Behavioral Measurement of Trust in Automation: The Trust Fall

David Miller[1], Mishel Johns[1], Brian Mok[1], Nikhil Gowda[2], David Sirkin[1], Key Lee[1], Wendy Ju[1]
[1]Center for Design Research at Stanford, [2]Renault North America
{davebmiller, mishel, brianmok, wendyju}@stanford.edu, nikhil.gowda@renault.com, keyjlee@gmail.com

ABSTRACT

Stating that one trusts a system is markedly different from demonstrating that trust. To investigate trust in automation, we introduce the trust fall: a two-stage behavioral test of trust. In the trust fall paradigm, first the one learns the capabilities of the system, and in the second phase, the 'fall,' one's choices demonstrate trust or distrust. Our first studies using this method suggest the value of measuring behaviors that demonstrate trust, compared with self-reports of one's trust. Designing interfaces that encourage appropriate trust in automation will be critical for the safe and successful deployment of partially automated vehicles, and this will rely on a solid understanding of whether these interfaces actually inspire trust and encourage supervision.

INTRODUCTION

Understanding how humans will interact with safety critical vehicle systems is a pressing concern, as automated driving systems will soon be putting complex robotic systems in the hands of the public. Drivers, in contrast to pilots, vary widely in their abilities, seldom have extensive training, and cannot be expected to understand complex literature about the automated driving features in their own vehicle. The potential diversity of vehicle systems also makes this a substantial challenge for designers to address.

A key challenge from a design standpoint is designing systems that individuals will trust appropriately; granting the system agency when appropriate, and taking control when human action may be warranted (J. D. Lee & See, 2004). The vehicle will have to communicate effectively on two levels: it must provide proper information relative to immediate actions, and also must supply information that helps drivers to build an accurate and usable mental model (Johnson-Laird, 1980; Norman, 1983). Automated driving systems that are highly capable in normal circumstances may invite intentional misuse due to overtrust—potentially leading to failure to properly supervise the system's operation. (Parasuraman & Riley, 1997; Sheridan, 2006).

We define trust as an antecedent to reliant behavior, a willingness to accept vulnerability in expectation of a positive outcome. This definition is based on that used by Verberne et al. (2015), which in turn was derived from the definition by Mayer et al. (1995), who describe trust as "the willingness to be vulnerable to the actions of another party."

Lee and See (2004) discuss the need for appropriate reliance, concluding that this is a challenge for both the design of systems and the design of interface affordances. They identify three components of appropriate trust: calibration, resolution, and specificity. Proper calibration, an accurate knowledge of the system's capabilities, will be necessary for drivers to know when the system will be able to handle situations presented by the environment, and when human action will be required. In situations where there is little time to select a course of action, a proper calibration of one's trust, developed over time and through experience, will influence or determine one's actions in that situation.

Designing a system that encourages appropriate trust requires addressing a two-sided problem: promoting trust when justified, and encouraging intervention when necessary. Considering that in the near future automated driving systems will require human oversight and intervention in some situations, humans will need to act in a supervisory control role (Sheridan, 2006; Sheridan & Verplank, 1978). High uncertainty situations, where an automated system will have difficulty determining a course of action, are common in the automotive field (Park, Jenkins, & Jiang, 2008), and human action may be required in order to resolve the ambiguities successfully. Overtrust in automation may delay a driver taking control in a situation where human intervention is warranted, and thus appropriate calibration will be necessary to ensure safety.

Pioneering work by Reeves and Nass (1996), Nass and Moon (2000), and Nass et al. (1994) suggests that humans nonconsciously treat computers and robotic systems as humanlike entities, developing a relationship through interaction. As situations requiring trust in automation may occur in the domain of seconds, this ongoing relationship between the driver and computer needs to be considered in the design of highly automated systems' communications and actions. The behavior of the automated system is communicative by itself, with visual and audible channels along with environmental factors occupying a secondary role in trust model formation (Hancock et al., 2011; Ju, 2014).

People may say they would trust an automated system, yet act in a way that demonstrates that they do not trust it. Instruments such as the questionnaire by Jian et al. (2000) inquire as to one's beliefs in the system's capabilities and trustworthiness, but one's beliefs may not translate to behaviors. In slow-developing situations, one can make reasoned decisions as to whether the system should be trusted to act properly. In a situation demanding a rapid reaction, one must act quickly, with biases and training exerting strong influence over the action (Kahneman, 2011).

