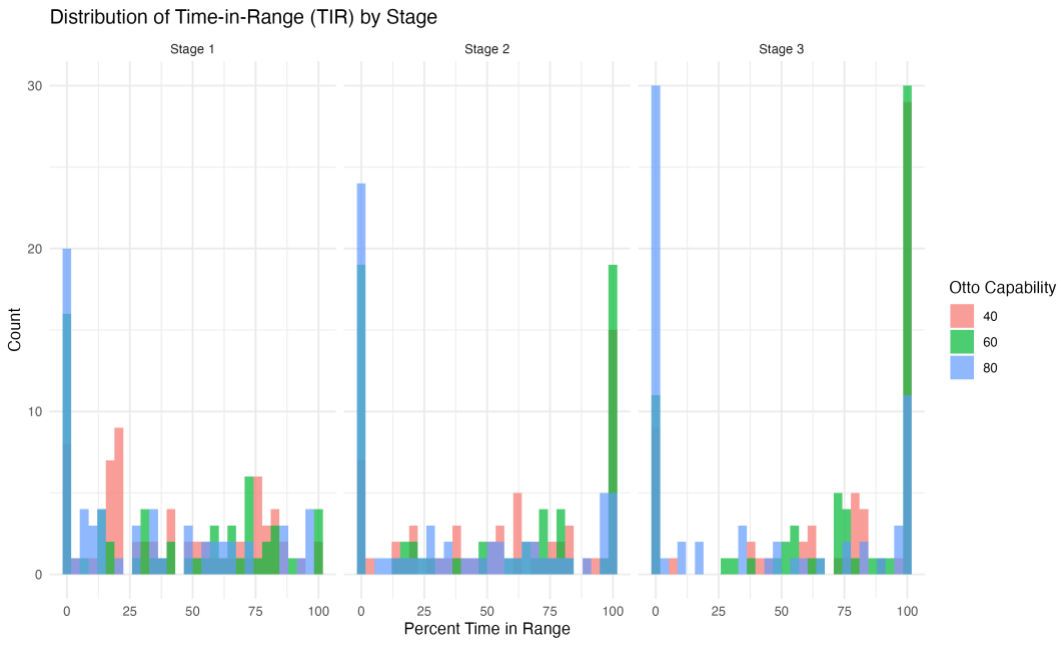
Correlation Matrix

	%TIR	T_{comp}	R_{conv}	cal_diff
%TIR	1.000	0.178	0.263	0.441

A frequency plot (**Figure 36**) depicts the frequency of calibrated use only, by stage and capability. In stage 1, operators who received the treatment of capability at 80 tended to spend less time in calibrated use compared to operators who received the 40 and 60 treatment levels at stage 1. In stage 2, operators who received the treatment level of 80 also showed this pattern of less time spent in calibrated use. By contrast, operators who received the treatment level of 60 at stage 2 tended to spend most time in calibrated use. The pattern in stage 2 is pronounced further

in stage 3. In all stages, operators tended to not spend any time in calibrated use when Otto's capability was at the lowest level of 40%. A small exception of this is depicted in stage 1, where some operators who received the lowest capability level on the first round of the game spent some (<25%) of the time in calibrated use.

Figure 36: Frequency Distribution of Calibrated Use - By Stage and Capability



Discussion

Despite the large body of data collected and analyzed in this empirical work, many clear patterns arose in analyzing both the impact of cultural and dispositional traits on trust and reliance in the adaptable system, Otto, and patterns of calibrated use, disuse, and misuse on Otto with respect to the stage and capability level of the automation.

The final analytical sample included 189 participants, totaling 567 observations after removing incomplete cases. The sample likely did not capture a full range of Hofstede's cultural values—Power Distance, Uncertainty Avoidance, Collectivism, Masculinity, and Long-Term Orientation— as there were observed biases across all traits. Other than the fact that the observed sample in this study was restricted to sampling from within the borders inside of the US, the low ratings in Collectivism (equivalent to high ratings for Individualism), low ratings on Masculinity, and high ratings on Propensity to Trust predominantly reflects individuals who align closer to the Dignity cultural syndrome (Leung and Cohen, 2011; Chien et al., 2016). However, the sample also tended to rate high on Long-Term Orientation and Uncertainty Avoidance, which are less characteristic of Dignity cultures. This aligns well with Leung and Cohen's original CuPS model, and thoughts from Hofstede (Hofstede, 2011), wherein variability is observed at the individual-level and contextual-level within a given culture at the population-level.

Interestingly, the sample also seems to be characterized by higher dispositional propensities to trust in technology due to observation of higher ratings on Faith in General Technology (McKnight, 2011) and Propensity to Trust in Machines (the adapted name from General Trusting Stance - General Technology; McKnight, 2011)—and, as suggested visually in the correlational matrices, these traits showed a moderately strong positive relationship (which is expected due to

the overlap of similarities between the scales). Perhaps, it is likely that the employment of the study through a web-based platform (Prolific) and the game in the description on the study posting facilitated the bias of higher ratings on both scales. However, it is unclear as to whether this speculation is the cause of this bias in the sample from the given data.

In a more general sense, the hypothetical case that this sample reflects a general population of individuals in the country, it would be interesting to see how general attitudes towards technologies trend alongside the advances in innovation. However, in the case that this sample is not representative of the general population (which I speculate is more likely the case due to the general expectancy for traits to be distributed normally across populations), this limitation of the biased sample would elicit further investigation on individuals who differ across dispositional trusting traits towards technology, people, and across cultural values—and perhaps, a more suitable method to capture this would indeed be on the global scale—for instance, one similar to the vast sample undertaken by the Moral Machine study (Awad et al., 2018).

Reliability and functionality subscales showed similar ratings across all stages. These findings indicate an overlap, or near-equivalence between the sample's attitudes on the reliability and functionality of Otto. This finding underscores a justification that in this context, these two constructs could be taken as a composite score— however, this understanding does not scale to any context or agent. In fact, a separate suggestion can be drawn from this finding, in that because Otto is a very simple embodiment of automation (he can only automate sorting, with respect to three levels of capability), there is a likely smaller range of the resolution in trust calibration which can be drawn from this atomically-decomposed singular-function agent. Regarding the dimension of automation capability, there is also a smaller range of the specificity of trust calibration, as Otto's three, distinct levels of capability (40, 60, and 80) are not indicative

of real-world systems which have broader LoA for separate functions and modes. Paired with findings from the Exploratory Analysis, it is interesting that the higher Otto's capability, the poorer the resolution of reliance calibration— which aligns with the interpretation of McDermott and Brink's (2023) point (demonstrated in Figure X) that more complex systems may have a larger margin of error allowed by human operators.

