In the Rough: Evaluation of Convergence Across Trust Assessment Techniques Using an Autonomous Golf Cart

Nathan L. Tenhundfeld, The University of Alabama in Huntsville, USA, **Jason Forsyth, Nathan R. Sprague** and **Samy El-Tawab**, James Madison University, USA, **Jenna E. Cotter,** and **Lisa Vangsness**, The University of Alabama in Huntsville, USA

As automated and autonomous systems become more widely available, the ability to integrate them into environments seamlessly becomes more important. One cognitive construct that can predict the use, misuse, and disuse of automated and autonomous systems is trust that a user has in the system. The literature has explored not only the predictive nature of trust but also the ways in which it can be evaluated. As a result, various measures, such as physiological and behavioral measures, have been proposed as ways to evaluate trust in real-time. However, inherent differences in the measurement approaches (e.g., task dependencies and timescales) raise questions about whether the use of these approaches will converge upon each other. If they do, then the selection of any given proven approach to trust assessment may not matter. However, if they do not converge, it raises questions about the ability of these measures to assess trust equally and whether discrepancies are attributable to discriminant validity or other factors. The present study used various trust assessment techniques for passengers in a self-driving golf-cart. We find little to no convergence across measures, raising questions that need to be addressed in future research.

Keywords: Trust, self-driving vehicles, trust in automation

Remarkable advances have been made in self-driving vehicle technology in recent years. While these vehicles fall short of what names like "full self-driving" insinuate (i.e., they are not autonomous), their high level of automation makes it such that the human in the "driver's seat" can be largely uninvolved. As these vehicles are used in increasingly complex environments, safe and appropriate use must be prioritized. One of the greatest predictors of the safe and appropriate use of a system is an understanding of the operator's trust in that system (de Visser et al., 2020; Dzindolet et al., 2003, p. 200; Hancock et al., 2020; Wiczorek & Meyer, 2019). Trust is a complex, dynamic attitude that consists of initial learned, situational, and dispositional trust (Hoff & Bashir, 2015; Lee & See, 2004). It is not simply that greater levels of trust result in safer use of a system, but rather that appropriately calibrated trust is needed to ensure that users are not using the systems in a way that is unsafe.

Users who place a level of trust in the system which is warranted by the system's actual capabilities (i.e., trustworthiness) are said to be calibrated in their trust (Estepp et al., 2018; Kraus et al., 2019). Miscalibrated users, on the other hand, can either overtrust or undertrust a system. Overtrust exists when one's trust in the system exceeds what is warranted by the system's actual capabilities (M. S. Cohen et al., 1997; Robinette et al., 2016). This overtrust can be pernicious in that users can become complacent and fail to provide appropriate supervisory control over the system (Merritt et al., 2019; Wickens et al., 2015), which can lead to degraded situation awareness, failure to detect system errors, and worse overall human–machine team performance (Manzey et al., 2012; Sebok & Wickens, 2017). Unsettling real-world examples of overtrust can be seen in images of drivers asleep behind the wheel of their vehicle while in self-driving mode.

Address correspondence to Nathan L. Tenhundfeld, Department of Psychology, University of Alabama in Huntsville, 301 Sparkman Dr NW, Huntsville, AL 35899, USA. Email: NLT0006@uah.edu

Journal of Cognitive Engineering and Decision Making
Vol. 18, No. 1, March 2024, pp. 3–21
DOI:10.1177/15553434231206422
Article reuse guidelines: sagepub.com/journals-permissions

Undertrust refers to instances wherein a user's trust in a system is less than what the system's capabilities warrant (de Visser et al., 2020; Kohn et al., 2018). Undertrust can result in failure to accept and adopt technology (Ghazizadeh et al., 2012; Matsuyama et al., 2021; Zhang et al., 2019) or disuse of a system after it has already been adopted (Dzindolet et al., 2003; Parasuraman & Riley, 1997). Distrust is similar to undertrust; however, the principal difference is that distrust can be warranted (Mirnig et al., 2016). In fact, research has suggested that trust and distrust represent two distinct constructs (Kramer, 1999; Lewicki et al., 1998; Tenhundfeld et al., 2019).

Under-reliance, be it caused by undertrust or distrust, can present an issue if it is presupposed that the system would otherwise provide a performance or safety benefit to the user. As such, product designers and developers may have an interest in working to repair trust through approaches like having the system apologize (Fratczak et al., 2021; Kim & Song, 2021), increasing its transparency (Hussein et al., 2020; Yang et al., 2017), or even reassuring the user of its (the system's) proficiency (Israelsen & Ahmed, 2019). However, these repair approaches can only be implemented if distrust or undertrust can be detected. What is more, the efficacy of these approaches may be tied to the actual timing of the repair strategy (Robinette et al., 2015). It, therefore, seems obvious that to promote highly efficacious human-automation interaction, systems need the ability to detect, in real-time, changes in a user's trust.

Humans are inherently good at inferring intention, state of mind, and emotions from both verbal and non-verbal social cues. While a system could routinely ask its user to evaluate his/her trust, there are potential methodological concerns (e.g., anchoring and response bias) in addition to concerns over whether such a redundant request may be seen as an annoying behavior which could impact the use of such a system (Segura et al., 2012). In an effort to develop a real-time assessment of trust, research has examined the utility of physiological and behavioral measures such as heart rate (Khalid et al., 2016; Mitkidis et al., 2015; Perello-March et al., 2022; Tolston et al., 2018), galvanic skin response (Akash et al., 2018; Chen et al., 2015), interventions (Tenhundfeld et al., 2019, 2020), monitoring behaviors (Bahner et al., 2008; Bailey & Scerbo, 2007; Banks et al., 2018; Endsley, 2017), and eye-tracking (Hergeth et al., 2015; Lu & Sarter, 2019).

It is worth noting that these trust assessment techniques are not directly assessing trust but rather believed correlate with trust. In the case of behavioral measures such as interventions, monitoring behaviors, and eye tracking, each is believed to be downstream of trust such that low trust may lead to more frequent or earlier interventions and monitoring/verification behaviors (Bahner et al., 2008; Tenhundfeld et al., 2019, 2020; Walker et al., 2018). The idea is that someone who trusts the vehicle less will be more likely to take over control in the face of uncertainty or will be more directly engaged in the primary task of supervisory control. Similarly, physiological monitoring of things like heart rate, heart rate variability, and galvanic skin response are all believed to be directly influenced through trust by a couple of mechanisms. The first mechanism wherein changes in trust would be expected in the physiological data is through stress. Each of these aforementioned variables has been shown to be a part of the human stress response (Arsalan & Majid, 2021; Nickel & Nachreiner, 2003; Thayer et al., 2012). The belief is that as trust decreases, the stress associated with reliance on an automated system increases, and as such, so too do the physiological indicators of stress. The second proposed mechanism would be through an individual's workload. In cases of high trust, there is more likely to be complacency and the failure to monitor the automated system (Bahner et al., 2008; Merritt et al., 2019; Parasuraman & Manzey, 2010; Sauer et al., 2016). On the other hand, in instances where there is low trust in a system, the individual will likely be more engaged in the supervisory control tasks required of him/her, resulting in higher task load and subsequent increases in workload. This increase in workload correlates with changes to physiological metrics like heart rate and heart rate variability (Eilebrecht et al., 2012; Hoover et al., 2012). However, it is important to note that there are other factors that may affect user workload, beyond trust. One's assessment of task demands may impact his/her workload. For example, one may believe that driving becomes a secondary task for which the automation is in control, and thus the "driver" can engage in his/her self-assigned primary task of engaging in conversation, reading, texting, etc. Alternatively, a driver may stay engaged in maintaining supervisory control over the automation, if he/she believes it to be the primary task, in a way that increases workload regardless of his/her trust in the system (Warm et al., 2008). This introduces noise into the potential interpretation of physiological assessments of trust if workload is the main factor that contributes to the changes in physiological response.

As detailed above, the use of non-subjective measures can be useful for the real-time and real-world assessment of trust. However, beyond the fact that much of the research highlighting the use of physiological and behavioral measures relies on theoretical assumptions as to why they are measures of trust, the reality is that there are very few studies that compare these approaches within a single study. This comparative analysis is important for several reasons.

First, physiological and behavioral measures of trust assess along different timescales than subjective assessments. Whereas subjective assessments are most frequently given following a manipulation, physiological and behavioral data are collected throughout the duration of an interaction. While the data are not always evaluated through dynamic modeling techniques (Tenhundfeld et al., 2022), even aggregated data (e.g., number of interventions and average heart rate) are still an aggregation of data collected over the course of the interaction, rather than following the interaction. This different timescale poses a potential theoretical problem as trust is not a monolithic construct.

Trust itself can be broken into categories of dispositional, situational, initial learned, and dynamic learned (Hoff & Bashir, 2015). These trust factors play a different role in reliance strategies such that dispositional, situational, and initial learned trust all affect one's initial reliance, whereas the use of a system informs one's dynamic learned trust which subsequently impacts their reliance thereafter. As such, trust assessment approaches that evaluate trust over the course of an experiment, as you see with behavioral and physiological measures, may be capturing one's shift from this pre-interaction trust to his/her during-/post-interaction trust. Because of this, complex interactions between, for example, dispositional trust and system performance may yield results that would otherwise not be captured by a post-interaction subjective questionnaire. So whereas measures could converge, there would also be discriminant validity to each measure such that they are unlikely to yield the same results because of the unique variance they are explaining (Campbell & Fiske, 1959). What is more, aggregated data may obfuscate actual shifts in user trust as the shifts in trust may occur following a series of interactions, which then get averaged with pre-shift trust (thereby dampening the effect).

