The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction, and System Transparency

Johannes Kraus, **David Scholz**, **Dina Stiegemeier**, and **Martin Baumann**,
Ulm University, Germany

Objective: This paper presents a theoretical model and two simulator studies on the psychological processes during early trust calibration in automated vehicles.

Background: The positive outcomes of automation can only reach their full potential if a calibrated level of trust is achieved. In this process, information on system capabilities and limitations plays a crucial role.

Method: In two simulator experiments, trust was repeatedly measured during an automated drive. In Study 1, all participants in a two-group experiment experienced a system-initiated take-over, and the occurrence of a system malfunction was manipulated. In Study 2 in a 2 × 2 between-subject design, system transparency was manipulated as an additional factor.

Results: Trust was found to increase during the first interactions progressively. In Study 1, take-overs led to a temporary decrease in trust, as did malfunctions in both studies. Interestingly, trust was reestablished in the course of interaction for take-overs and malfunctions. In Study 2, the high transparency condition did not show a temporary decline in trust after a malfunction.

Conclusion: Trust is calibrated along provided information prior to and during the initial drive with an automated vehicle. The experience of take-overs and malfunctions leads to a temporary decline in trust that was recovered in the course of error-free interaction. The temporary decrease can be prevented by providing transparent information prior to system interaction.

Application: Transparency, also about potential limitations of the system, plays an important role in this process and should be considered in the design of tutorials and human-machine interaction (HMI) concepts of automated vehicles.

Keywords: trust in automation, compliance and reliance, human-automation interaction, function allocation, trust formation

Address correspondence to Johannes Kraus, Department of Human Factors, Ulm University, Albert-Einstein-Allee 41, 89081 Ulm, Germany; e-mail: johannes.kraus@uni-ulm.de.

HUMAN FACTORS
Vol. 62, No. 5, August 2020, pp. 718–736
DOI: 10.1177/0018720819853686
Article reuse guidelines: sagepub.com/journals-permissions

During the transition from manual to fully automated driving, concepts at the Society of Automotive Engineers (SAE; SAE International, 2014) automation levels 3 (conditional automation) and 4 (high automation) call for "intermediate, coordinative modes of interaction, which allow human operators to focus the power of the automation on particular sub-problems" (Woods, 2001, p. 3). In this time of highly—but not fully—automated driving, the automation is able to take over control in those situations that it is able to handle (e.g., driving on the motorway)—restricted by system limitations and system malfunctions. A *system limitation* is reached if road type or conditions change into an environment the system has not been designed for. In this case, control is handed back to the driver with a take-over request (TOR; Gold, Damböck, Lorenz, & Bengler, 2013). In contrast to system limitations, *malfunctions* are sudden, unpredicted errors related to a system's reliability within its area of application. These properties of highly automated driving make the task of supervising and monitoring complex and challenging (Parasuraman, Sheridan, & Wickens, 2000).

To enable drivers to use the system in an appropriate way, they have to gain an understanding of the system's capabilities and shortcomings through available information prior to and during system interaction (Van den Beukel, van der Voort, & Eger, 2016). In this regard, an appropriate usage pattern corresponds to a calibrated level of trust: a situation in which the level of trust is exactly reflecting a system's capabilities and its actual performance (Muir, 1987). Although there is a growing body of work in the area of trust in automation (i.e., Forster, Kraus, Feinauer, & Baumann, 2018; Hartwich,

Witzlack, Beggiato, & Krems, 2018; Körber, Baseler, & Bengler, 2018; Wintersberger, von Sawitzky, Frison, & Riener, 2017; Yang, Unhelkar, Li, & Shah, 2017) up until now, the questions of how trust is built up, how it dynamically changes during system use, and which variables influence this process are not fully understood (e.g., Lee & See, 2004; Walker, Stanton, & Salmon, 2016). At the same time, an understanding of these processes seems important, as calibrated trust may serve as a benchmark for safe and efficient design of driving automation and the associated interaction strategies in terms of information provided prior to (tutorials, user guides, marketing, and training) and during system use (human-machine interface [HMI] concepts).