Another finding from the metric analysis was the high composite in-task trust and baseline trust scores, which align with higher ratings on propensities to trust in both people and technology, as mentioned. However, reliance tended to cluster around the midpoint. In the model, trust modestly translates to reliance—i.e., capability moves both trust and reliance up, while self-assessment of performance moves both down. This is supported robustly by the distinction between attitudes and action, and the mediation of intent in both Ajzen and Fishbein's (1987) original behavioral model, the *Theory of Reasoned Action*, as well as Lee and See's (2004) model. Further, compliance to rely on automation is also distinct from voluntary use of automation (Nof,)— and ultimately, the incidence of whether an instance is characterized by rational, planned, or even preferred decision-making is harder to disentangle in real-world systems.

More findings from the metrics are that people who rated lower baseline trust in Otto tended to increase their trust after experiencing the gameplay. In-task trust and converged reliance in Otto remained generally stable across different stages of the experiment, but people tended to rate higher trust in Otto as its capability increased from 40% to 60% to 80%, regardless of order.

A prominent suggestion from these aggregated metric findings was that the in-task trust and self-assessment of performance suggested an opposite relationship, with in-task trust increasing as

self-assessment of performance decreased. Further, the variability in the operators' self-assessment of performance showed a negative relationship with automation capability and (weaker, but still negative) converged reliance. In addition, the covariance matrices visually suggested strong positive relationships between baseline trust, in-task trust, and converged reliance, and a negative relationship between these variables and self-assessment of performance. Despite these relationships supposed from the high-level view and not confirmed via significance testing, the latter should be warranted prior to making any claims about the definite relationship between these variables. However, the speculation I propose here is that there is a cyclical model: more information is available when automation capability increases, because operators rely on the automation more, which in turn enables them to allocate reliance to the agent, and continuous positive feedback will enforce the attitudes that the agent is reliable, therefore they rely on it more, and so forth. This cycle is also a significant part of Lee and See's (2004) original model of the observational trust approach. The opposite is also true, in cases where operators are more capable than automation, they may rely on it less, and therefore less feedback is given to enforce reliance. In conjunction, in cases where automation is merely at a low capability level and therefore it is a low bar for the human to surpass it in capability— which is equivalent to the prior paradigm, but perhaps may be regarded differently in context.

The three-tiered structural equation model fits the study data well based on all model fit indices. The most significant predictors of each equation are as follows: Collectivism and Faith in General Technology are the most significant predictors of baseline trust; Baseline trust, Otto's capability, and self-assessment of performance are significant predictors of in-task trust; and In-task trust, Otto's capability, and SA significantly predict the final path. However, the model also suggested that each set of predictors only predict a little over 20% of the variability in each of

their respective outcomes (T_Comp0=.208; T_Comp=.228; converged reliance=.222). This is expected, as it suggests that there are many predictors unaccounted for which compose the remaining 80% of each layer (and these are not necessarily mutually-exclusive, as each layer may have an effect on the following). It is interesting to speculate on which variables could have been dimensionalized in addition to the already hefty load of predictors in this study set, but iterating on different contexts in the real world, or again in this same context, warrants better ideas of which predictors are more relevant to retain in future models.

Finally, the exploratory analysis quantified the percentage of time spent in calibrated use, misuse, and disuse across conditions. It is interesting to apply the technique across disciplines (%TIR is, as discussed, from diabetes medical technology), and results yield just as useful indices in a different context of use. For instance, this technique enabled findings as follows: operators calibrated their reliance on Otto to the actual capability level about half the time. Operators disused Otto about two-fifths of the time. And, operators rarely misused Otto, relying on it more than the capabilities warranted less than one-tenth of the time. Operators generally spent more time calibrating reliance on Otto's capabilities, followed by disuse, and least on misuse. As operators progressed through stages, they tended to calibrate use to Otto's capabilities better, especially at capability levels 40 and 60. Operators with higher capability levels (60 and 80) tended to spend more time in the calibrated use band, compared to when operating with Otto at 40% capability. Together, these findings indicate that not only was the original research question addressed in terms of quantifying states of calibrated use, misuse, and disuse—but it also gave insights as to how operators perceived Otto's design. The high transparency in Otto's reliability likely had an influence on operator calibration strategies, but not to the full extent which may be speculated during the design process. This suggests that high

transparency does not directly translate to calibrated use— other factors, as discussed in the background, including affective cues such as Otto's lack of actual embodiment or truly robotic, impersonal characterization may have prevented operators from relying on Otto. Further, I speculate that operator behaviors such as mere exploration of the delegation interface may have influenced the reliance strategies.

Limitations

This study was conducted in a controlled laboratory environment using a single-function, simulation-based task. While this design allowed for precise control of variables and minimized confounds, it necessarily limits ecological validity. The constrained setting does not replicate all of the contextual factors present in operational environments, such as realistic or context-specific time pressure, multitasking demands, and the consequences of real-world decision errors. Similarly, the automation capability manipulations (40%, 60%, and 80%) were deliberately chosen to be distinct and transparent for the purposes of calibration analysis; in applied contexts, capability changes are often subtler, more granular, and less explicitly communicated (Sarter, Woods, & Billings, 1997).

Furthermore, the participant pool was drawn entirely from a U.S.-based sample, which restricts the generalizability of the cultural findings. Using the CVscale (Yoo et al., 2011), Hofstede's dimensions was measured at the individual level, which captures dispositional variance but does not account for broader organizational or national cultural contexts. Broader cross-national sampling would be necessary in future research to fully validate the findings due to cultural effects.

Although these factors limit the scope of generalization, the metrics and predictor relationships identified here—continuous reliance, calibration classification, %TIR, trust, self-assessment of performance, and cultural values—are not inherently tied to this task or sample. The methodological constraints of the present study should therefore be viewed as a trade-off between experimental control and ecological fidelity, with the measures themselves remaining portable to more complex environments, which I discuss in the next section.

Future Directions

Several extensions of this work are warranted. From a research perspective, replication with more diverse and international samples would enable a fuller understanding of how cultural values like Power Distance influence calibration across contexts. In applied terms, such studies could be embedded within existing operational training environments to assess whether the relationships observed here hold under domain-specific conditions.

Future work should also examine calibration in multi-function automation scenarios, where the system can vary in capability across different sub-tasks. The influence of individual differences also presents an opportunity for targeted intervention research through different contexts. Training modules or interface features could be experimentally varied to test their effectiveness in reducing miscalibration for operators with different trait profiles. For example, operators high in Power Distance might benefit from prompts that invite verification of automation outputs, while those with low self-assessment might benefit from confidence-building manual practice embedded within automated workflows.

Finally, longitudinal field studies would be valuable for assessing how calibration metrics and predictive models evolve over time in real-world operations. This would allow practitioners to detect emerging patterns of misuse or disuse early and adjust training, interface design, or operating procedures accordingly, closing the loop between laboratory research and operational safety.