For these reasons, it is imperative that comparative analyses be done on the different approaches to trust assessment. If we are to accept that the measurements used are all assessing trust, it is critical to understand how results may vary as a function of the measurement approach. Because trust is not a monolithic construct, the field must understand how different measurement approaches are mapping onto the multidimensionality of trust. The first step, however, is to see whether there are even disparate results as a function of measurement used. If there is a disagreement between the results from different measurements, this would suggest that each measure is uniquely assessing different facets of trust. It would therefore be necessary for future research to understand the nuance of how different timescales and measurement approaches map onto the multidimensionality of trust as a construct. Additionally, this would also warrant more careful consideration of the measurement approach used for any given study, as different measurement approaches may yield different results. If, on the other hand, there is complete alignment between results, regardless of the measurement technique used, this would suggest that researchers would be able to select whichever technique best fits within their paradigm as the results would be the same.

Ultimately, the understanding of how these different trust assessment techniques map onto trust is not only important for advancing theory but also for the application of trust assessment in the "real world." Recently, there has been consideration about ways in which to assess trust in the "real world" (Dorton & Harper, 2022; Tenhundfeld et al., 2022). Part of this push has focused on ways in which to assess trust in non-intrusive ways that allow for naturalistic interactions between the human and automation (i.e., not simply relying on subjective assessments). This push makes it all the more imperative that there is a comprehensive understanding of the alignment between results which approach trust assessment differently.

For a comparative analysis to be successful, different assessment techniques must be collected within the same study. We have therefore run a study in which physiological, behavioral, and subjective self-report measures are used, to assess trust of participants interacting with an autonomous golf-cart. Some of the techniques we used have been routinely relied upon for trust assessment, such as subjective self-reports (Brown & Galster, 2004; Foroughi et al., 2021; Seet et al., 2022; Wojton et al., 2020), heart rate (Mitkidis et al., 2015; Tolston et al., 2018), and monitoring behaviors (Bailey & Scerbo, 2007; Ferraro et al., 2018; Tenhundfeld et al., 2019). However, we also selected measures that have been linked to trust through being measures of stress but have not had the same breadth of use in empirical studies, such as heart-rate variability (HRV) (Petersen et al., 2019) and displacement behaviors (Fratczak et al., 2021). Inclusion of these measures also allows for exploratory analyses to be run which can assess these proposed measures. This sort of exploratory analysis can serve to evaluate whether proposed approaches like HRV and displacement behaviors are in fact aligning with other measures of trust, even though they are less substantiated. If these measures do converge with more frequently used measures, they may represent additional, non-invasive, trust assessment techniques that could be used by practitioners (Tenhundfeld et al., 2022).

By using an autonomous golf-cart, we are able to assess trust in an ecologically valid way that may prove beneficial for researchers and practitioners alike (Tenhundfeld et al., 2022). The use of a self-driving golf cart is comparatively novel to the literature which has relied upon self-driving cars (de Visser et al., 2023; Dikmen & Burns, 2016; Endsley, 2017; Koskinen et al., 2019; Morando et al., 2020; Tenhundfeld et al., 2019; Tomzcak et al., 2019). However, this golf cart provides a real-world analogue to the experience of a car, while simultaneously giving the researchers greater control over vehicle behaviors and participant safety. While negligible, this golf cart also provides a degree of risk that is not found in a laboratory environment, that is believed necessary for the formulation of trust (Li et al., 2019), and, as such, is most likely to manifest differences in physiological and behavioral indicators of trust. If there is convergence across the measures used, this would suggest that the differing timescales inherent with various trust assessment approaches may not be relevant when considering which approach to use. Alternatively, if there is a lack of convergence across measures, this suggests there needs to be further exploration into the nature of trust assessment, and greater consideration should be paid to measures used when assessing trust.

Methods

Participants and Recruitment

A total of 27 participants were recruited at James Madison University (JMU) by faculty members associated with the project through recruitment emails to school Listserv and word of mouth. The call for participants stated, "the purpose of this study is to better understand how participants perceive trust and reliability in autonomous vehicles before and after riding in such a vehicle" and "… along with surveys and interview questions, we will assess what aspects of the ride were enjoyable and/or potentially stressful for you." An a priori power analysis indicated that we needed 29 subjects in order to adequately power a study with an expected correlation coefficient of .5 which is the recommended minimum threshold for assessing convergent validity, although others advocate for a much higher correlation (Cheah et al., 2018). A total of 58 persons completed an "interest form" to schedule an initial meeting and experimental time with the investigators of which 27 of these persons attended the meeting and completed the study. The remaining participants never followed up to meet with the investigators. No one dropped out of the study after meeting with the investigator. At their initial meeting, the participants were advised regarding the research purpose, procedures, and safety protocols. The IRB consent form required several things to be listed regarding the features/ safety of the vehicle:

- “During the ride, a safety supervisor will be present in the back of the vehicle. The supervisor will be watching the vehicle’s path for any deviations or potential objects in its way. If something is observed, then a ‘kill switch’ will be manually thrown that will bring the vehicle to a halt within approximately 2 seconds. The vehicle will not resume until the supervisor throws the switch again.”
- “The vehicle is programmed to operate at a slow rate approximately the speed of a brisk walk. It cannot be commanded to move faster in autonomous mode.”
- “The vehicle is equipped with a 3D lidar system that can detect objects in front of the vehicle. It is programmed to stop in the event an object is in its path and resume once the object has been removed.”

> “The vehicle will only operate on known and pre-mapped paths on the JMU campus that have been previously tested by the research team.”

These meetings generally occurred in a faculty member’s office. Participants were not compensated. This study was approved by the IRB and followed ethical research guidelines.

Testbed

The JMU Autonomous Vehicle[1] (Figure 1) is a modified EZ-Go Golf Cart that has been adapted to autonomous navigation by controlling its electronic braking and throttle through custom-designed circuit boards and by utilizing an electric DC motor to automatically drive/turn the steering wheel. These electronic controls are governed by a custom Robot Operation System (ROS) program that utilizes as inputs a 3-D LiDAR for localization and obstacle detection, electronic sensors for steering column position, and user-selected destination from the GUI. For additional technical details on the project, please see El-Tawab, Sprague, and Mufti (2020); El-Tawab, Sprague, and Stewart, et al. (2020).

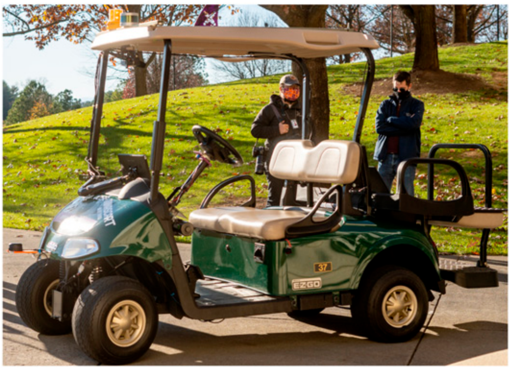

Figure 1. Autonomous vehicle testbed.

The autonomous vehicle operates on defined routes that are pre-mapped with a 3-D LiDAR with known locations as labeled in Figure 2. Given these locations, the vehicle can autonomously determine routes between destinations, execute those routes at a fixed speed, and stop for large obstacles such as people and vehicles. The vehicle operates at speeds of <3 MPH (4.8 KPH), which is generally the speed of a brisk walk. While the vehicle can stop for obstacles (vehicles, people, etc.), it cannot autonomously navigate around them. It will either wait until the obstacle has moved or the cart is manually driven around the obstacle.

Test Procedures

After receiving informed consent, the participants were provided the Automation Induced Complacency Potential – Revised survey (AICP-R) (Merritt et al., 2019) pre-ride and then a wearable device measuring heart rate (HR) and heart rate variability (HRV) (Maxim Integrated REFDES2103) (Figure 3) was placed on their wrist to monitor heart rate. The AICP-R consists of 10 questions evaluated using a 5-point Likert scale ranging from “strongly disagree” to “strongly agree.” Past research has suggested

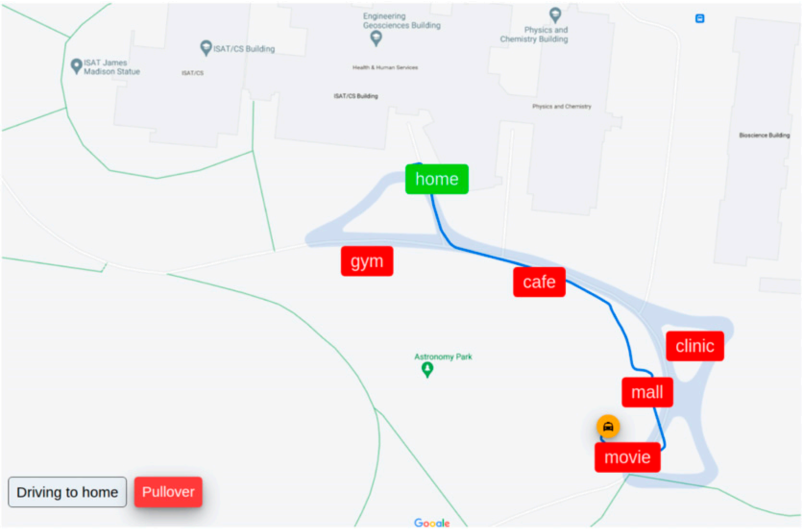

Figure 2. Graphical User Interface. *Note.* GUI for autonomous vehicle showing selectable destinations, vehicle position, destination, and pull-over controls.

that HRV and HR may increase when a user's workload is higher due to lower trust resulting in a greater degree of monitoring of the system (Khalid et al., 2016; Mitkidis et al., 2015; Perello-March et al., 2022; Tolston et al., 2018). Additionally, these measures are indications of stress which may be caused by decreased levels of trust (Reimer et al., 2010, 2016). After verifying the wearable device was operational, the faculty member and participant proceeded out of the building to the waiting autonomous vehicle that was in a courtyard outside the JMU EnGeo Building (Figure 4).