This paper presents a psychological model for dynamic trust calibration and two simulator studies investigating the dynamics of trust during early system use under error-free conditions and in the case of TORs and system malfunction. In addition, the effect of increased system transparency is investigated. The presented studies provide a cohesive investigation of early trust development over time in a controlled driving simulator experiment.

TRUST IN HIGHLY AUTOMATED VEHICLES

In Lee and See (2004), trust in automated technology is conceptualized as "the attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" (p. 51). As trust builds the basis of the decision of when to make use of the driving automation and when to drive manually, it is crucial that the level of trust corresponds with the actual capabilities of the automation. Otherwise, the interaction with automated vehicles will neither fully benefit from the advantages associated with automated driving (*distrust*) nor will it be safe for the driver or other road users (*overtrust*; Lee & See, 2004). Exemplary, in case of system distrust, drivers would maintain control of the vehicle although the automation could overall realize a more efficient and safer driving.

TRUST DYNAMICS AND TRUST CALIBRATION

Trust is a mediating variable between system properties and a user's allocation decisions (Muir & Moray, 1996). It evolves and adapts over time along the accumulating knowledge of the user about a system (*trust calibration*, Lee & See, 2004; Parasuraman & Miller, 2004). While in the beginning of the interaction—especially with new technologies—this assessment is vague, unstable, and prone to relatively fast dynamic changes as a function of new information (Merritt & Ilgen, 2008), it stabilizes in the course of interaction. From a normative perspective, calibrated trust describes the optimal result of this process as a condition in which a user's trust in a system corresponds exactly with its objective capabilities and its actual performance (Muir, 1987). This situation is reflected by an absence of both disuse and misuse of the system (Parasuraman & Riley, 1997). At this point, research on the psychological processes leading to trust calibration is rather scarce, but two theoretical frameworks provide direction.

Trust and reliance are established in a dynamic interaction of characteristics of the user, the system, and the situation in which the system is used (Lee & See, 2004). In their conceptual model of trust in automation, Lee and See (2004) integrated concepts of trust from different domains with the theory of planned behavior (Ajzen, 1991; Ajzen & Fishbein, 1980; Fishbein & Ajzen, 1975). The authors conceive of trust as an attitude that affects behavioral decisions with automated systems in a closed-loop dynamic decision-making process, which is moderated by variables in the environment of interaction, properties of the automation itself and its interface, as well as by user characteristics. Although the model constitutes a step toward a well-founded psychological theory of the emergence of trust in automation, the model still needs empirical validation and further specification to explain how trust is established and which information is used for trust calibration.

Recently, Hoff and Bashir (2015) proposed four layers of trust that are formed prior to and during the interaction with an automated system.