Example Operational Application: Mammography Clinical Decision Support

> The examples used in this section show contributions of this work extending beyond a controlled research setting simulating industrial operation. The calibration metrics and predictive relationships of the findings in this study—i.e., continuous reliance, calibrated/misuse/disuse classification, %TIR, trust ratings, and self-assessment—offer a portable framework for assessing and improving calibration. Some high-stakes domains such as Medicine already capture similar data in routine practice, therefore these measures might be readily deployable without requiring interface redesign or new hardware.
>
> Perhaps when paired with the automation—such as through the deployment through a pre-programmed, accessible interface—the described computational examples below could be smoothly incorporated into existing operational workflows.

Large prospective and population-based evaluations (Dembrower et al., 2023; Lång et al., 2023) showed that AI-supported protocols can maintain or improve breast cancer detection while reducing radiologist workload (Lauritzen et al., 2024), but these effects depend on the workflow role and operating point (triage, second reader, independent reader). Furthermore, controlled experiments (Dratsch et al., 2023) found automation bias in mammography reading across experience levels, resulting in radiologists sometimes following weak AI suggestions, and sometimes ignoring strong ones.

Screening programs already log the necessary signals—AI case/region scores and marks, display state, reader actions (e.g., recall or BI-RADS assignment), and timestamps—as part of integrated reading workflows (Lång et al., 2023; Lauritzen et al., 2024). These logs can already support: (a) continuous reliance—the proportion of cases (or marked regions) in which the radiologist's action aligns with AI within a defined reading segment; (b) calibration classification—per-epoch labels of calibrated use, misuse (following low-confidence marks), or disuse (ignoring high-confidence marks) by comparing behavior to confidence-tier capability ranges set by the site's operating point and validation performance; and (c) %Time-in-Range (%TIR)—the share of reading time/cases within the appropriate confidence band. This can help evaluating the screening data, by showing where miscalibration lives—by tier, reader, and context (e.g., lesion type, density, view)— rather than averaging it away.

Conclusion

This work explored the impact of dispositional traits and cultural values on trust in and reliance on automation, and quantified instances of calibrated use, misuse, and disuse. I found significant predictors of trust and converged reliance in cultural values and the extent to which operators calibrate use of an adaptable automation system, but future work must be done to holistically define cognitive modeling of trust with respect to operator dispositional factors and real-world modeling of calibrated use, misuse, and disuse. I presented both applied implications for design and training, as well as potential operational examples of the analytical approaches to framing calibrated use.

As prophesized by Bainbridge's seminal work, *The Ironies of Automation* (Bainbridge, 1983b), there exists the pitfall of system designs which increase, rather than eliminate, problems for the human operator (Bainbridge, 1983b). Operationally, this aligns with the onus on designers and implementors of extant and novel technologies to design automation technologies which foster Lee and See's cornerstone pillars of calibrated trust and appropriate reliance (J. D. Lee & See, 2004b).

Appendix

Survey Instruments

Figure 37: CVSCALE (Yoo et al., 2011)

Dimension	Internal Item Code	Item Wording	Response Scale
Power Distance	PO1	People in higher positions should not ask the opinions of people in lower positions too frequently.	1 = Strongly Disagree … 5 = Strongly Agree
	PO2	People in higher positions should avoid social interaction with people in lower positions.	
	PO3	People in lower positions should not disagree with decisions by people in higher positions.	
Uncertainty Avoidance	UA1	It is important to closely follow instructions and procedures.	
	UA2	Rules and regulations are important because they inform me of what is expected of me.	
	UA3	Instructions for operations are important.	

Collectivism	CO1	Individuals should sacrifice self-interest for the group.	
	CO2	Group welfare is more important than individual rewards.	
	CO3	Group success is more important than individual success.	
Masculinity	MA1	It is more important for men to have a professional career than it is for women.	
	MA2	Men usually solve problems with logical analysis; women usually solve problems with intuition.	
	MA3	Solving difficult problems usually requires an active, forcible approach, which is typical of men.	
Long-Term Orientation	LTO1	How important is the value long-term planning to you?	1 = Very Unimportant … 5 = Very Important
	LTO2	How important is the value giving up today's fun for success in the future to you?	
	LTO3	How important is the value working hard for success in the future to you?	

Figure 38: Trust in Otto (TIO) adapted from Trusting Belief – Specific Technology (McKnight et al., 2011)

Dimension	Item Code	Item Wording	Response Scale
Reliability	TIO1	Otto is a very reliable piece of software.	Integer [0-100]
	TIO2	Otto does not fail me.	
	TIO3	Otto is extremely dependable.	
	TIO4	Otto does not malfunction for me.	
Functionality	TIO5	Otto has the functionality I need.	
	TIO6	Otto has the features required for my tasks.	
	TIO7	Otto has the ability to do what I want it to do.	

Figure 39: Faith in General Technology (McKnight, 2011)

Item Code	Item Wording	Response Scale
FIGT1	I believe that most technologies are effective at what they are designed to do.	1 = Strongly Disagree … 5 = Strongly Agree

FIGT2	A large majority of technologies are excellent.	
FIGT3	Most technologies have the features needed for their domain.	
FIGT4	I think most technologies enable me to do what I need to do.	

Figure 40: Trusting Stance – General Technology (McKnight, 2011)

Item Code	Item Wording	Response Scale
GPTM1	My typical approach is to trust new technologies until they prove to me that I shouldn't trust them.	1 = Strongly Disagree … 5 = Strongly Agree
GPTM2	I usually trust a technology until it gives me a reason not to trust it.	
GPTM3	I generally give a technology the benefit of the doubt when I first use it.	

Figure 41: Propensity to Trust (Frazier et al., 2013)

Item Code	Item Wording	Response Scale
PTT1	I usually trust people until they give me a reason not to trust them.	1 = Strongly Disagree … 5 = Strongly Agree
PTT2	Trusting another person is not difficult for me.	
PTT3	My typical approach is to trust new acquaintances until they prove I should not trust them.	
PTT4	My tendency to trust others is high.	

Figure 42: In-Task Measures

Dimension	Item Code	Item Wording			Response Scale
Assessment of Otto's Performance (# in-time)	Q8	Considering Otto's performance in the last round of the game:	If Otto sorted 100 shapes, how many would Otto sort in time?	I think Otto would sort __ shapes in time out of 100 shapes.	Integer [0-100]
Assessment of Otto's Performance (# correct)	Q9		If Otto sorted 100 shapes, how many would Otto sort correctly?	I think Otto would sort __ shapes correctly out of 100 shapes.	
Self-Assessment of Performance (# in-time)	Q10	Considering your performance in the last round of the game:	If I sorted 100 shapes, how many would I sort in time?	I think Otto would sort __ shapes correctly out of 100 shapes.	