The participant was asked to sit in the passenger seat of the vehicle and then was provided a short demo on how to operate the graphical user interface (GUI). We chose to have the participant sit in the passenger seat for three reasons. First, we wanted participants to have ready access to the GUI which was installed on the passenger side of the vehicle. Secondly, for participant safety, we did not want them to be able to accidentally touch one of the pedals or get their hands caught in the steering wheel when it was being turned by the actuator. Finally, because we wanted the user's interaction with the vehicle to mimic the experience of a fully autonomous vehicle, we did not want the participants to have input beyond telling the vehicle to pull over (mechanism discussed below). Through the interface, a passenger can select a destination around the JMU East Campus area that the vehicle will autonomously navigate to. Upon reaching the destination, the passenger can select a new location. Also, the participants were

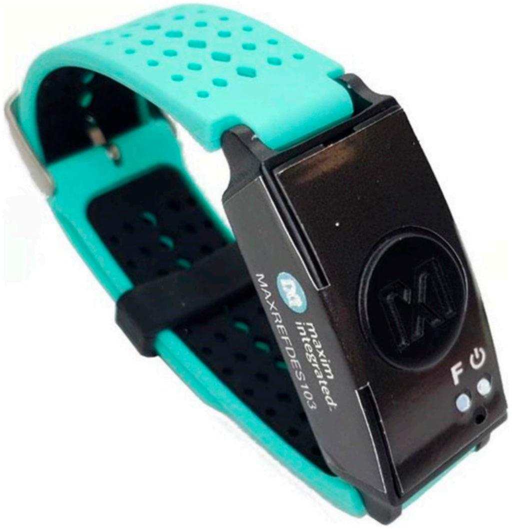

Figure 3. Maxim Integrated Wearable Device. *Note.* The Maxim Integrated MAXREFDES103 was used to detect user HR and HRV throughout the experiment.

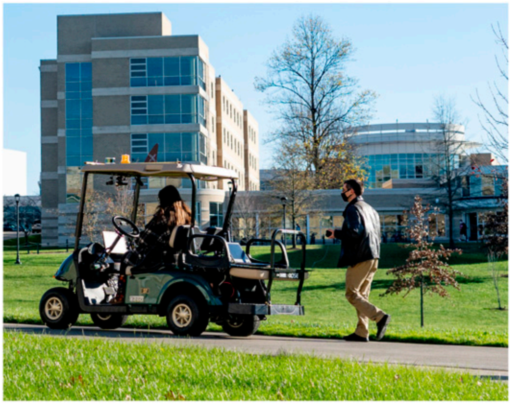

Figure 4. Testbed environment in action. *Note.* The figure shows the JMU autonomous vehicle traveling on a walking path with a "safety officer" following to manually turn off the vehicle if needed.

shown the presence of a "Pullover" button that would halt the vehicle en route. The destination labels utilized in the GUI were fictional in that the labels did not actually correspond to what was in that location (i.e., there was no movie theater at the destination labeled "movie"); however, they represented available stopping points around the larger campus area.

After completing the demonstration and answering any questions, the participants were instructed to complete two trips: one to a specific location on campus (often to "mall" as that one was farthest) and then return to the starting location "home," and the second trip was a "free choice" that could be to any destination available that the participant desired along the known routes. Throughout the trips, one of the principal investigators walked behind the vehicle as a "safety officer" to manually stop the vehicle to prevent any unanticipated collisions, if needed (see Figure 4). While the passenger was told this person was present, they were instructed to envision themselves as riding alone in the vehicle, and in general, the "safety officer" would not interact with them unless necessary. This additional safety measure was necessary as the experimental times varied throughout the week and interactions among other students, facilities vehicles, and on-campus delivery robots were possible. While the autonomous vehicle has a forward-looking LiDAR, and can automatically stop in many situations, the researchers were unsure if it would act reliably given the many different testing procedures and additional precautions that were added. No collisions between the autonomous vehicle and other objects/vehicles occurred during the experimental period. Following their ride in the cart, participants were asked to fill out the Trust of Automated Systems Test (TOAST) (Wojton et al., 2020) and were asked a series of follow-up questions. The TOAST is assessed using a 7-point Likert scale ranging from "strongly disagree" to "strongly agree." Participants were then thanked and sent on their way.

After all data had been collected, researchers went through and coded for specific events in each video. These videos were captured by two cameras, one facing forward and mounted on the front of the vehicle and the second mounted inside the vehicle facing the participant. These events fell into three different categories: obstacles, GUI interaction, and other participant behaviors (Table 1). We use the term obstacles to reference any potential impedance to the travel of the cart, but that is not to say that the cart needed to slow down for the obstacles coded. Said another way, obstacles were those things which one may reasonably assume the cart *may* have to alter its course or speed to deal with but that does not mean that the cart in fact had to do so. These obstacles were truly naturalistic (i.e., were not designated to occur at a specific time or place). Participants experienced an average of just over 17 obstacles during the course of the ride. While the GUI did not present information from the cart's sensors (Figure 2), we believed that coding for glances at the GUI still represented a facet of supervisory control for two reasons. First, the GUI still presented information about the cart's projected path and participants may have been looking to see if there were indications that the path had changed. Secondly, the GUI housed the "pullover" button, which participants may have been looking toward in order to cue the necessary action should they feel they needed to hit it (Gottlieb, 2007). This sort of "verification" behavior of looking at a central GUI has been used as a measure of trust/distrust in other self-driving vehicle research (Tenhundfeld et al., 2019) and stems from an individual's level of trust and subsequent

TABLE 1: The Categories, and Subsequent Qualifying Events, Coded for.

Obstacles	GUI interaction	Participant behaviors	
- Food robots	- Looking at the GUI	- Bracing	- Noticeable gasp
- Pedestrians	- Interacting with the GUI	- Glancing	- Fidgeting
- Cars		- Darting eyes	- Touching his/her face
- Curves in the road		- Noticeable/pronounced exhale	- Messing with hands
- Construction		- Talking to self	
		- Mouth movement (e.g., biting lips and pursed lips)	

complacency when interacting with the system (Bahner et al., 2008). The idea is that those who trust a system more are more likely to become complacent, and thus monitor/verify system behavior less. Additional participant "displacement" behaviors were coded, as these are indicators of stress response, and are used to evaluate stress in humans (Troisi, 2002). The term displacement behaviors is used to categorize behaviors that occur in circumstances in which they would not be otherwise expected (McFarland, 1966; Troisi, 2002). Because of the relationship between stress and trust, these displacement behaviors were believed to be elicited from low trusting individuals who experienced stress in the face of the uncertainty evoked by not trusting the system (Morris et al., 2017; Reimer et al., 2010, 2016). Additionally, secondary task engagement behaviors (such as checking one's phone) were coded. These secondary task engagements were coded because of the literature suggesting that secondary task engagement is an indication of complacency and overtrust in an automated system (Manzey et al., 2012; Noble et al., 2021). Each video was coded by three researchers independently by identifying the time that corresponded to each of the events. From those researcher codes, an event was determined to have happened at the earliest timestamp given by a researcher, provided at least one other researcher coded that event within the following approximately 2 s, and it had been at least 15 s since that type of event had been coded for. We selected the earliest timestamp as we wanted to reduce any potential latency between the onset of physiological identifiers and the potential behavioral or obstacle event. This helped ensure that there was an agreement amongst researchers and avoided situations where the same event was coded for at several different times depending on researcher interpretations. This yielded timestamped codes for every event which could then be synchronized with the physiological data.

To evaluate the physiological responses to coded events, we looked at the percentage change for the measure of interest from 5 s before the coded event to 15 s after the coded event. This is in line with previous research which has relied upon a 20-s window following an event (Waytz et al., 2014). We chose to shift the 20-s window to include 5 s before the coded event to account for the fact that there were likely upcoming obstacles that the participants were able to see before they came into view of the outward-facing camera. Note that the outward-facing camera was to be used by the experimenters to record the trials so that they could code events that happened in the world, and the feed from the camera was not shown to the subjects.

Results

Statistical analyses and data cleaning were conducted in both Python and R (R Core Team, 2022) using functions from the tidyverse (Wickham et al., 2019), lme4 (Bates et al., 2015), emmeans (Lenth, 2022), effects (Fox & Weisberg, 2018), ggplot2 (Wickham, 2016), and base packages.

Physiological Assessment

Data Cleaning. Three subjects' data were omitted from the analyses because they did not encounter any obstacles during their drives (N = 2) or was the only subject who encountered a car (N = 1). Data from the remaining 24 participants was subset into unique *event windows*, 50 s intervals of time 25 s before and after participants encountered an obstacle. Although the length of the event window was arbitrary, it gave us the opportunity to track changes in *heart rate*, measured in beats per minute (BPM), that occurred as participants saw, drove toward, encountered, and passed obstacles. We also captured the *time* during the session (in seconds since start) at which each event window was recorded. This allowed us to track co-occurring obstacles and control for drift in heart rate over time. Finally, we captured the *type of event* that participants encountered during each window: curves in the road, pedestrians, and food robots. All predictors were effect-coded, means-centered, and scaled prior to analysis.