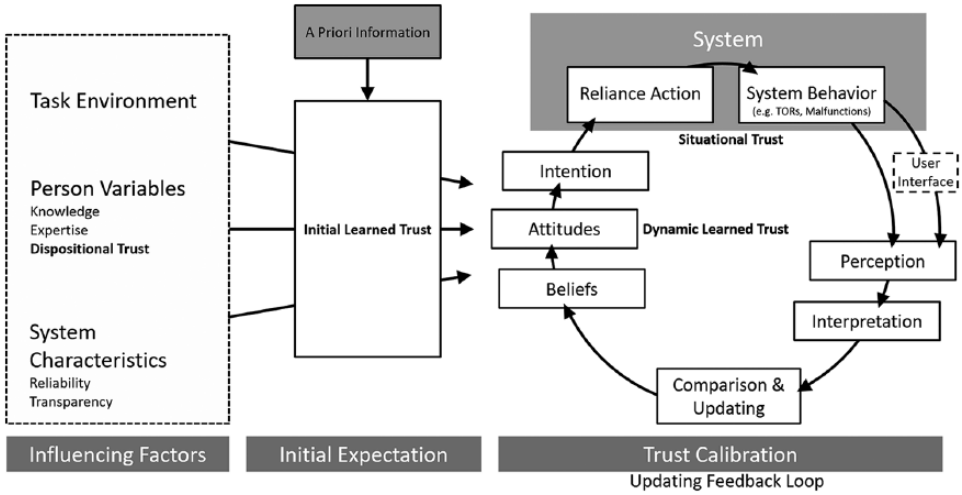

Figure 1. Model of dynamic trust calibration integrating different stages of trust attitude formation.
Note. Different trust forms reflect attitudes at different stages of this process. Initial learned trust is built up before the actual interaction with a system along the available information. In the course of interaction, dynamic learned trust is steadily updated in a dynamic calibration feedback loop. In this process, interpretation of system output is influenced by user personality and current beliefs and attitudes about the system. The model integrates assumptions of both Lee and See (2004) and Hoff and Bashir (2015).

Dispositional trust is conceptualized as a general personality trait determining the initial level of a person's trust. On a more specific level, initial learned trust represents an evaluation of the system from knowledge prior to actual system use, and dynamic learned trust is established by the sum of experiences with a specific system during the interaction. In addition, situational trust describes trust as a combination of interaction context and operator state. This theory also awaits empirical validation. At this point, both theories can serve as sources for drawing hypotheses for trust research, which in turn may support their implications.

Figure 1 presents a model of dynamic trust calibration prior to and during the interaction with an automated system. This model synthesizes some central assumptions of the two models by Lee and See (2004) and Hoff and Bashir (2015) and advances these with several specifications. In addition, it includes system, environmental, and individual factors influencing trust establishment (Hancock et al., 2011).

A first key implication of the model is the prediction that trust is calibrated in accordance with information provided both prior to and during the interaction with an automated system (see also, for example, Lee & Moray, 1992). Prior to the interaction, available information (e.g., marketing) guides the establishment of general expectations toward system functioning (e.g., initial learned trust). During the interaction, the user relies on real-time information about the automation's performance (experienced or displayed; Hoff & Bashir, 2015). Second, the model predicts that this continuous balancing of expectations and perceived system performance takes place in an updating feedback loop establishing dynamic learned trust in the system (Lee & See, 2004). Hereby, both the behavior of the system and the information from the user interface are integrated in a real-time assessment of the capabilities of the automation in terms of trust. In this process, all information and all stages are predicted to be influenced by prior knowledge, personality, and current beliefs and attitudes. Third, the model predicts that in the early phase of interaction, beliefs build the basis of trust. Following Lee and See (2004), beliefs can be defined as subjectively attributed

system characteristics based on the information available about the object under consideration (Ajzen & Fishbein, 1980). Relevant beliefs for trust calibration include perceived predictability, dependability, faith, competence, and reliability of the system (Muir, 1994; Muir & Moray, 1996). The depicted model serves as a framework for the design and hypotheses of the reported studies.

TRUST DEVELOPMENT IN HIGHLY AUTOMATED DRIVING

Early studies measuring trust in automated process control microworlds show that trust increases over time in the course of error-free interaction (e.g., Moray & Inagaki, 1999; Muir & Moray, 1996). A similar trust increase was found in studies assessing trust in adaptive cruise control (ACC) during the initial days (Kazi, Stanton, Walker, & Young, 2007) and weeks (Beggiato & Krems, 2013) of system interaction. Preliminary evidence for a similar trust development during the first encounter with highly automated driving stems from studies with repeated single-item trust measures (Hergeth, Lorenz, & Krems, 2017; Hergeth, Lorenz, Vilimek, & Krems, 2016) and a pre-post measurement with a trust scale (Hergeth, Lorenz, Krems, & Toenert, 2015), showing an average increase in trust from before and after system interaction.

The research presented adds to these findings by assessing trust repeatedly over time in the early interaction with a highly automated vehicle. Based on the assumptions of the depicted trust calibration model's feedback loop, it is hypothesized that the experience of a system that works reliably leads to a gradual increase in trust over time.

Hypothesis 1 (H1): If an automated system works without malfunctions, trust increases over the course of system interaction.