Self-Assessment of Performance (# correct)	SA		If I sorted 100 shapes, how many would I sort correctly?	I think I would sort __ shapes in time out of 100 shapes.	

Figure 43: Demographic & Background Items

Item Type	Item Wording	Response Options
Demographics	What is your gender?	Male; Female; Other (Please Specify: [text entry])
	What is your age?	18–24; 25–34; 35–44; 45–54; 55–64; 65+
	What is your annual income?	Less than $5,000; $5,001–$25,000; $25,001–$50,000; $50,001–$75,000; $75,001–$100,000; More than $100,000
	What is your highest level of education?	Attended High School; High School Diploma; Attended College; Bachelor's Degree; Graduate Degree
	What are your political views?	Integer Scale [0-100] with Intervals: Very liberal; Liberal; Moderate; Conservative; Very conservative

	How religious are you?	Integer Scale [0-100] with Intervals: Not religious at all; Slightly religious; Moderately religious; Very religious
	What is your native language?	[text entry]
	What is your current city of residence?	City, ST, Country
	Where did you grow up?	City, ST, Country
Handedness	Which is your dominant hand while writing?	Right; Left; Ambidextrous
Gaming Proficiency Questionnaire	In the past month, approximately how many hours per week do you play video games?	Integer [0+]
	Please indicate what types of games you play most often. Select all that apply.	Action; Action-adventure; Adventure; Puzzle; Role-playing; Simulation; Strategy; Sports

SEM Model Outcomes

Figure 44: Lavaan Summary

lavaan 0.6-19 ended normally after 2 iterations Estimator ML

Optimization method	NLMINB
Number of model parameters	28
Number of observations	567
Number of missing patterns	1

Model Test User Model:

	Standard	Scaled
Test Statistic	16.548	16.187
Degrees of freedom	11	11
P-value (Chi-square)	0.122	0.134
Scaling correction factor		1.022
Yuan-Bentler correction (Mplus variant)		

Model Test Baseline Model:

Test statistic	437.795	367.715
Degrees of freedom	33	33
P-value	0.000	0.000
Scaling correction factor		1.191

User Model versus Baseline Model:

Comparative Fit Index (CFI)	0.986	0.985
Tucker-Lewis Index (TLI)	0.959	0.954
Robust Comparative Fit Index (CFI)		0.987
Robust Tucker-Lewis Index (TLI)		0.961

Loglikelihood and Information Criteria:

Loglikelihood user model (H0)	-7444.214	-7444.214
Scaling correction factor for the MLR correction		1.300
Loglikelihood unrestricted model (H1)	-7435.940	-7435.940
Scaling correction factor for the MLR correction		1.222
Akaike (AIC)	14944.427	14944.427
Bayesian (BIC)	15065.958	15065.958
Sample-size adjusted Bayesian (SABIC)	14977.071	14977.071

Root Mean Square Error of Approximation:

RMSEA	0.030	0.029
90 Percent confidence interval - lower	0.000	0.000

90 Percent confidence interval - upper	0.058	0.056
P-value H_0: RMSEA <= 0.050	0.872	0.886
P-value H_0: RMSEA >= 0.080	0.001	0.000
Robust RMSEA	0.029	
90 Percent confidence interval - lower	0.000	
90 Percent confidence interval - upper	0.057	
P-value H_0: Robust RMSEA <= 0.050	0.877	
P-value H_0: Robust RMSEA >= 0.080	0.001	

Standardized Root Mean Square Residual:

SRMR	0.017	0.017

Parameter Estimates:

Standard errors	Sandwich
Information bread	Observed
Observed information based on	Hessian

Regressions:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
T_Comp0 ~						

PD	(b_PD)	-1.913	1.393	-1.373	0.170	-1.913	-0.055
UA	(b_UA)	5.904	1.970	2.996	0.003	5.904	0.135
CO	(b_CO)	6.144	1.026	5.990	0.000	6.144	0.231
MA	(b_MA)	0.621	1.140	0.545	0.586	0.621	0.027
LTO	(b_LT)	-1.929	1.498	-1.287	0.198	-1.929	-0.050
PTT	(b_PT)	0.929	0.855	1.087	0.277	0.929	0.043
FIGT	(b_FI)	5.582	1.430	3.905	0.000	5.582	0.185
GPTM	(b_GP)	2.972	1.224	2.428	0.015	2.972	0.136
T_Comp ~							
T_Comp0	(a_T0)	0.199	0.044	4.513	0.000	0.199	0.217
ottocap	(a_OC)	0.405	0.053	7.658	0.000	0.405	0.304
SA	(a_SA)	-0.356	0.086	-4.140	0.000	-0.356	-0.272
converged_reliance ~							
T_Comp	(c_TC)	0.139	0.062	2.245	0.025	0.139	0.151
ottocap	(c_OC)	0.357	0.055	6.436	0.000	0.357	0.291
SA	(c_SA)	-0.227	0.057	-3.978	0.000	-0.227	-0.188
PD	(c_PD)	2.555	1.129	2.263	0.024	2.555	0.088
UA	(c_UA)	-2.436	1.409	-1.729	0.084	-2.436	-0.066
CO	(c_CO)	0.187	0.864	0.217	0.828	0.187	0.008
MA	(c_MA)	-2.038	0.889	-2.291	0.022	-2.038	-0.107
LTO	(c_LT)	-2.042	1.395	-1.464	0.143	-2.042	-0.063
PTT	(c_PT)	-0.249	0.837	-0.297	0.766	-0.249	-0.014
FIGT	(c_FI)	0.276	1.270	0.218	0.828	0.276	0.011

GPTM (c_GP) 0.591 0.870 0.679 0.497 0.591 0.032

Intercepts:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.T_Comp0	6.498	9.669	0.672	0.502	6.498	0.274
.T_Comp	47.756	5.205	9.175	0.000	47.756	2.195
.converged_rlnc	38.780	9.038	4.291	0.000	38.780	1.936

Variances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.T_Comp0	446.181	31.865	14.002	0.000	446.181	0.792
.T_Comp	365.176	35.240	10.363	0.000	365.176	0.772
.converged_rlnc	312.159	25.633	12.178	0.000	312.159	0.778

R-Square:

	Estimate
T_Comp0	0.208
T_Comp	0.228
converged_rlnc	0.222

Defined Parameters:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
ind_PD	-0.053	0.049	-1.075	0.283	-0.053	-0.002

ind_UA	0.163	0.108	1.512	0.131	0.163	0.004
ind_CO	0.170	0.097	1.746	0.081	0.170	0.008
ind_MA	0.017	0.033	0.522	0.602	0.017	0.001
ind_LTO	-0.053	0.051	-1.049	0.294	-0.053	-0.002
ind_PTT	0.026	0.026	0.973	0.331	0.026	0.001
ind_FIGT	0.155	0.092	1.679	0.093	0.155	0.006
ind_GPTM	0.082	0.058	1.411	0.158	0.082	0.004
ind_ottocap	0.056	0.024	2.310	0.021	0.056	0.046
ind_SA	-0.050	0.026	-1.944	0.052	-0.050	-0.041
tot_PD	2.502	1.145	2.186	0.029	2.502	0.086
tot_UA	-2.272	1.401	-1.622	0.105	-2.272	-0.062
tot_CO	0.357	0.871	0.410	0.682	0.357	0.016
tot_MA	-2.021	0.900	-2.246	0.025	-2.021	-0.106
tot_LTO	-2.095	1.400	-1.497	0.134	-2.095	-0.065
tot_PTT	-0.223	0.839	-0.266	0.790	-0.223	-0.012
tot_FIGT	0.431	1.273	0.338	0.735	0.431	0.017
tot_GPTM	0.673	0.869	0.774	0.439	0.673	0.036
tot_ottocap	0.413	0.049	8.494	0.000	0.413	0.337
tot_SA	-0.277	0.063	-4.399	0.000	-0.277	-0.230

Figure 45: Lavaan Bootstrap

lavaan 0.6-19 ended normally after 2 iterations

Estimator	ML
Optimization method	NLMINB
Number of model parameters	28
Number of observations	567
Number of missing patterns	1

Model Test User Model:

Test statistic	16.548
Degrees of freedom	11
P-value (Chi-square)	0.122

Parameter Estimates:

Standard errors	Bootstrap
Number of requested bootstrap draws	1000
Number of successful bootstrap draws	1000

Regressions:

	Estimate	Std.Err	z-value	P(>|z|)	ci.lower	ci.upper
T_Comp0 ~						
PD (b_PD)	-1.913	1.460	-1.310	0.190	-4.831	1.149

UA	(b_UA)	5.904	1.972	2.994	0.003	2.197	9.666
CO	(b_CO)	6.144	1.046	5.876	0.000	4.011	8.209
MA	(b_MA)	0.621	1.195	0.520	0.603	-1.521	3.063
LTO	(b_LT)	-1.929	1.515	-1.273	0.203	-4.801	1.186
PTT	(b_PT)	0.929	0.897	1.036	0.300	-0.757	2.782
FIGT	(b_FI)	5.582	1.467	3.806	0.000	2.805	8.507
GPTM	(b_GP)	2.972	1.234	2.408	0.016	0.503	5.268
T_Comp ~							
T_Comp0	(a_T0)	0.199	0.046	4.327	0.000	0.108	0.285
ottocap	(a_OC)	0.405	0.054	7.521	0.000	0.297	0.506
SA	(a_SA)	-0.356	0.090	-3.962	0.000	-0.541	-0.192
converged_reliance ~							
T_Comp	(c_TC)	0.139	0.061	2.288	0.022	0.017	0.258
ottocap	(c_OC)	0.357	0.056	6.402	0.000	0.239	0.461
SA	(c_SA)	-0.227	0.056	-4.037	0.000	-0.337	-0.113
PD	(c_PD)	2.555	1.128	2.266	0.023	0.453	4.814
UA	(c_UA)	-2.436	1.466	-1.661	0.097	-5.242	0.538
CO	(c_CO)	0.187	0.865	0.217	0.829	-1.457	1.953
MA	(c_MA)	-2.038	0.911	-2.237	0.025	-3.907	-0.316
LTO	(c_LT)	-2.042	1.409	-1.449	0.147	-4.721	0.712
PTT	(c_PT)	-0.249	0.836	-0.298	0.766	-1.892	1.384
FIGT	(c_FI)	0.276	1.288	0.214	0.830	-2.326	2.787
GPTM	(c_GP)	0.591	0.876	0.675	0.500	-1.039	2.282

Std.lv	Std.all
-1.913	-0.055
5.904	0.135
6.144	0.231
0.621	0.027
-1.929	-0.050
0.929	0.043
5.582	0.185
2.972	0.136
0.199	0.217
0.405	0.304
-0.356	-0.272
0.139	0.151
0.357	0.291
-0.227	-0.188
2.555	0.088
-2.436	-0.066
0.187	0.008
-2.038	-0.107
-2.042	-0.063

-0.249 -0.014

0.276 0.011

0.591 0.032

Intercepts:

	Estimate	Std.Err	z-value	P(>|z|)	ci.lower	ci.upper
.T_Comp0	6.498	9.641	0.674	0.500	-12.966	24.311
.T_Comp	47.756	5.509	8.669	0.000	36.985	58.384
.converged_rlnc	38.780	9.262	4.187	0.000	21.121	57.294

Std.lv	Std.all
6.498	0.274
47.756	2.195
38.780	1.936

Variances:

	Estimate	Std.Err	z-value	P(>|z|)	ci.lower	ci.upper
.T_Comp0	446.181	31.575	14.131	0.000	379.844	501.270
.T_Comp	365.176	35.275	10.352	0.000	297.011	435.085
.converged_rlnc	312.159	26.341	11.851	0.000	250.051	359.057

Std.lv	Std.all
446.181	0.792
365.176	0.772
312.159	0.778

R-Square:

	Estimate
T_Comp0	0.208
T_Comp	0.228
converged_rlnc	0.222

Defined Parameters:

	Estimate	Std.Err	z-value	P(>|z|)	ci.lower	ci.upper
ind_PD	-0.053	0.055	-0.967	0.334	-0.191	0.033
ind_UA	0.163	0.115	1.422	0.155	0.008	0.427
ind_CO	0.170	0.100	1.694	0.090	0.016	0.408
ind_MA	0.017	0.038	0.450	0.653	-0.048	0.107
ind_LTO	-0.053	0.056	-0.961	0.336	-0.189	0.033
ind_PTT	0.026	0.031	0.839	0.401	-0.027	0.101
ind_FIGT	0.155	0.097	1.594	0.111	0.013	0.394
ind_GPTM	0.082	0.064	1.295	0.195	0.002	0.245
ind_ottocap	0.056	0.025	2.290	0.022	0.007	0.108
ind_SA	-0.050	0.027	-1.874	0.061	-0.107	-0.005
tot_PD	2.502	1.140	2.195	0.028	0.352	4.756
tot_UA	-2.272	1.459	-1.557	0.119	-5.051	0.608
tot_CO	0.357	0.875	0.409	0.683	-1.320	2.150
tot_MA	-2.021	0.921	-2.194	0.028	-3.910	-0.237

tot_LTO	-2.095	1.414	-1.482	0.138	-4.743	0.648	-0.053	-0.002
tot_PTT	-0.223	0.840	-0.266	0.790	-1.919	1.415	0.163	0.004
tot_FIGT	0.431	1.290	0.334	0.738	-2.212	2.968	0.170	0.008
tot_GPTM	0.673	0.876	0.769	0.442	-0.932	2.370	0.017	0.001
tot_ottocap	0.413	0.049	8.504	0.000	0.313	0.503	-0.053	-0.002
tot_SA	-0.277	0.063	-4.364	0.000	-0.400	-0.154	0.026	0.001
							0.155	0.006
							0.082	0.004
							0.056	0.046
							-0.050	-0.041
							2.502	0.086
							-2.272	-0.062
							0.357	0.016
							-2.021	-0.106
							-2.095	-0.065
							-0.223	-0.012

0.431	0.017
0.673	0.036
0.413	0.337
-0.277	-0.230

References

Adler, P. S. (2001). Market, Hierarchy, and Trust: The Knowledge Economy and the Future of Capitalism. *Organization Science*, *12*(2), 215–234. https://doi.org/10.1287/orsc.12.2.215.10117

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Prentice Hall, Inc.