Statistical Control for Individual Differences. Because physiological data is prone to noise from individual differences, we used a multilevel model to analyze our data. This model included a random-effect structure that statistically controlled for individual differences in heart rate that may have occurred during the experiment. Four possible random-effect structures were proposed that differed in their assumptions. All of the models statistically controlled for individual differences in resting heart rate (i.e., heart rate intercept). Additional models were developed to also control for individual differences in heart rate response to obstacles (i.e., event window slope), changes in heart rate throughout the study (i.e., time slope), or both. AIC values (Akaike, 1973; see Table 2) were used to identify the random-effect structure that was most likely to have produced our data. This random-effect structure was included in the final model of heart rate response to automation.

Modeling Heart Rate Response to Automation. Exploratory visualizations suggested that fluctuations in heart rate could be modeled using linear regression. Therefore, a full-factorial multilevel linear model was fitted to the data. This model included the event window, time, and event type predictors in the fixed-effect structure and the intercept, event window slope, and time slope in the random-effect structure.

This model indicated that subjects' heart rate changed when they encountered food robots but not when they encountered curves in the road or pedestrians (see 95% confidence intervals in Table 3). This effect is illustrated by the negative slopes in Figure 1: subjects' heart rate was higher when they first encountered a food robot relative to when they passed it. The model indicated that this change in heart rate became slightly more pronounced over time (B = −.07, SE = .02, t = −4.26), as seen across the panels of Figure 5, even as overall heart rate remained stable throughout the course of the experiment (B = .84, SE = 1.51, t = .56).

Subjective Assessments.

The time between when the cart started driving and when it returned to the start/stop location took an average of 620.56 s (SD = 134.25). The participants reported an average overall score on the AICP-R of 3.552 (SD = .336), with averages for the alleviating workload (which assesses an individual's attitudes toward delegating tasks to automation in order to alleviate workload) and monitoring (which assesses an individual's attitudes toward the need to monitor automation when it is being used) subscales of 4.024 (SD = .543) and 3.080 (SD = .370), respectively. Higher scores indicate a greater propensity to delegate tasks to automation (alleviating workload subscale) and to not monitor it (monitoring subscale), both of which contribute to the complacency potential. There was an average of 17.080 (SD = 7.810, range: 3–34) obstacles coded for each participant during the participant's time in the cart. Notably, the cart was perfect in avoiding all obstacles for every participant. Participants reported very high levels of trust in the system following their trial with an average score on the TOAST of 5.893 (SD = .466) (out of 7).

TABLE 2: AIC Comparisons of Possible Model Random-Effect Structures.

Model	AIC
Intercept	3499510
Intercept, event window slope	3490937
Intercept, time slope	3446890
Intercept, event window slope and time slope	3437710

Note. Because AIC values are grounded in the -2LL, lower values are indicative of a better fit.

TABLE 3: Event Window Slopes for Heart Rate as a Function of Event Type.

Event type	B	SE	95% Confidence interval
Curves	−.07	.27	[−.60, .46]
Food robots	−.55	.27	[−1.08, −.01]
Pedestrians	−.42	.27	[−.95, .11]

There was no significant correlation between complacency potential and participant trust, r (25) = .206, p = .324, and BF_{10} = .393. While we anticipated that the more experience a participant would have with the cart successfully avoiding obstacles would elevate trust (Tenhundfeld et al., 2019, 2020), the correlation between numbers of obstacles encountered and self-reported levels of trust, via the TOAST, was not significant, r (25) = −.337, p = .100, and BF_{10} = .896 (Figure 6).

Behavioral Assessments. Given that the approach to obstacles represented a period of uncertainty for which participants would be more likely to display greater displacement behaviors as well as verification behaviors (glances at/interactions with the UI), we assessed whether there was in fact a relationship. There was a significant correlation between the number of obstacles coded and the number of behaviors coded, r (25) = .477, p = .016, and BF_{10} = 3.907, as well as between the number of obstacles coded and the number of glances at/interactions with the user interface (UI), r (25) = .543, p = .005, and BF_{10} = 10.234. Because it was possible that this positive correlation could be caused by a longer duration, such that trials that lasted longer had a potentially greater number of opportunities for obstacles, behaviors, and glances we evaluated whether this was the case. To evaluate this, we standardized the counts by averaging the counts and getting a "count per minute" measure for obstacles, behaviors, and glances at/interactions with the UI. Doing this yielded even stronger results; there was a significant correlation between the number of obstacles coded (per minute) and the number of behaviors coded (per minute), r (25) = .531, p = .006, and BF_{10} = 8.480, as well as between the number of obstacles coded and the number of glances at/interactions with the UI, r (25) = .618, p < .001, and BF_{10} = 41.788. There was no correlation between duration and obstacles, r (25) = .222, p = .287, and BF_{10} = .424, duration and number of behaviors, r (25) = .029, p = .892, and BF_{10} = .250, nor duration and glances at/interactions with the UI, r (25) = −.024, p = .908, and BF_{10} = .250.

Convergence Assessments. Given that the displacement behaviors and UI glances/interactions were thought to be indications of stress which manifests in the face of low trust, we assessed whether there was a relationship between the number of these behaviors and the subjective self-report of trust (Figure 7). There was no correlation between self-reported trust (TOAST) and the number of behaviors coded, r (25) = −.066, p = .754, and BF_{10} = .260, nor was there any correlation between self-reported trust and the number of glances at/interactions with the UI, r (25) = −.122, p = .561, and BF_{10} = .291. Given that both the number of behaviors and number of glances at/interactions with the UI are hypothesized to be negatively correlated with trust, we evaluated the convergence between these two measures and found a significant correlation, r (25) = .609, p = .001, and BF_{10} = 34.593.

There was no significant correlation between participant scores on the monitoring subscale

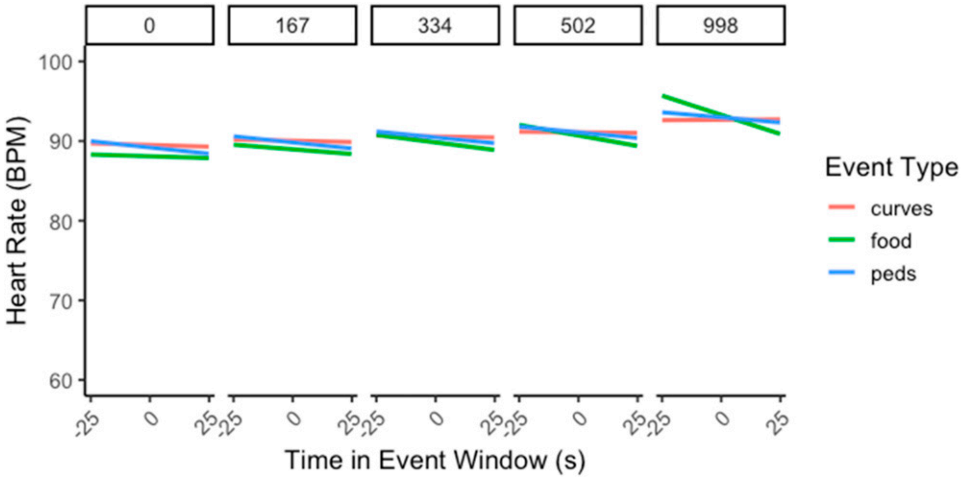

Figure 5. Heart rate during the event window changed as a function of event type over drive time. *Note.* Panels represent consecutive time slices (s) throughout the drive; narrow error ribbons represent ± 1*SE*.

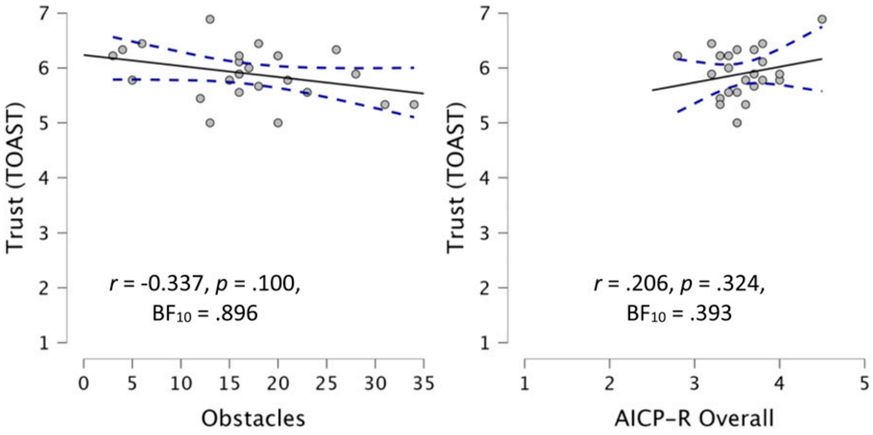

Figure 6. Scatterplots for self-reported trust. *Note.* Scatterplots between the number of obstacles encountered and self-reported trust (left) as well as AICP-R and self-reported trust (right). A solid line is regressed upon the data, and dashed lines represent the 95% CI for that regression line.

(which is supposed to predict the frequency with which an individual believes automated systems should be monitored) and the number of times they monitored the system by glancing at/ interacting with the UI, r (25) = −.179, p = .391, and BF_{10} = .352.[2]

While there was not a difference between BPM changes at baseline and those in response to obstacles, BPM changes are still thought to indirectly assess trust (Khalid et al., 2016). We therefore assessed whether there was a relationship between BPM changes and the other subjective and behavioral indicators of trust (Figure 8). There was no significant correlation between the average BPM in response to obstacles and self-reported trust (TOAST), r (25) = .225, p = .279, and BF_{10} = .432, the number of UI glances/interactions and average BPM change in response to obstacles, r (25) = −.039, p = .851, and BF_{10} = .252, or number of displacement behaviors, r (25) = −.043, p = .838, and BF_{10} = .253.