Much of the research into trust in automation has been performed in a third-person frame, where the participant has no perception of direct risk in case of automation failure (J. D. Lee, 1991; Muir & Moray, 1996). A first-person framing, where the participant's life is on the (simulated) line, explores trust in automation behaviorally. In a simulation where the driver perceives she or he may die if they improperly rely on the automation, the test of trust has more gravity, and this design may offer greater validity.

Studying Trust Behavior in Virtual Reality

In a highly immersive virtual reality environment, participants react as if the situations they are placed in similarly to the way they would act in real situations, and this effect has been used for social science research to good effect (Blascovich et al., 2002). The concept of presence, specifically telepresence, forms a foundation for considering virtual reality simulation research as comparable to real-world situations. (IJsselsteijn, de Ridder, Freeman, & Avons, 2000; K. M. Lee, 2004). Vehicle simulation has been shown to elicit similar reactions to on-road testing in critical situations (McGehee, Mazzae, & Baldwin, 2000), and the trust fall relies on the driver being highly present in the virtual environment, reacting as if threats to safety were real.

Self-reports of trust in automation measure one's beliefs in the trustworthiness of a system, but this does not necessarily map to one's behavior in situations requiring reliance on another agent. In a simulation, one reacts as if the situation were real, taking reflexive action before one can deliberate. In a real critical situation, one may not have sufficient time or cognitive resources (Young & Stanton, 2002) available to make a reasoned decision, thus systems need to be designed so to avoid drivers inadvertently defeating safety features.

The Trust Fall

The trust fall is a common team trust building exercise, where an individual falls backward into the arms of colleagues or teammates. The person taking the fall must invest trust in his or her teammates, risking potential injury to demonstrate that trust.

The metaphor of the trust fall is extended to a behavioral test of trust in automation: a situation is presented where people must trust a robotic system to act in a (simulated) safety critical situation. Based on orientation towards trust in automation and previous experience, participants must choose whether to trust the system, or to exert agency, behaviorally expressing their distrust.

The Trust Fall relies on first establishing a mental model of the system's capability or incapability to navigate non-critical situations without human input, and then testing whether participants continue to trust the system in a critical situation. It seems natural to exert agency when a system's failure is likely to have drastic consequences, but if the system appears trustworthy, one may not act.

METHODS

This study was conducted in a full-vehicle driving simulator from Realtime Technologies. The automated driving system could be enabled/disabled with steering wheel mounted buttons, or by placing a hand on the capacitive touch-sensitive steering wheel. The throttle and brake pedals would similarly override the automated control.

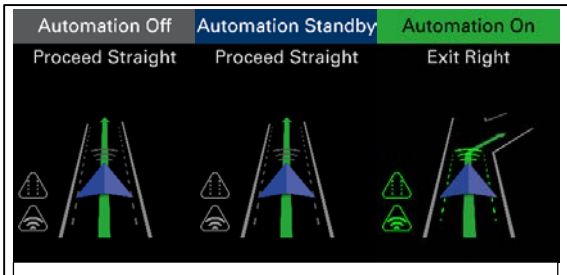
A 2x3 between-participants design was employed, with two levels of automation capability (low/high) and three levels of interfaces (navigation, perception, and planning). Forty-two participants, split evenly by gender (21 M, 21 F) ranging in age from 18 to 74 years (M=34.4 years, SD=14.0) were recruited from Stanford University and the surrounding area. Each participant was presented with a sequence of six challenges over the course of the experiment (see Table 1).

Automation Capability levels

Two levels of automation capability were tested, varying between participants: a high capability level that could handle more of the situations presented without driver intervention, and a low level of automation capability that required the participant to take control of the vehicle in more of the challenge situations (see Table 1).

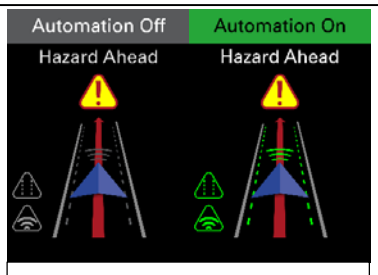
Interface Communication Levels

All three interface levels offered visual and voice navigation cues, and accurately informed the driver of the state of the automated driving system. The navigation interface displayed only navigation and system status; the perception interface added nonspecific warnings of hazards ahead; and the planning interface illustrated the future actions that the system would take, or that the driver should take (see Figure 1).