SYSTEM LIMITATIONS AND TORS

In highly automated driving, the automation can only perform the driving task in an array of predefined situations, and a foreknown approach to system limitations leads to a TOR (e.g., Gold, Körber, Lechner, & Bengler, 2016; Walch et al., 2017). In line with the "perfect automation schema," users expect an automation to work nearly flawlessly (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Merritt, Unnerstall, Lee, & Huber, 2015). In the same vein, it can be hypothesized that an automation that has limitations falls short of that expectation and should lead to a temporal reduction in trust in an automated system (e.g., Dzindolet et al., 2003; Madhavan & Wiegmann, 2007). For a single-item trust measure, such a temporal reduction of trust subsequent to TORs was shown in the work of Hergeth and colleagues (2015). Our studies add to these findings and investigate the real-time effects of a TOR in highly automated driving on trust assessed with a trust scale. In line with the information feedback loop of the proposed model, we hypothesize that an experience of a TOR is associated with the perception of reduced system reliability which results in trust impairment, subsequently reestablished with an experience of sustained safety and subsequent system reliability.

Hypothesis 2 (H2): Trust decreases temporarily after a take-over of manual control in face of a system limitation.

How TORs are perceived after the experience of a system malfunction is as yet unanswered. In the current research, we investigated whether a second TOR affects trust development differently with the prior experience of an error-free system as compared to an experience with the occurrence of a system malfunction.

MALFUNCTIONS

Although system limitations are part of the design of a system and refer to its scope of application, system malfunctions include mechanical failures, errors in data acquisition, hazardous weather conditions, and hardware failures (Emzivat, Ibanez-Gutman, Martinet, & Roux, 2017; Molina et al., 2017). Thus, while system limitations are preceded by a TOR of the system and a controlled transition of control, system malfunctions occur suddenly and prevent further safe functioning of the automation. To protect the vehicle from collision, excessive speed, or

other hazardous conditions, such as a malfunction, is resolved through a fallback strategy, for example, leading to a safe halt in SAE level 4 (Emzivat et al., 2017; Molina et al., 2017).

As a prediction from the depicted trust calibration feedback loop, if a driver experiences a malfunction blindsided, the automated system should be perceived as less reliable and competent than expected, and trust should decline as an initial reaction (Madhavan & Wiegmann, 2007; Moray & Inagaki, 1999). In line with this reasoning, a number of studies show an immediate decrease in trust following a system malfunction in different domains (e.g., Dzindolet et al., 2003 [military decision aid]; Lacson, Wiegmann, & Madhavan, 2005 [decision aid]; Wiegmann, Rich, & Zhang, 2001 [diagnostic aid]; Lee & Moray, 1992 [plant simulation]). If a malfunction did not corrupt trust to a degree that hinders further system usage, prolonged error-free operation was found to lead to remediation of trust in automated plant simulations (Lee & Moray, 1992) and automated robots (Desai, Kaniarasu, Medvedev, Steinfeld, & Yanco, 2013) over time. As whether this reestablishment of trust (trust recovery) also takes place in highly automated driving remains an open question, we hypothesize that also in this context a decrease in trust occurs after a system malfunction which is recovered in subsequent system use. Although system malfunctions have traditionally been investigated as a hindrance for trust development, they may also be associated with a positive effect on long-term trust calibration (e.g., Helldin, Falkman, Riveiro, & Davidsson, 2013). In line with the hypothesized feedback loop of the dynamic trust model, a malfunction may enable a driver to collect new information and reduce some uncertainties and potential anxieties in regard to system functioning. In this sense, the availability of new information may even lead to a long-term increase in trust. Accordingly, we hypothesize that in the case of a low-risk malfunction, trust is recovered to the level before the malfunction.

Hypothesis 3 (H3): An unexpected malfunction of the system leads to a significant temporary decrease in trust in the automated system (H3.1). Trust recovers in case of continued error-free system operation (H3.2).