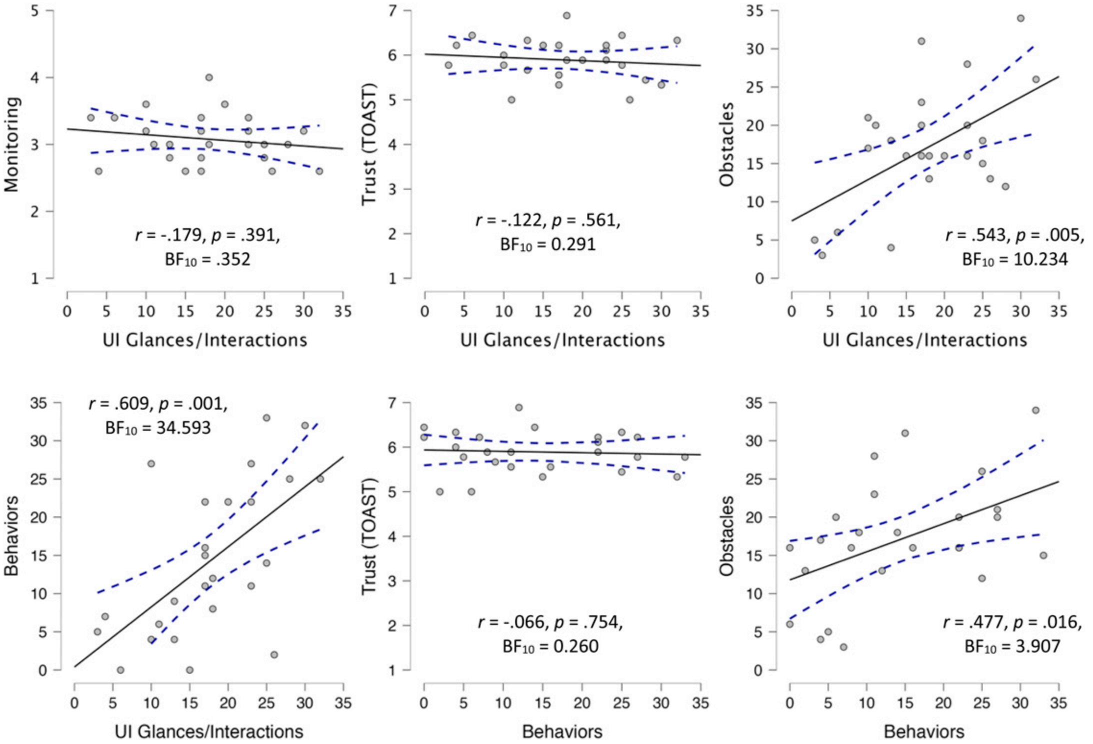

Figure 7. Scatterplots between behavioral and self-report measures used. *Note.* Scatterplots between the number of UI glances/interactions and monitoring subscale of AICP-R (top-left), self-reported trust (top-middle), number of obstacles encountered (top-right), and number of behaviors coded for (bottom-left), along with scatterplots between the number of behaviors coded for and self-reported trust (bottom-middle) and number of obstacles encountered (bottom-right). A solid line is regressed upon the data, and dashed lines represent the 95% CI for that regression line.

Similarly, we compared HRV changes in response to obstacles to the other subjective and behavioral indicators of trust to see whether there was an agreement amongst these different approaches. There was no significant difference between the average HRV in response to obstacles and self-reported trust, r (25) = −.155, p = .461, and BF_{10} = .321. While there was a significant correlation, the Bayes Factors reported only anecdotal evidence in favor of a correlation between average HRV in response to obstacles and the number of UI glances/interactions, r (25) = .418, p = .038, and BF_{10} = 1.922, and moderate in favor of the correlation between average HRV in response to obstacles and the number of displacement behaviors, r (25) = .475, p = .016, and BF_{10} = 3.768. Despite both heartrate and HRV being thought to be predictive of trust, there was no significant correlation between average BPM change in response to obstacles and average HRV change in response to obstacles, r (25) = .051, p = .808, and BF_{10} = .255.

Discussion

The current study was a comparative analysis of different trust assessment techniques. Various trust measures have been used and proposed as effective ways to evaluate one's trust; however, little research to date has explored their ability to converge upon a single finding. Not only is trust not a monolithic construct, but the different assessment approaches also used by the field rely on different assumptions and timescales for assessment. As such, it is important to understand whether there is an agreement between

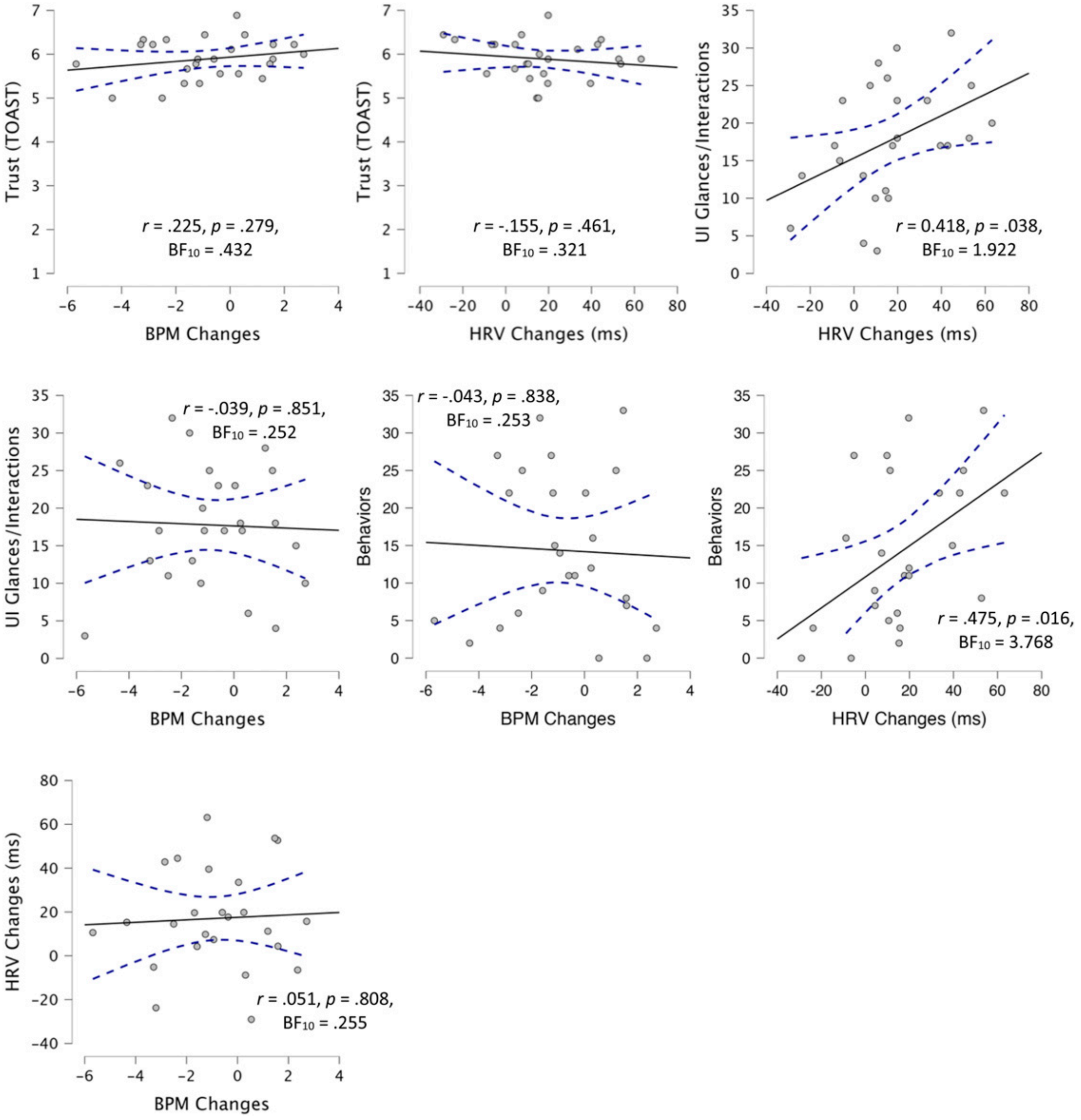

Figure 8. Scatterplots for physiological measures. *Note*. Scatterplots for trust, number of UI glances/interactions, and behaviors, with BPM and HRV changes in response to obstacles in addition to scatter plot for the relationship between HRV and BPM. A solid line is regressed upon the data, and dashed lines represent the 95% CI for that regression line.

results for these different approaches as disparate results may arise in any given study as a function of the assessment approach used. Additionally, with new measures being proposed or called for, careful consideration should be paid to the ability of the measures to align with existing approaches that are more empirically substantiated.