Alarcon, G. M., Gibson, A. M., Jessup, S. A., & Capiola, A. (2021). Exploring the differential effects of trust violations in human-human and human-robot interactions. *Applied Ergonomics*, *93*, 103350. https://doi.org/10.1016/j.apergo.2020.103350

Alves, J., Lima, T. M., & Gaspar, P. D. (2023). Is Industry 5.0 a Human-Centred Approach? A Systematic Review. *Processes*, *11*(1), Article 1. https://doi.org/10.3390/pr11010193

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J.-F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, *117*(5), 2332–2337. https://doi.org/10.1073/pnas.1911517117

Bainbridge, L. (1983a). Ironies of automation. *Automatica*, *19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Bernabei, M., & Costantino, F. (2024). Adaptive automation: Status of research and future challenges. *Robotics and Computer-Integrated Manufacturing*, *88*, 102724. https://doi.org/10.1016/j.rcim.2024.102724

Burge, D. S. (n.d.). *The Systems Engineering Tool Box*.

Capiola, A., Hamdan, I. aldin, Lyons, J. B., Lewis, M., Alarcon, G. M., & Sycara, K. (2024). The Effect of Asset Degradation on Trust in Swarms: A Reexamination of System-Wide Trust in Human-Swarm Interaction. *Human Factors*, *66*(5), 1475–1489. https://doi.org/10.1177/00187208221145261

Casner, S. M., & Schooler, J. W. (2014). Thoughts in Flight: Automation Use and Pilots' Task-Related and Task-Unrelated Thought. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*(3), 433–442. https://doi.org/10.1177/0018720813501550

Chien, S.-Y., Lewis, M., Semnani-Azad, Z., & Sycara, K. (2014). An Empirical Model of Cultural Factors on Trust in Automation. *Proceedings of the Human Factors and*

Ergonomics Society Annual Meeting, *58*(1), 859–863. https://doi.org/10.1177/1541931214581181

Chien, S.-Y., Lewis, M., Sycara, K., Jyi-Shane Liu, & Kumru, A. (2016). Influence of cultural factors in dynamic trust in automation. *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 002884–002889. https://doi.org/10.1109/smc.2016.7844677

Chien, S.-Y., Lewis, M., Sycara, K., Kumru, A., & Liu, J.-S. (2020a). Influence of Culture, Transparency, Trust, and Degree of Automation on Automation Use. *IEEE Transactions on Human-Machine Systems*, *50*(3), 205–214. IEEE Transactions on Human-Machine Systems. https://doi.org/10.1109/THMS.2019.2931755

Chien, S.-Y., Lewis, M., Sycara, K., Kumru, A., & Liu, J.-S. (2020b). Influence of Culture, Transparency, Trust, and Degree of Automation on Automation Use. *IEEE Transactions on Human-Machine Systems*, *50*(3), 205–214. https://doi.org/10.1109/THMS.2019.2931755

Chien, S.-Y., Lewis, M., Sycara, K., Liu, J.-S., & Kumru, A. (2018). The Effect of Culture on Trust in Automation: Reliability and Workload. *ACM Transactions on Interactive Intelligent Systems*, *8*(4), 29:1-29:31. https://doi.org/10.1145/3230736

Chien, S.-Y., Sycara, K., Liu, J.-S., & Kumru, A. (2016). Relation between Trust Attitudes Toward Automation, Hofstede's Cultural Dimensions, and Big Five Personality Traits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 841–845. https://doi.org/10.1177/1541931213601192

Chiou, E. K., & Lee, J. D. (2023). Trusting Automation: Designing for Responsivity and Resilience. *Human Factors*, *65*(1), 137–165. https://doi.org/10.1177/00187208211009995

Cohen, J. (2009). *Statistical power analysis for the behavioral sciences* (2. ed., reprint). Psychology Press.

Dara, R., Hazrati Fard, S. M., & Kaur, J. (2022). Recommendations for ethical and responsible use of artificial intelligence in digital agriculture. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.884192

Davis, S. E. (2019). *Individual Differences in Operators' Trust in Autonomous Systems: A Review of the Literature*.

De Oliveira, E., Reynaud, E., & Osiurak, F. (2019). Roles of Technical Reasoning, Theory of Mind, Creativity, and Fluid Cognition in Cumulative Technological Culture. *Human Nature*, *30*(3), 326–340. https://doi.org/10.1007/s12110-019-09349-1

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience

in cognitive agents. *Journal of Experimental Psychology. Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

Dembrower, K., Crippa, A., Colón, E., Eklund, M., & Strand, F. (2023). Artificial intelligence for breast cancer detection in screening mammography in Sweden: A prospective, population-based, paired-reader, non-inferiority study. *The Lancet Digital Health*, *5*(10), e703–e711. https://doi.org/10.1016/S2589-7500(23)00153-X

Dergaa, I., Ben Saad, H., Glenn, J. M., Amamou, B., Ben Aissa, M., Guelmami, N., Fekih-Romdhane, F., & Chamari, K. (2024). From tools to threats: A reflection on the impact of artificial-intelligence chatbots on cognitive health. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1259845

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour & Information Technology*, *18*(6), 399–411. https://doi.org/10.1080/014492999118832

Dijkstra, J. J., Liebrand, W. B. G., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology*, *17*(3), 155–163. https://doi.org/10.1080/014492998119526

Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., & Pinto Dos Santos, D. (2023). Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology*, *307*(4), e222176. https://doi.org/10.1148/radiol.222176

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/s1071-5819(03)00038-7

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, 789–795 vol.3. https://doi.org/10.1109/NAECON.1988.195097

Endsley, M. R. (2017). Autonomous Driving Systems: A Preliminary Naturalistic Study of the Tesla Model S. *Journal of Cognitive Engineering and Decision Making*, *11*(3), 225–238. https://doi.org/10.1177/1555343417695197

Endsley, M. R. (2018). Level of Automation Forms a Key Aspect of Autonomy Design. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 29–34. https://doi.org/10.1177/1555343417723432

Federal Aviation Administration. (2017). *Instrument procedures handbook (FAA-H-8083-16B)*. U.S. Department of Transportation.