TRANSPARENCY

The concept of transparency can be defined as the extent to which the inner processes and decisions of an automation are made accessible (Seong & Bisantz, 2008). According to Chen and colleagues (2014), transparency enables a user's comprehension of intention, performance, future actions, and reasoning of automated processes. Information fostering system transparency in advance or during the drive should facilitate trust calibration (Muir, 1987). In line with this, it was shown in empirical studies that continuous information about system reliability favors an establishment of appropriate levels of trust, especially in the context of system malfunctions (in automated decision aids: Dzindolet et al., 2003; Wang, Jamieson, & Hollands, 2009; in driving automation: Beggiato, Pereira, Petzoldt, & Krems, 2015; Ekman, Johansson, & Sochor, 2016). In addition, Körber and colleagues (2018) showed that prior information on the trustworthiness of driving automation affected trust in the system in subsequent interaction.

At this point, it is unclear how information on the character and consequences of malfunctions of an automated system provided prior to system use influences trust calibration in automated driving. On one hand, one could argue that providing information about system shortcomings may lead to an initial trust reduction and a potential obstacle to using the system at all. On the other hand, following the reasoning of the depicted trust calibration model, a priori transparency information may serve as a safeguard against trust reduction in the face of an actual malfunction. This can be assumed because the initial expectation of the system functioning is no longer violated in this case. Based on this reasoning and the reported research, we assume that a malfunction of the automation does not lead to a trust reduction if information on the reasons for the malfunction are made transparent and the consequences of the safe mode are experienced prior to system use.

Hypothesis 4 (H4): Higher system transparency (providing a priori information about reasons for system malfunctions and the character of the safe mode) leads to less trust reduction subsequent to a malfunction of an automated system.

Following the mediation processes specified in the trust calibration model (qua Lee & See, 2004), beliefs about a system build the basis of trust and should be strongly influenced both by information prior to system interaction (e.g., advertisement) and the experiences of the system behavior during the drive. Thus, in the early phase of interaction, beliefs about a system should change in the face of new information and experiences with a certain system and provide the basis for trust calibration as assumed in the proposed model.

Hypothesis 5 (H5): Beliefs show a significant correlation with trust.

The presented hypotheses were tested in two experimental driving simulator studies. In Study 1, the nature of initial trust development during a drive with and without system malfunction was investigated to test H1-3 and H5. In a one factorial mixed design, at the halfway point of a drive on the motorway, one study group experienced a system malfunction, while in the second group, the system worked flawlessly. In both groups, two TORs were included prior to and after the system malfunction. Trust was measured repeatedly. In Study 2, in a similar paradigm, two additional study groups were introduced to investigate the effect of system transparency on early trust development. Thus, Study 2 provides a further investigation of the findings on trust dynamics (H1, 3, 5) and additionally investigates H4.

STUDY 1

In this study, trust was measured over time during the initial interaction with an automated vehicle to identify the character of early trust development during error-free functioning and in the case of a system malfunction. Furthermore, the effect of a TOR in case of an error-free system was assessed. In addition, the effect of a second TOR—either after further error-free functioning or after a malfunction—was explored.

Method

In a driving simulator experiment, participants drove on a highway with a highly automated vehicle capable of overtaking automatically (SAE level 3). The study was conducted in a dual task paradigm to provide a natural setup and to reduce possible bias from boredom. As a primary task, participants were instructed to drive safely. As a secondary task, they played a game on a tablet PC. In a one factorial experimental design, the independent variable system malfunction was manipulated, and the dependent variables trust, reliability, predictability, and competence were measured repeatedly using self-report questionnaires. Participants were randomly assigned to one of the two study groups. While in one group, no system malfunctions occurred (MF–), the second group experienced a malfunction during the drive (MF+). This research complied with the tenets of the Declaration of Helsinki. Informed consent was obtained from each participant.

Sample. Participants were recruited at Ulm University and had to hold a driver's license. For their participation, they received 8€ or study credit. The study sample consisted of 31 participants (18 female, 13 male) with an average age of $M = 23.84$ ($SD = 3.66$) years. They had held their driver's license for the average of $M = 5.81$ ($SD = 2.55$) years and rated their experience with any automated driving assistance assessed with a single-item measure on a 7-point Likert-type scale, with $M = 2.58$ ($SD = 2.08$), as comparatively low.