As such, we chose to use subjective self-report, behavioral, and physiological data to assess trust, each of which has a body of literature detailing its use for assessment of trust (Ajenaghughrure et al., 2020; Akash et al., 2018; Banks et al., 2018; Mitkidis et al., 2015; Sauer et al., 2016; Schwarz et al., 2019; Thayer et al., 2012; Tolston et al., 2018; Wang et al., 2018;

Wojton et al., 2020). In this study, we selected the TOAST for our subjective assessment measure because of its good intrasubject reliability, which is needed for individual differences research (Wojton et al., 2020). Additionally, we used the AICP-R scale to assess participant complacency potential when interacting with an automated system (Merritt et al., 2019). For physiological data, we collected heart rate and HRV. Heart rate has been used to assess trust in multiple ways (Mitkidis et al., 2015; Tolston et al., 2018); however, as abovementioned, we used the approach of Waytz et al. (2014), which involved evaluating the physiological changes over a 20-s window following an event. While the use of HRV has not been used, to our knowledge, to assess trust, it has been shown to correlate with stress and workload (Nickel & Nachreiner, 2003; Thayer et al., 2012), both of which are impacted by one's trust in and reliance on a system (Parasuraman & Riley, 1997; Sauer et al., 2011), and thus could be considered a non-invasive approach to trust assessment (Tenhundfeld et al., 2022). Finally, for behavioral measures, we relied on checking/verification behaviors (Bahner et al., 2008; Ezer et al., 2007; Tenhundfeld et al., 2019), as well as the coding of "displacement" behaviors which have been shown to appear in response to stress (Burgoon et al., 2021; E. J. Cohen et al., 2018; Mohiyeddini et al., 2013; Mohiyeddini & Semple, 2013; Troisi, 2002),but have also been used in trust research (Fratczak et al., 2021; Hald et al., 2019).

Contrary to our expectations, our data did not show any real degree of convergence across measures. In fact, there was moderate evidence in favor of the null hypothesis (as evidenced by Bayes Factor values below .333) for the relationship between an individual's TOAST score and the UI glances/interactions, behaviors, and HRV. The only relationship for which there was strong evidence in favor of a correlation was between the number of UI glances/interactions and the number of coded behaviors. There was even moderate evidence against a relationship between BPM and HRV in response to real-world obstacles.

There are several possible explanations for this lack of convergence. The first possible explanation is that some of these measures were not actually assessing trust. The measure of heart rate variability has not, to our knowledge, been used before to assess trust in this way. While there is a theoretical reason to believe HRV may be able to provide some information about the user's trust, this is theoretical in nature and thus it may simply be the case that HRV is not a reliable way of assessing trust as there are other factors which directly impact it (Fatisson et al., 2016). This is supported by the fact that a more well-recognized physiological measure of trust (heart rate) did not correlate with HRV even on the same timescale. Additionally, while there has been some limited research on the use of displacement behaviors as a trust assessment technique (Fratczak et al., 2021; Hald et al., 2019), there has been comparatively little research validating it, and the displacement behaviors we used were different than those used before. The displacement behaviors we selected were deemed more relevant to the task at hand and are well-established in the literature on displacement behaviors (Troisi, 2002).

The second possible explanation for the lack of convergence is the fact that these trust assessment approaches represented fundamentally different timescales. Whereas the subjective assessment (TOAST) was administered after the participant completed his/her drive, the physiological and behavioral data were collected throughout the experiment. As mentioned in the introduction, these differing timescales present problems for trust assessment. Surveys which are administered after trials are complete only collect trust data after the entirety of exposure to the system. On the other hand, behavioral and physiological data are collected throughout the experiment and therefore may be picking up the process of trust formation (more on this below). Additionally, as evidenced by our analyses, there are a variety of ways in which to analyze physiological and behavioral data. Some approaches involve processing data continuously, while others deal with the data in aggregate.

A third possible explanation is that, regardless of the timescales, there are different categories of trust (i.e., situational, dispositional, and learned) which may be what is being assessed by any given measure. Studies that have demonstrated an assessment technique to be effective may be able to map the results onto trust but would be unable to establish what category of trust is being measured. Said another way, if assessment A maps onto situational trust, but assessment B maps onto learned trust, it is appropriate to say that both assessment approaches measure trust, but they may vary independently from one another. This is in line with existing research on construct and discriminant validity and could explain the lack of convergence between widely used measures of trust like subjective assessments and monitoring behaviors (Campbell & Fiske, 1959). In order to ascertain whether these assessment techniques are mapping onto the same aspect of trust, comparative studies, like this one, are needed.

The fourth possible explanation for the lack of convergence pertains to the methodology used here. Because this paradigm had not been used before, it is difficult to establish whether the null results were paradigmatic or due to one of the other abovementioned factors. There is some evidence for this as there was nearly uniform high trust on the subjective assessment (which admittedly took the authors by surprise). This could have been a function of either the recruitment (i.e., self-selection bias)/consent documentation or the system's performance. While positive news for the developers of the system, this uniformity of trust (at least in subjective assessment) means that there may have been insufficient variability to truly tease apart degrees of convergence in the measures. This may have been a function of the system performance, the presence of a safety officer, or simply the relatively little amount of time that the participants spent interacting with the vehicle. Finally, the nature of the paradigm was such that participants were outside during the course of the experiment. This means that they were exposed to temperature/weather which may have impacted the physiological measures in ways that would not have affected subjective responses. Being outside means that they were also subjected to passing social interactions during which non-participants may have been staring at the participant/cart as the presence of an autonomous golf cart with various sensors is novel and attention-grabbing. These factors could have affected behavioral and physiological data in ways that were not related to trust, thereby making the data noisier and obfuscating any results that may have otherwise been detectable.

Ultimately, more research is needed in order to understand the reasons that these data did not converge. Future research should work to tease apart the contributions of different timescales and how these assessment approaches may map onto the different components of trust, especially in order to establish whether proposed approaches actually do measure trust rather than some other factor such as workload or general stress. This also may be better done in an established paradigm in which these various assessment techniques have been reported to assess trust.

Conclusion

In conclusion, we ran a comparative and exploratory analysis of different trust measurement approaches. Had measures converged upon one another, we would have had reason to believe that the trust assessment selected for any study would not matter all that much, as they would be yielding the same result. However, we found very little, to no, convergence across measures. This suggests that certain measures (e.g., HRV and displacement behaviors) may not be a reliable assessment approach for trust, while others may be task-dependent, measuring different components of trust, or results may differ simply as a function of the timescales in which these measurements were used. However, future research should be done in a more controlled environment that may be able to more concretely establish whether there is an alignment between these different trust assessment techniques. Such efforts are needed in order to provide trust assessment approaches for both researchers and practitioners.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Nathan L. Tenhundfeld https://orcid.org/0000-0002-3753-8096

Notes

1. https://av.cise.jmu.edu/#home
2. Results for correlations between numbers of behaviors, interactions with/glances at the UI, number of obstacles, TOAST, and AICP-R have previously been reported in Cotter et al., 2022.

References

Ajenaghughrure, I. B., Da Costa Sousa, S., & Lamas, D. (2020). Measuring trust with psychophysiological signals: A systematic mapping study of approaches used. *Multimodal Technologies and Interaction*, *4*(3). 63. https://doi.org/10.3390/mti4030063

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Proceedings of the Section International Symposium on Information Theory, Tsahkadsor, Armenia. September 1971. 267–281.

Akash, K., Hu, W. L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems*, *8*(4), 1–20. https://doi.org/10.1145/3132743

Arsalan, A., & Majid, M. (2021). Human stress classification during public speaking using physiological signals. *Computers in Biology and Medicine*, *133*, 104377. https://doi.org/10.1016/j.compbiomed.2021.104377

Bahner, J. E. E., Hüper, A. D., & Manzey, D. H. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies*, *66*(9), 688–699. https://doi.org/10.1016/j.ijhcs.2008.06.001

Bailey, N. R., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science*, *8*(4), 321–348. https://doi.org/10.1080/14639220500535301

Banks, V. A., Eriksson, A., O'Donoghue, J., & Stanton, N. A. (2018). Is partially automated driving a bad idea? Observations from an on-road study. *Applied Ergonomics*, *68*, 138–145.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brown, R. D., & Galster, S. M. (2004). Effects of reliable and unreliable automation on subjective measures of mental workload, situation awareness, trust and confidence in a dynamic flight task. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting*, *48*, 147–151. https://doi.org/10.1177/154193120404800132

Burgoon, J. K., Wang, X., Chen, X., Pentland, S. J., Dunbar, N. E., & Bond, G. D. (2021). Nonverbal behaviors " speak " relational messages of dominance, trust, and composure. *Frontiers in Psychology*, *12*(624177), 1–17. https://doi.org/10.3389/fpsyg.2021.624177

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cheah, J. H., Sarstedt, M., Ringle, C. M., Ramayah, T., & Ting, H. (2018). Convergent validity assessment of formatively measured constructs in PLS-SEM: On using single-item versus multi-item measures in redundancy analyses. *International Journal of Contemporary Hospitality Management*, *30*(11), 3192–3210. https://doi.org/10.1108/IJCHM-10-2017-0649.