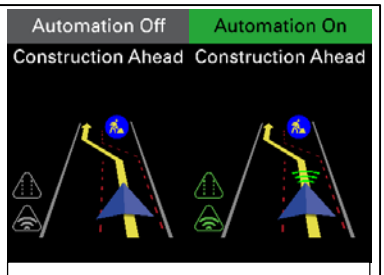
Navigation Interface

The navigation interface presented navigation cues and information about the state of the automated driving system (whether it was enabled and driving, disabled, or in standby mode).



Perception Interface

The Perception interface displayed automation state and navigation cues, and added a display that the vehicle was aware of potential hazards ahead.



Planning Interface

The planning interface showed awareness of the environment; and provided information about what the computer *would do* or what the driver *should do*, based on automation capability level.

Figure 1. Instrument panel display of automated system state, navigation, and the awareness of environmental features.

Figure 2. Road segment without lane markings. In the high-capability automation conditon, the system would continue to drive; in the low-capability automation condition, the driver woiuld have to take control.

Table 1. Experimental Challenges and System Responses

Challenge	Low Capability Automation	High Capability Automation
No lane markings	System shuts down delegating control to driver	System continues to drive, 'snakes' within lane
Construction zone/closed right lane	Driver must change lanes to left lane	Automated lane change to left lane
Highway exit	Driver must exit and follow navigation directions	Driver must exit and follow navigation directions
Red traffic light	Automatic stop, system shuts down	Automated stop, system resumes driving when light turns green
Police car on right shoulder	Driver should change lanes to left	Automated lane change to left lane
Car cut off	System will avoid collision	System will avoid collision

Testing Method

Participants were provided with a tablet computer on which to complete a questionnaire collecting demographic information and providing instructions describing the driving experience, information on the automated driving system and the instrumentation panel display. Participants brought the same tablet in to the simulator, and used it to answer questions after each event sequence, to avoid the nonconscious politeness effect (Nass & Moon, 2000).

The first stage of the trust fall is to establish a mental model relative to the trustworthiness of the automated system. The events in the testing sequence (see Table 1) were intended to help participants form a strong mental model regarding the capability level of the system. An example is shown in Figure 2, where the high capability system would continue to drive in the area without lane markings, while the low capability system would require the participant to take control of the vehicle.

The final challenge of the sequence, the "fall," featured a short-gap lane incursion, arranged such that it was impossible for the participant to collide with the cut-off car ahead (see Figure 3). If the driver does not take control, that is a strong vote of confidence in the automated driving system, indicating a high level of trust. During the driving component of the study, participants first drove the simulated vehicle to establish a model of vehicle behavior, and then were instructed to enable automated driving, actively supervising the automated system. Drivers were required to take control when the automated system shut down, or could take control if they did not trust the system.

Figure 3. Participant reaction to lane incursion.

After each challenge and subsequent return to automated driving, participants were verbally prompted to complete the survey on the tablet, which asked a single question: "What did you think the computer would do in this situation?" The question had two possible answers: "I expected the vehicle to handle the situation independently." and "I would have to take control of the vehicle." This prediction, influenced by the participant's having seen the episode having unfolded, is necessarily contaminated by the actions taken by the computer and by the participant's actions. This statement can be compared with the participants' actions to gain insight into the disparity between their thoughts and actions. Following the driving component of the study, participants completed a questionnaire which included measures of trust adapted from Jian et al.'s inventory.

RESULTS

Driver behavior varied considerably across all events in the sequence. Many drivers did not trust the automated driving even in cases where it would perform flawlessly. This proved a hindrance, as it interfered with the formation of the mental model of system capability.

Driver Behavior and Expected Behavior Questions

Comparing participants' expectations of the automated driving system's behavior to their own inputs did not yield significant correspondences between expectations and actions. There were no statistically significant differences across automation levels or interface conditions. This is in itself important: if a driver cannot predict accurately what an automated system will do a few seconds into the future, and does not respond in a way that is appropriate, disaster can result.