Procedure. The study procedure is illustrated in Figure 2. Altogether, the study took 75 minutes to complete. After being welcomed, participants filled out informed consent sheets. Then they were introduced to the experimental task and the driving automation. Participants were instructed that the system has two modes (manual and automated) and that the automated mode allows automated driving under certain conditions on the motorway. After this, they filled out a questionnaire with the dependent variables prior to any interaction with the

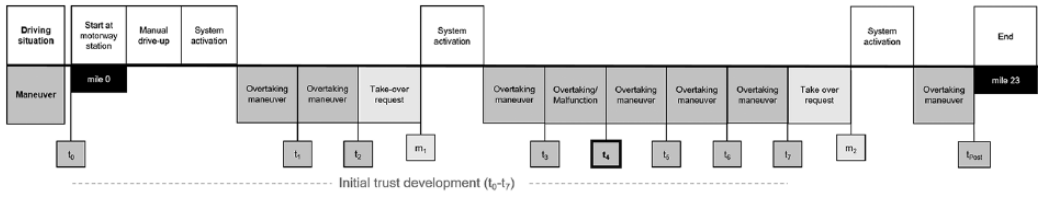

Figure 2. Experimental drive of Study 1.

Figure 3. Driving simulator of the Human Factors Department at Ulm University.
Note. Photo by: H. Grandel/Ulm University.

highly automated vehicle (t_{pre}). Participants went through two practice trials to familiarize themselves with manual and automated driving. In the latter, participants experienced autonomous overtaking, a TOR and a low-consequence system malfunction. When the malfunction occurred, the system provided a visual and auditory signal, activated the right indicator, reduced speed, and stopped on the side-strip (safe mode). No information on the reasons for this system behavior was provided. Subsequently, participants drove manually until the automation indicated that it was available again. After a baseline assessment (t_0) of the dependent variables, participants were introduced to the secondary task—a Tetris game (Electronic Arts, 2013) on a tablet computer. For the experimental drive, participants were told that driving safely (accident-free driving in compliance with traffic rules) is the primary goal and that they should try to achieve a high score in the secondary task at the same time to win a prize. It was clearly explained that engagement in the secondary task is only allowed during automated driving, while the driver is responsible for system supervision. Accordingly, the Tetris score would be set to zero in case of an accident or traffic violations. They were further instructed that control from automation can be taken over at all times and that safety is the highest priority of the system. After the instruction, the experimental scenario was started with a total track length of 23 miles.

In the experimental drive, the simulation started in manual mode before the automated mode was activated, and participants experienced seven automated overtaking maneuvers, after each of which the simulation was paused and a questionnaire was presented (t_1 to t_7; Figure 2). Participants also experienced a TOR with a subsequent short phase of manual driving, after which the dependent variables were assessed again (m_1). Before t_4, the MF+ group experienced a malfunction of the system with a subsequent safe mode and a short phase of manual driving, while the MF– group experienced a normal autonomous overtake. After the experimental drive, an additional TOR (before m_2) was implemented to gain more understanding of the trust implications of a repeated TOR. After an additional automated take-over of the system, trust was assessed a last time (t_{post}). After finishing the course, participants filled out a demographic questionnaire, received their reward, and were debriefed, thanked, and dismissed. Trust development over time is reflected in the interval t_0-t_7. t_{pre} constituted a general pre-measurement of trust before system introduction and t_{post} assessed effects of the second TOR on trust.