Chen, F., Marcus, N., Khawaji, A., & Zhou, J. (2015). Using galvanic skin response (GSR) to measure trust and cognitive load in the text-chat environment. *Conference on Human Factors in Computing Systems - Proceedings*, *18*, 1989–1994. https://doi.org/10.1145/2702613.2732766

Cohen, E. J., Bravi, R., & Minciacchi, D. (2018). The effect of fidget spinners on fine motor control. *Scientific Reports*, *8*(3144), 3144–3149. https://doi.org/10.1038/s41598-018-21529-0

Cohen, M. S., Parasuraman, R., Serfaty, D., & Andes, R. C. (1997). Trust in decision aids: A model and its training implications. *U.S. Army aviation and troop command*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.90.2591

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

de Visser, E. J., Phillips, E., Tenhundfeld, N., Donadio, B., Barentine, C., Kim, B., Madison, A., Ries, A., & Tossell, C. C. (2023). Trust in automated parking systems: A mixed methods evaluation. *Transportation Research Part F: Traffic Psychology and Behaviour*, *96*, 185–199. https://doi.org/10.1016/j.trf.2023.05.018

Dikmen, M., & Burns, C. M. (2016). Autonomous driving in the real world: Experiences with tesla autopilot and summon. *Proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications* (pp. 225–228). AutomotiveUI '16). https://doi.org/10.1093/ndt/gfn515

Dorton, S. L., & Harper, S. B. (2022). A naturalistic investigation of trust, AI, and intelligence work. *Journal of Cognitive Engineering and Decision Making*, *16*(4), 222–236. https://doi.org/10.1177/15553434221103718

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

Eilebrecht, B., Wolter, S., Lem, J., Lindner, H. J., Vogt, R., Walter, M., & Leonhardt, S. (2012). The relevance of HRV parameters for driver workload detection in real world driving. 2012 Computers in Cardiology, Krakow, Poland, 09-12 September 2012. *39*, 409–412.

El-Tawab, S. S., Sprague, N., & Mufti, A. (2020). Autonomous vehicles: Building a test-bed prototype at a controlled environment. In IEEE world forum on Internet of things, WF-IoT 2020—symposium proceedings. New Orleans, LA, 02-16 June 2020. https://doi.org/10.1109/WF-IoT48130.2020.9221222

El-Tawab, S. S., Sprague, N., Stewart, M., Pareek, M., & Zubov, P. (2020). Enhanced interface for autonomously driven golf cart in a networked controlled environment. Proceedings of the 2020 9th International Connference on Software and Information Engineering. 174–179. https://doi.org/10.1145/3436829.3436875

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

Estepp, J. R., de Visser, E. J., Klosterman, S. L., & Galster, S. M. (2018). Predicting trust calibration and workload using machine-learning classification of neurophysiological measurement during the monitoring of automation. *ACM Transactions on Interactive Intelligent Systems*.

Ezer, N., Fisk, A. D., & Rogers, W. A. (2007). Reliance on automation as a function of expectation of reliability, cost of verification, and age. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting*, *51*(1), 6–10. https://doi.org/10.1177/154193120705100102

Fatisson, J., Oswald, V., & Lalonde, F. (2016). Influence diagram of physiological and environmental factors affecting heart rate variability: An extended literature overview. *Heart International*, *11*(1), e32. https://doi.org/10.5301/heartint.5000232

Ferraro, J., Clark, L., Christy, N., & Mouloua, M. (2018). Effects of automation reliability and trust on system monitoring performance in simulated flight tasks. *Proceedings of the Human*

Factors and Ergonomics Society - Annual Meeting, *62*, 1232–1236. https://doi.org/10.1177/1541931218621283

Foroughi, C. K., Devlin, S., Pak, R., Brown, N. L., Sibley, C., & Coyne, J. T. (2021). Near-perfect automation: Investigating performance, trust, and visual attention allocation. *Human Factors*, *65*(4), 546–561. https://doi.org/10.1177/00187208211032889

Fox, J., & Weisberg, S. (2018). *An R companion to applied regression*. Sage publications.

Fratczak, P., Goh, Y. M., Kinnell, P., Justham, L., & Soltoggio, A. (2021). Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction. *International Journal of Industrial Ergonomics*, *82*, 103078. https://doi.org/10.1016/j.ergon.2020.103078

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the technology acceptance model to assess automation. *Cognition, Technology and Work*, *14*(1), 39–49. https://doi.org/10.1007/s10111-011-0194-3

Gottlieb, J. (2007). From thought to action: The parietal cortex as a bridge between perception, action, and cognition. *Neuron*, *53*(1), 9–16. https://doi.org/10.1016/j.neuron.2006.12.009

Hald, K., Rehm, M., & Moeslund, T. B. (2019). *Proposing human-robot trust assessment through tracking physical apprehension signals in close-proximity human-robot collaboration. 2019 28th IEEE international conference on robot and human interactive communication*. https://doi.org/10.1109/RO-MAN46459.2019.8956335

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2020). *Evolving trust in robots: Specification through sequential and comparative meta-analyses. Human factors*. https://doi.org/10.1177/0018720820922080

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2015). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors*, *58*(3), 509–519. https://doi.org/10.1177/0018720815625744

Hoff, K. A., & Bashir, M. (2015). Trust in automation. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hoover, A., Singh, A., Fishel-Brown, S., & Muth, E. (2012). Real-time detection of workload changes using heart rate variability. *Biomedical Signal Processing and Control*, *7*(4), 333–341. https://doi.org/10.1016/j.bspc.2011.07.004

Hussein, A., Elsawah, S., & Abbass, H. A. (2020). The reliability and transparency bases of trust in human-swarm interaction: Principles and implications. *Ergonomics*, *63*(9), 1116–1132. https://doi.org/10.1080/00140139.2020.1764112

Israelsen, B. W., & Ahmed, N. R. (2019). "Dave...I can assure you ...that it's going to be all right ..." A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. *ACM Computing Surveys*, *51*(6), 1–37. https://doi.org/10.1145/3267338

Khalid, H. M., Shiung, L. W., Nooralishahi, P., Rasool, Z., Helander, M. G., Kiong, L. C., & Ai-Vyrn, C. (2016). Exploring psycho-physiological correlates to trust: Implications for human-robot-human interaction. *Proceedings of the Human Factors and Ergonomics Society*, *60*(1), 696–700. https://doi.org/10.1177/1541931213601160

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, *61*, 101595. https://doi.org/10.1016/j.tele.2021.101595

Kohn, S. C., Quinn, D., Pak, R., de Visser, E. J., & Shaw, T. H. (2018). Trust repair strategies with self-driving vehicles: An exploratory study. *Proceedings of the Human Factors and Ergonomics Society*, 2(1), 1108–1112. https://doi.org/10.1177/1541931218621254

Koskinen, K., Lyyra, A., Mallat, N., & Tuunainen, V. K. (2019). Trust and risky technologies: Aligning and coping with tesla autopilot. *Proceedings of the 52nd Hawaii International Conference on System Sciences,* Hawaii. 2019. 5777–5786.

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, *50*(1), 569–598. https://doi.org/10.1146/annurev.psych.50.1.569

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2019). The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human Factors*, *62*(5), 718–736. https://doi.org/10.1177/0018720819853686

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lenth, R. (2022). *Emmeans: Estimated marginal means, aka least-squares means R package version 1.4.7*.

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *Academy of Management Review*, *23*(3), 438–458. https://doi.org/10.2307/259288

Li, M., Holthausen, B. E., Stuck, R. E., & Walker, B. N. (2019). No risk no trust: Investigating perceived risk in highly automated driving. *Proceedings of the Annual Automotive UI Conference*, 177–185. https://doi.org/10.1145/3342197.3344525

Lu, Y., & Sarter, N. (2019). Eye tracking: A process-oriented method for inferring trust in automation as a function of priming and system reliability. *IEEE Transactions on Human-Machine Systems*, *49*(6), 560–568. https://doi.org/10.1109/THMS.2019.2930980

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87. https://doi.org/10.1177/1555343411433844

Matsuyama, L., Zimmerman, R., Eaton, C., Weger, K., Mesmer, B., Tenhundfeld, N. L., Van Bossuyt, D., & Semmens, R. (2021). Determinants that influence the acceptance and adoption of mission critical autonomous systems. Proceedings of txhe AIAA SciTech Forum. Virtual Event. January 2021. https://doi.org/10.2514/6.2021-1156

McFarland, D. (1966). On the causal and functional significance of displacement activities. *Zeitschrift fur Tierpsychologie*, *23*(2), 217–235. https://doi.org/10.1111/j.1439-0310.1966.tb01600.x

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., & Shirase, L. (2019). Automation-induced complacency potential: Development and validation of a new scale. *Frontiers in Psychology*, *10*(225), 1–13. https://doi.org/10.3389/fpsyg.2019.00225

Mirnig, A. G., Wintersberger, P., Sutter, C., & Ziegler, J. (2016). A framework for analyzing and calibrating trust in automated vehicles. AutomotiveUI 2016 - 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Adjunct Proceedings, Ann Arbor, MI. October 2016. 33–38. https://doi.org/10.1145/3004323.3004326

Mitkidis, P., McGraw, J. J., Roepstorff, A., & Wallot, S. (2015). Building trust: Heart rate synchrony and arousal during joint action increased by public goods game. *Physiology and Behavior*, *149*(1), 101–106. https://doi.org/10.1016/j.physbeh.2015.05.033

Mohiyeddini, C., Bauer, S., & Semple, S. (2013). Displacement behaviour is associated with reduced stress levels among men but not women. *PLoS One*, *8*(2), e56355–e56359. https://doi.org/10.1371/journal.pone.0056355

Mohiyeddini, C., & Semple, S. (2013). Displacement behaviour regulates the experience of stress in men. *Stress: The International Journal on the Biology of Stress*, *16*(2), 163–171. https://doi.org/10.3109/10253890.2012.707709

Morando, A., Gershon, P., Mehler, B., & Reimer, B. (2020). Driver-initiated tesla autopilot disengagements in naturalistic driving. *Proceedings - 12th international ACM conference on automotive user interfaces and interactive vehicular applications*, AutomotiveUI, 57–65. https://doi.org/10.1145/3409120.3410644

Morris, D. M., Erno, J. M., & Pilcher, J. J. (2017). Electrodermal response and automation trust during simulated self-driving car use. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting*, *61*(1), 1759–1762. https://doi.org/10.1177/1541931213601921