Analyzing the predictions participants made compared with their behavior shows a bias towards conservatism: drivers will intervene even when they say they expected the computer to have handled the situation, but few did not intervene when they felt the computer expected them to take action. For all conditions combined, this was statistically significant, $\chi^2(1)=3.953$, $p=.047$, but the individual conditions did not show statically significant differences at the $p<.05$ level. Comparing all participants' behavior and predictions across the automation capability dimension,

more participants in the high capability condition stated that they expected the computer to not act (17/21), but they held back and trusted the computer to act (13/21), irrespective of their stated prediction. This indicates that the participants behaved by trusting the car, even though they stated they did not expect the computer to act and thus would need to act themselves (see Table 2). This split between self-reported trust and trusting behavior validates the concept of the trust fall as a measure of trust that is independent of self-reported trust in automation.

Table 2. Car Cut-Off Expectations and Actions

Condition			Driver Action: No Action	Driver Action: Action	Total
Low-Capability Automation	I expected the computer to act	No	2	6	8
		Yes	2	11	13
	Total		4	17	21
High-Capability Automation	I expected the computer to act	No	8	9	17
		Yes	0	4	4
	Total		8	13	21
All Participants	I expected the computer to act	No	10	15	25
		Yes	2	15	17
	Total		12	30	42

Driver Attention Focus

Using eyetracking to observe driver behavior during the event sequences showed that drivers looked to the ambient environment first, as they had to establish situation awareness and react to the situation requiring their input. As a result, the head-down display on the instrument panel was consulted as a supplement to the environmental cues, rather than as the source of information on the status of the environment and the automated system's reaction to situations on the road. The audible signals in the Perception and Planning level interfaces provided a cue to participants that there was a change in the state of the automated driving system, and thus made them aware of a need to assess the situation around them. While the beep was an effective general alert, more specialized alert using speech may be more effective at aiding drivers to establish situation awareness and take appropriate actions.

Self-Reported Trust in Automation

As part of the post-drive questionnaire, a set of questions derived from the set developed by Jian et al. was asked of participants. The set was repeated three times, participants being asked to first assess the automated driving system's ability to sense the ambient environment, then its ability to make decisions, and finally its ability to carry out vehicle control actions. Each of the question sets was independently assessed by factor analysis, and indices comprised of the highly correlated items were developed by summing the items. The three indices are highly reliable, each component showing a Cronbach's α over .9. The three indices are highly correlated, but do not show statistically significant differences between interface displays or automation capability levels (see Figure 4).

Discussion

Observing the drivers' actions, specifically that they did not rely on the instrumentation to a high degree, illustrates the need for nonvisual and possibly non-auditory interfaces—and the need for research into interface design for highly automated vehicles. Considering the need for drivers to be engaged in supervision of near-future

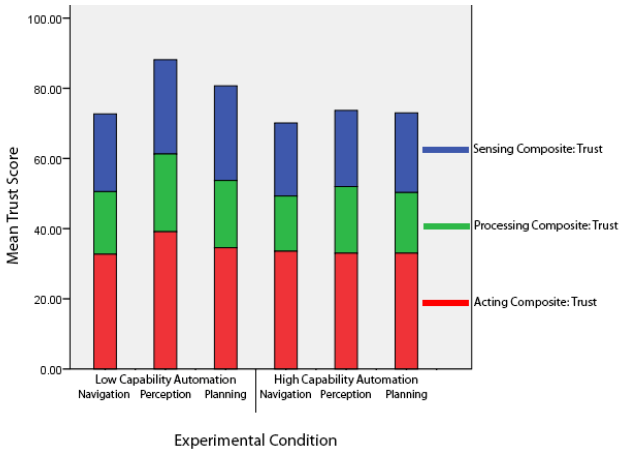


Figure 4. Self-reported trust in automation, broken down into sensing (blue), processing (green), and acting (red).

automated driving systems, and to be able to take effective control within a short time, with further future automated systems, appropriate trust will be a serious issue.

This study is an experimental trial of the trust fall method, where the mental model of the participant is in part shaped by the behavior of the automation. While the overall trust level as reported by participants did not differ as a function of automation capability level or interface affordances, there is evidence for the value of testing trust using a behavioral measure, in addition to questionnaire measures. Using a first-person frame, where the participant feels a (simulated) threat to self is likely to elicit a true response, which may contrast with what one says or believes that she or he would do in that situation. Further study will be necessary to better assess the disparity between instinctive actions in situations requiring trust and reasoned actions in analogous situations.