Federal Aviation Administration. (2024). *LNAV/VNAV approaches*. FAA GNSS—Satellite Navigation. https://www.faa.gov/about/office_org/headquarters_offices/ato/service_units/techops/navservices/gnss/nas/procedures/vnav?utm_source=chatgpt.com

Frazier, M. L., Johnson, P. D., & Fainshmidt, S. (2013). Development and validation of a propensity to trust scale. *Journal of Trust Research*, *3*(2), 76–97. https://doi.org/10.1080/21515581.2013.820026

Freeth, T., Bitsakis, Y., Moussas, X., Seiradakis, J. H., Tselikas, A., Mangou, H., Zafeiropoulou, M., Hadland, R., Bate, D., Ramsey, A., Allen, M., Crawley, A., Hockley, P., Malzbender, T., Gelb, D., Ambrisco, W., & Edmunds, M. G. (2006). Decoding the ancient Greek astronomical calculator known as the Antikythera Mechanism. *Nature*, *444*(7119), 587–591. https://doi.org/10.1038/nature05357

Fu, E., Sibi, S., Miller, D., Johns, M., Mok, B., Fischer, M., & Sirkin, D. (2019). The Car That Cried Wolf: Driver Responses to Missing, Perfectly Performing, and Oversensitive Collision Avoidance Systems. *2019 IEEE Intelligent Vehicles Symposium (IV)*, 1830–1836. https://doi.org/10.1109/IVS.2019.8814190

Gibson, K. R., & Ingold, T. (1993). *Tools, Language and Cognition in Human Evolution*. Cambridge University Press.

Goldenberg, G., & Spatt, J. (2009). The neural basis of tool use. *Brain*, *132*(6), 1645–1655. https://doi.org/10.1093/brain/awp080

Groumpos, P. P. (2021). A Critical Historical and Scientific Overview of all Industrial Revolutions. *IFAC-PapersOnLine*, *54*(13), 464–471. https://doi.org/10.1016/j.ifacol.2021.10.492

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Stowers, K., Brill, J. C., Billings, D. R., Schaefer, K. E., & Szalma, J. L. (2023). How and why humans trust: A meta-analysis and elaborated model. *Frontiers in Psychology*, *14*, 1081086. https://doi.org/10.3389/fpsyg.2023.1081086

Hayes, S. (2016). Industrial automation and stress, c. 1945–79. *Stress in Post-War Britain, 1945–85*, 75–94.

Hoff, K. A., & Bashir, M. (2015b). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture*, *2*(1). https://doi.org/10.9707/2307-0919.1014

Hofstede, G., & Bond, M. H. (1984). Hofstede's Culture Dimensions: An Independent Validation Using Rokeach's Value Survey. *Journal of Cross-Cultural Psychology*, *15*(4), 417–433. https://doi.org/10.1177/0022002184015004003

Huang, S., Wang, B., Li, X., Zheng, P., Mourtzis, D., & Wang, L. (2022). Industry 5.0 and Society 5.0—Comparison, complementation and co-evolution. *Journal of Manufacturing Systems*, *64*, 424–428. https://doi.org/10.1016/j.jmsy.2022.07.010

Hussein, A., Elsawah, S., & Abbass, H. A. (2020). Trust Mediating Reliability–Reliance Relationship in Supervisory Control of Human–Swarm Interactions. *Human Factors*, *62*(8), 1237–1248. https://doi.org/10.1177/0018720819879273

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Johnson-Frey, S. H. (2004). The neural bases of complex tool use in humans. *Trends in Cognitive Sciences*, *8*(2), 71–78. https://doi.org/10.1016/j.tics.2003.12.002

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (Fourth edition). The Guilford Press.

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021a). Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.604977

Lång, K., Josefsson, V., Larsson, A.-M., Larsson, S., Högberg, C., Sartor, H., Hofvind, S., Andersson, I., & Rosso, A. (2023). Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): A clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology*, *24*(8), 936–944. https://doi.org/10.1016/S1470-2045(23)00298-X

Lauritzen, A. D., Lillholm, M., Lynge, E., Nielsen, M., Karssemeijer, N., & Vejborg, I. (2024). Early Indicators of the Impact of Using AI in Mammography Screening for Breast Cancer. *Radiology*, *311*(3), e232479. https://doi.org/10.1148/radiol.232479

Lee, J. D., & Moray, N. (1994a). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. https://doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & Moray, N. (1994b). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. https://doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004a). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392

Leung, A. K.-Y., & Cohen, D. (2024). Within- and Between-Culture Variation: Individual Differences and the Cultural Logics of Honor, Face, and Dignity Cultures. *ResearchGate*. https://doi.org/10.1037/a0022151

Lin, R., Ma, L., & Zhang, W. (2018). An interview study exploring Tesla drivers' behavioural adaptation. *Applied Ergonomics*, *72*, 37–47. https://doi.org/10.1016/j.apergo.2018.04.006

Lucas, G. M., Becerik-Gerber, B., & Roll, S. C. (2024). Calibrating workers' trust in intelligent automated systems. *Patterns*, *5*(9), 101045. https://doi.org/10.1016/j.patter.2024.101045

Madhavan, G. (2024). *Wicked problems: How to engineer a better world.* (First). W. W. Norton & Company, Inc.

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. https://doi.org/10.1080/14639220500337708

Mangalam, M., Fragaszy, D. M., Wagman, J. B., Day, B. M., Kelty-Stephen, D. G., Bongers, R. M., Stout, D. W., & Osiurak, F. (2022). On the psychological origins of tool use. *Neuroscience & Biobehavioral Reviews*, *134*, 104521. https://doi.org/10.1016/j.neubiorev.2022.104521

Marocco, S., Talamo, A., & Quintiliani, F. (2024). From service design thinking to the third generation of activity theory: A new model for designing AI-based decision-support systems. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1303691

Mathur, A., Dabas, A., & Sharma, N. (2022). Evolution From Industry 1.0 to Industry 5.0. *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 1390–1394. https://doi.org/10.1109/ICAC3N56670.2022.10074274

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

McDermott, P. L., & Brink, R. N. ten. (2019). Practical Guidance for Evaluating Calibrated Trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 362–366. https://doi.org/10.1177/1071181319631379

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Trans. Manage. Inf. Syst.*, *2*(2), 12:1-12:25. https://doi.org/10.1145/1985347.1985353

McWilliams, T., & Ward, N. (2021). Underload on the Road: Measuring Vigilance Decrements During Partially Automated Driving. *Frontiers in Psychology*, *12*, 631364. https://doi.org/10.3389/fpsyg.2021.631364