Nickel, P., & Nachreiner, F. (2003). Sensitivity and diagnosticity of the 0.1-hz component of heart rate variability as an indicator of mental workload. *Human Factors: The Journal of the Human*

Factors and Ergonomics Society, *45*(4), 575–590. https://doi.org/10.1518/hfes.45.4.575.27094

Noble, A. M., Miles, M., Perez, M. A., Guo, F., & Klauer, S. G. (2021). Evaluating driver eye glance behavior and secondary task engagement while using driving automation systems. *Accident Analysis and Prevention*, *151*, 105959. https://doi.org/10.1016/j.aap.2020.105959

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. *52*(3), 381–410 https://doi.org/10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Perello-March, J. R., Burns, C. G., Woodman, R., Elliott, M. T., & Birrell, S. A. (2022). Driver state monitoring: Manipulating reliability expectations in simulated automated driving scenarios. *IEEE Transactions on Intelligent Transportation Systems*, *23*(6), 5187–5197. https://doi.org/10.1109/TITS.2021.3050518

Petersen, L., Robert, L., Yang, X. J., & Tilbury, D. M. (2019). Situational awareness, driver's trust in automated driving systems and secondary task performance. *SAE International Journal of Connected and Autonomous Vehicles*, *2*(2), 1–26.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reimer, B., Mehler, B., & Coughlin, J. F. (2010). *An evaluation of driver reactions to new vehicle parking assist technologies developed to reduce driver stress*. New England University Transportation Center, https://doi.org/10.13140/RG.2.1.3317.3361

Reimer, B., Mehler, B., & Coughlin, J. F. (2016). Reductions in self-reported stress and anticipatory heart rate with the use of a semi-automated parallel parking system. *Applied Ergonomics*, *52*, 120–127. https://doi.org/10.1016/j.apergo.2015.07.008

Robinette, P., Howard, A. M., & Wagner, A. R. (2015). Timing is key for robot trust repair. International Conference on Social Robotics, 61–71. https://doi.org/10.1007/978-3-319-25554-5

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. ACM/IEEE International Conference on Human-Robot Interaction. Christchurch, *New Zealand. 07-10* March 2016, https://doi.org/10.1109/HRI.2016.7451740

Sauer, J., Chavaillaz, A., & Wastell, D. (2016). Experience of automation failures in training: Effects on trust, automation bias, complacency and performance. *Ergonomics*, *59*(6), 767–780. https://doi.org/10.1080/00140139.2015.1094577

Sauer, J., Kao, C. S., Wastell, D., & Nickel, P. (2011). Explicit control of adaptive automation under different levels of environmental stress. *Ergonomics*, *54*(8), 755–766. https://doi.org/10.1080/00140139.2011.592606

Schwarz, C., Gaspar, J., & Brown, T. (2019). The effect of reliability on drivers' trust and behavior in conditional automation. *Cognition, Technology and Work*, *21*(1), 41–54. https://doi.org/10.1007/s10111-018-0522-y

Sebok, A., & Wickens, C. D. (2017). Implementing lumberjacks and black swans into model-based tools to support human-automation interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(2), 189–203. https://doi.org/10.1177/0018720816665201

Seet, M., Harvy, J., Bose, R., Dragomir, A., Bezerianos, A., & Thakor, N. (2022). Differential impact of autonomous vehicle malfunctions on human trust. *IEEE Transactions on Intelligent Transportation Systems*, *23*(1), 548–557. https://doi.org/10.1109/TITS.2020.3013278

Segura, E. M., Kriegel, M., Aylett, R., Deshmukh, A., & Cramer, H. (2012). How do you like me in this: User embodiment preferences for companion agents. *Intelligent Virtual Agents*, *7502*, 112–125. https://doi.org/10.1007/978-3-642-33197-8-12

Tenhundfeld, N. L., Demir, M., & de Visser, E. (2022). Assessment of trust in automation in the "real world": Requirements for new trust in automation measurement techniques for use by practitioners. *Journal of Cognitive Engineering and Decision Making*, *16*(2), 101–118. https://doi.org/10.1177/15553434221096261

Tenhundfeld, N. L. ., de Visser, E. J., Haring, K. S. ., Ries, A. J., Finomore, V. S., & Tossell, C. C. (2019). Calibrating trust in automation through familiarity with the autoparking feature of a Tesla Model X. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 279–294. https://doi.org/10.1177/1555343419869083

Tenhundfeld, N. L., de Visser, E. J., Ries, A. J., Finomore, V. S., & Tossell, C. C. (2020). Trust and distrust of automated parking in a tesla model X. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *62*(2), 194–210. https://doi.org/10.1177/0018720819865412

Thayer, J. F., Åhs, F., Fredrikson, M., Sollers, J. J., & Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience and Biobehavioral Reviews*, *36*(2), 747–756. https://doi.org/10.1016/j.neubiorev.2011.11.009

Tolston, M. T., Funke, G. J., Alarcon, G. M., Miller, B., Bowers, M. A., Gruenwald, C., & Capiola, A. (2018). Have a heart: Predictability of trust in an autonomous agent teammate through team-level measures of heart rate synchrony and arousal. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting*, *62*(1), 714–715. https://doi.org/10.1177/1541931218621162

Tomzcak, K., Pelter, A., Gutierrez, C., Stretch, T., Hilf, D., Donadio, B., Tenhundfeld, N. L., de Visser, E. J., & Tossell, C. C. (2019). Let tesla park your tesla: Driver trust in a semi-automated car. Proceedings of the annual systems and information engineering design symposium (SIEDS) conference. Charlottesville, VA, 26-26 April 2019. https://doi.org/10.1109/SIEDS.2019.8735647

Troisi, A. (2002). Displacement activities as a behavioral measure of stress in nonhuman primates and human subjects. *Stress: The International Journal on the Biology of Stress*, *5*(1), 47–54. https://doi.org/10.1080/102538902900012378

Walker, F., Verwey, W., & Martens, M. H. (2018). *Gaze behaviour as a measure of trust in automated vehicles*. *Proceedings of the 6th Humanist Conference* (pp. 13–14). The Hague, Netherlands.

Wang, M., Hussein, A., Rojas, R. F., Shafi, K., & Abbass, H. A. (2018). EEG-based neural correlates of trust in human-autonomy interaction. Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, Bangalore, 18-21 November 2018, 350–357. https://doi.org/10.1109/SSCI.2018.8628649

Warm, J. S., Matthews, G., & Finomore, V. S. (2008). Vigilance, workload, and stress. In *Performance under stress* (pp. 115–141). CRC Press.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and automation bias in the use of imperfect automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(5), 728–739. https://doi.org/10.1177/0018720815581940

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., & Spinu, V., …, (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wiczorek, R., & Meyer, J. (2019). Effects of trust, self-confidence, and feedback on the use of decision automation. *Frontiers in Psychology*, *10*(519), 1–12. https://doi.org/10.3389/fpsyg.2019.00519

Wojton, H., Porter, S., T Lane, D., Bieber, C., & Madhavan, P.. (2020). Initial validation of the trust of automated systems test (TOAST). *The Journal of Social Psychology*, *160*(6), 735–750. https://doi.org/10.1080/00224545.2020.1749020

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI*, *17*, 408–416. https://doi.org/10.1145/2909824.3020230

Zhang, T., Tao, D., Qu, X., Zhang, X., Lin, R., & Zhang, W. (2019). The roles of initial trust and perceived risk in public's acceptance of automated vehicles. *Transportation Research Part C: Emerging Technologies*, *98*, 207–220. https://doi.org/10.1016/j.trc.2018.11.018

Nathan L. Tenhundfeld is an Associate Professor at the University of Alabama in Huntsville. He did his postdoctoral research in the Warfighter Effectiveness Research Center at the United States Air Force Academy. He received a PhD in Cognitive Psychology from Colorado State University in 2017.

Jason Forsyth is an Associate Professor of Engineering at James Madison University. He received his PhD in Computer Engineering from Virginia Tech in May 2015. His major research interests are in wearable/ubiquitous computing and engineering education.

Nathan Sprague received his Sc.B. in Computer Science from Brown University in 1997 and a PhD in Computer Science from the University of Rochester in 2004. He is an Associate Professor in the Department of Computer Science at James Madison University. His research has spanned multiple areas of machine learning and robotics, including reinforcement learning in multiple-goal and high-dimensional domains, deep learning, and intelligent user interfaces for autonomous vehicles.

Dr Samy El-Tawab received his PhD in Computer Science from Old Dominion University in 2012. Dr Samy El-Tawab is an Associate Professor and the Information Technology Program Director, Department of Computer Science, in the College of Integrated Science and Engineering at James Madison University. His class "Autonomous Vehicles" won the Governor's Technology Award for 2018 for using "Technology in Education." Dr El-Tawab has been working on the issues surrounding Intelligent Transportation, Vehicular Ad-Hoc Networks, internet of Things (IoT), and Cyber Security. Dr El-Tawab is the co-founder of JMU Autonomous Cart (JACart) Research Group.

Jenna E. Cotter is currently pursuing a Master's degree in Psychology at the University of Alabama in Huntsville, with plans to continue on to a doctoral program. She received a Bachelor of Science in Psychology from Sam Houston State University in 2020.

Lisa Vangsness is an Assistant Professor at the University of Alabama in Huntsville. She researches metacognition (e.g., human-automation teaming and student engagement) and frequently serves as a statistical consultant on a variety of projects. Dr Vangsness received her PhD in Psychology from Kansas State University in 2019.