That 71% of participants (30/42) took evasive action when cut off in the final test of trust in the automated system is unsurprising; this illustrates the need for automated safety systems to employ mechanisms to avoid interference by drivers in safety critical situations. If a driver's action will decrease safety or lead to a collision, the system needs to override or ignore driver inputs, but also allow driver inputs to override its actions when it is possible the automated system is acting in error. This will necessitate the development of a meta-trust mechanism to assess the trustworthiness of the driver's actions, in addition to an assessment of the system's trust in its own sensing, processing, and behavior abilities. While difficult to implement, this meta-trust mechanism may be necessary for achieving optimal safety in a human-machine system such as an automobile in the hands of a relatively untrained driver.

CONCLUSIONS

Overtrust in systems that are commonly reliable but prone to rare, unpredictable, and hazardous failures can present a significant danger. In the case of highly automated vehicles, such situations that may cause the system to transfer control, or where the driver may have to take control, are likely to be rare and unpredictable, presenting serious safety risks.

For high levels of vehicle automation to be successful (e.g. automated driving without constant supervision), systems may have to feature significant adaptive capabilities, tailoring their behavior to the abilities of the driver. This may be necessary to adjust for driver states, such as fatigue or impairment, or to compensate for inexperience or age-related faculty decline. This adaptive automation can be considered a joint-cognitive system (Woods, 1985), requiring sophisticated computing and in-vehicle sensing in order to determine driver state, in addition to developing a knowledge of the driver's capabilities over time. Using the driver as a sensing and decision making element of a driver-vehicle system (Miller & Ju, 2015), with automation acting as an aid to the driver creates a joint cognitive system which can greatly enhance overall safety and performance.

ACKNOWLEDGEMENTS

This research was supported by the Toyota CSRC, and we thank Jim Foley, Josh Domeyer, and Larry Cathey.

REFERENCES

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive Virtual Environment Technology as a Methodological Tool for Social Psychology. *Psychological Inquiry*, *13*(2), 103–124.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., Visser, E. J. de, & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 517–527.

IJsselsteijn, W. A., de Ridder, H., Freeman, J., & Avons, S. E. (2000). Presence: concept, determinants, and measurement (Vol. 3959, pp. 520–529).

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.

Johnson-Laird, P. N. (1980). Mental Models in Cognitive Science. *Cognitive Science*, *4*(1), 71–115.

Ju, W. (2014). *Design of Implicit Interactions.* San Rafael: Morgan & Claypool.

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed). New York: Farrar, Straus and Giroux.

Lee, J. D. (1991). The Dynamics of Trust in a Supervisory Control Simulation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *35*(17), 1228–1232.

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80.

Lee, K. M. (2004). Presence, Explicated. *Communication Theory*, *14*(1), 27–50.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709.

McGehee, D. V., Mazzae, E. N., & Baldwin, G. H. S. (2000). Driver Reaction Time in Crash Avoidance Research: Validation of a Driving Simulator Study on a Test Track. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*(20), 3-320-3–323.

Miller, D. B., & Ju, W. (2015). Joint Cognition in Automated Driving: Combining Human and Machine Intelligence to Address Novel Problems. In *2015 AAAI Spring Symposium Series*.

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

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

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72–78). New York, NY, USA: ACM.

Norman, D. A. (1983). Some observations on mental models. *Mental Models*, *7*(112), 7–14.

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.

Park, E., Jenkins, Q., & Jiang, X. (2008). Measuring Trust of Human Operators in New Generation Rescue Robots. *Proceedings of the JFPS International Symposium on Fluid Power*, *2008*(7–2), 489–492.

Reeves, B., & Nass, C. I. (1996). *The media equation: how people treat computers, television, and new media like real people and places*. Stanford, Calif.; New York: CSLI Publications; Cambridge University Press.

Sheridan, T. B. (2006). Supervisory Control. *Handbook of Human Factors and Ergonomics* (pp. 1025–1052). John Wiley & Sons, Inc. Retrieved from

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

Verberne, F. M. F., Ham, J., & Midden, C. J. H. (2015). Trusting a Virtual Driver That Looks, Acts, and Thinks Like You. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 18720815580749.

Woods, D. D. (1985). Cognitive Technologies: The Design of Joint Human-Machine Cognitive Systems. *AI Magazine*, *6*(4), 86.

Young, M. S., & Stanton, N. A. (2002). Malleable Attentional Resources Theory: A New Explanation for the Effects of Mental Underload on Performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(3), 365–375.