Merritt, S. M. (2011). Affective Processes in Human–Automation Interactions. *Human Factors*, *53*(4), 356–370. https://doi.org/10.1177/0018720811411912

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

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

Meyer, J., & Lee, J. D. (2013). 109 Trust, Reliance, and Compliance. In J. D. Lee & A. Kirlik (Eds.), *The Oxford Handbook of Cognitive Engineering* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199757183.013.0007

Miller, C. A., & Parasuraman, R. (2007). Designing for Flexible Interaction Between Humans and Automation: Delegation Interfaces for Supervisory Control. *Human Factors*, *49*(1), 57–75. https://doi.org/10.1518/001872007779598037

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., & Ju, W. (2016). Behavioral Measurement of Trust in Automation The Trust Fall. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 1849–1853. https://doi.org/10.1177/1541931213601422

Morando, A., Gershon, P., Mehler, B., & Reimer, B. (2021). Visual attention and steering wheel control: From engagement to disengagement of Tesla Autopilot. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 1390–1394. https://doi.org/10.1177/1071181321651118

Mueller, A. S., Cicchino, J. B., & Calvanelli, J. V. (2024). Habits, attitudes, and expectations of regular users of partial driving automation systems. *Journal of Safety Research*, *88*, 125–134. https://doi.org/10.1016/j.jsr.2023.10.015

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5), 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

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. https://doi.org/10.1080/00140139608964474

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nof, S. Y. (Ed.). (2023). *Springer Handbook of Automation* (2nd ed. 2023). Springer International Publishing. https://doi.org/10.1007/978-3-030-96729-1

Nordhoff, S., Lee, J. D., Calvert, S. C., Berge, S., Hagenzieker, M., & Happee, R. (2023). (Mis-)use of standard Autopilot and Full Self-Driving (FSD) Beta: Results from interviews with users of Tesla's FSD Beta. *Frontiers in Psychology*, *14*, 1101520. https://doi.org/10.3389/fpsyg.2023.1101520

Osiurak, F., Navarro, J., & Reynaud, E. (2018). How Our Cognition Shapes and Is Shaped by Technology: A Common Framework for Understanding Human Tool-Use Interactions in the Past, Present, and Future. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00293

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Pinker, S. (n.d.). *The cognitive niche: Coevolution of intelligence, sociality, and language*. https://doi.org/10.1073/pnas.0914630107

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual Differences in the Calibration of Trust in Automation. *Human Factors*, *57*(4), 545–556. https://doi.org/10.1177/0018720814564422

Puttero, S., Verna, E., Genta, G., & Galetto, M. (2025). Collaborative robots for quality control: An overview of recent studies and emerging trends. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-025-02600-w

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*(3), 257–266. https://doi.org/10.1109/TSMC.1983.6313160

Razin, Y. S., & Feigh, K. M. (2024). Converging Measures and an Emergent Model: A Meta-Analysis of Human-Machine Trust Questionnaires. *J. Hum.-Robot Interact.*, *13*(4), 58:1-58:41. https://doi.org/10.1145/3677614

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places* (pp. xiv, 305). Cambridge University Press.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95–112. https://doi.org/10.1037/0022-3514.49.1.95

Sarter, N. B., & Woods, D. D. (1995). How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 5–19. https://doi.org/10.1518/001872095779049516

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 1926–1943). Wiley.

Schaefer, K. E. (2016). Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI." In R. Mittu, D. Sofge, A. Wagner, & W. F. Lawless (Eds.), *Robust Intelligence and Trust in Autonomous Systems* (pp. 191–218). Springer US. https://doi.org/10.1007/978-1-4899-7668-0_10

Sharma, A., Singh, B. J., & Professor, Department of Mechanical Engineering, MMDU Mullana, Haryana, India. (2020). Evolution of Industrial Revolutions: A Review. *International Journal of Innovative Technology and Exploring Engineering*, *9*(11), 66–73. https://doi.org/10.35940/ijitee.i7144.0991120

Sheridan, T., Verplank, W., & Brooks, T. (1978). *Human and Computer Control of Undersea Teleoperators*.

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction*, *12*(3), 109–124. https://doi.org/10.17705/1thci.00131

Soressi, M., Dibble, H. L., & University of Pennsylvania (Eds.). (2003). *Multiple approaches to the study of bifacial technologies* (1st ed). University of Pennsylvania, Museum of Archaeology and Anthropology.

Stotz, K. (2010). Human nature and cognitive–developmental niche construction. *Phenomenology and the Cognitive Sciences*, *9*(4), 483–501. https://doi.org/10.1007/s11097-010-9178-7

Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology. https://doi.org/10.6028/nist.ai.100-1

Taj, I., & Zaman, N. (2022). Towards Industrial Revolution 5.0 and Explainable Artificial Intelligence: Challenges and Opportunities. *International Journal of Computing and Digital Systems*, *12*(1), 285–310. https://doi.org/10.12785/ijcds/120124

Takayama, L., Groom, V., & Nass, C. (2009). I'm sorry, Dave: I'm afraid i won't do that: social aspects of human-agent conflict. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2099–2108. https://doi.org/10.1145/1518701.1519021

Triandis, H. C. (1996). The psychological measurement of cultural syndromes. *American Psychologist*, *51*(4), 407–415. https://doi.org/10.1037/0003-066X.51.4.407

Triberti, S., Di Fuccio, R., Scuotto, C., Marsico, E., & Limone, P. (2024). "Better than my professor?" How to develop artificial intelligence tools for higher education. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1329605

van de Merwe, K., Mallam, S., & Nazir, S. (2024). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. *Human Factors*, *66*(1), 180–208. https://doi.org/10.1177/00187208221077804

Villani, V., Pini, F., Leali, F., & Secchi, C. (2018). Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics*, *55*, 248–266. https://doi.org/10.1016/j.mechatronics.2018.02.009

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *The Journal of Applied Psychology*, *82*(2), 247–252. https://doi.org/10.1037/0021-9010.82.2.247

Wildman, J. L., Nguyen, D., Thayer, A. L., Robbins-Roth, V. T., Carroll, M., Carmody, K., Ficke, C., Akib, M., & Addis, A. (2024). Trust in Human-Agent Teams: A Multilevel Perspective and Future Research Agenda. *Organizational Psychology Review*, *14*(3), 373–402. https://doi.org/10.1177/20413866241253278

Wright, E. E., Morgan, K., Fu, D. K., Wilkins, N., & Guffey, W. J. (2020). Time in Range: How to Measure It, How to Report It, and Its Practical Application in Clinical Decision-Making. *Clinical Diabetes : A Publication of the American Diabetes Association*, *38*(5), 439–448. https://doi.org/10.2337/cd20-0042