Material. Apparatus. The study was conducted in the driving simulator of the Human Factors Department at Ulm University (see Figure 3), which is assembled by three 1920 × 1200px video

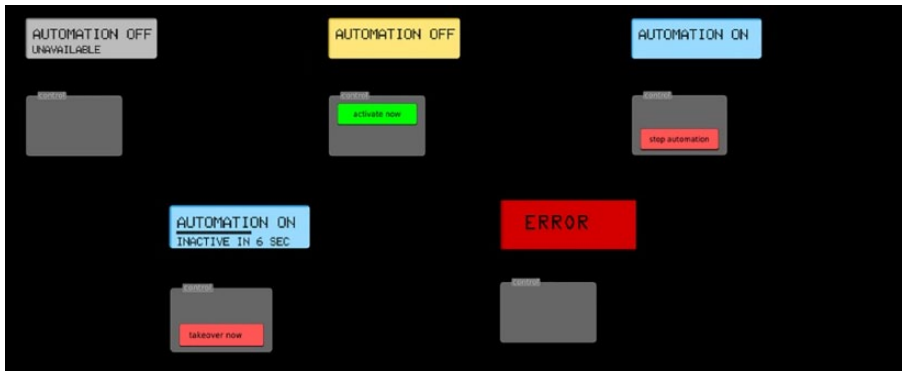

Figure 4. Interface for enabling and disabling the automation located in the center console.
Note. Each display consisted of information on the automation status in the upper part and a dynamic control panel with touch icons to turn the automation on or off in the lower part. In the upper row, the different states of normal functioning are depicted (from left to right: automation off/no automation available, automation off/automation available, automation on). In the lower row on the left, a TOR and on the right, the display for a system malfunction is presented. This simple and reduced interface design was chosen to minimize UI-related distractions.

projections onto three screens of 3.3 m × 2.1 m with a 200° viewing angle. Participants sit in a realistic vehicle mock-up. Rear view is implemented by two side screens and one back screen with 7″ displays (800 × 400px; 16:9). The highway simulation was programmed in SILAB 5.1 (Würzburg Institute for Traffic Sciences GmbH, 2014). A center display contained information on the automation as well as the possibility to engage and turn off the system (see Figure 4).

Questionnaires. Short scales were used to reduce frustration of the participants in face of repeated measurements. All questions were to be answered with a 7-point Likert-type scale (1 = *I do not agree at all*; 7 = *I totally agree*). Trust was measured with a shortened and translated German version of the questionnaire by Jian, Bisantz, and Drury (2000) aiming at assessing trust as a unidimensional construct. At critical times (t_{pre}, t_0, t_4, and t_{post}), a seven-item version, at the remaining times, a five-item version was used (items for the short version were selected on the basis of factor loadings from earlier studies). Perceived reliability (5 items) and predictability (4 items) were measured with translated versions of the questionnaires by Madsen and Gregor (2000). Competence was assessed with a translation of the questionnaire by Gong (2008; 6 items). Internal consistency was good for the trust scale with $\alpha > .83$ at every point of measurement and acceptable to good for perceived reliability, competence and predictability with $\alpha > .70$ (George & Mallery, 2003). The assessed variables at the different points of measurement are depicted in Table 1 along with their reliabilities.

Analysis. Data analysis was conducted with IBM SPSS Statistics (version 24; IBM Corp., 2016). Descriptives for all dependent variables are provided in Table 2 (also see Figure 5 for trust development over time in the different study groups). The complete sample reported an initial trust of $M = 4.59$ ($SD = 1.00$) at t_{pre} and $M = 4.96$ ($SD = 1.21$) at t_0. The groups did not differ significantly in any of the socio-demographic variables (age, gender, driving license, and experience). At both times of trust measurement prior to the experimental drive (t_{pre}, t_0), two-tailed *t*-tests showed no significant group differences (t_{pre}: $t(29) = 0.08$, $p = .937$, $d = -0.03$; t_0: $t(29) = -0.36$, $p = .720$, $d = 0.13$).

H1 was tested with a repeated-measure analysis of variance (ANOVA) with absolute trust ratings (t_0-t_7) as dependent variables. As H2-H4 predicted specific shapes of trust development over time, polynomial contrast analyses were used (e.g., Bühner & Ziegler, 2009; Rosenthal & Rosnow, 1985) investigating linear, quadratic, and cubic trends. A linear trend indicates a constantly increasing (decreasing) mean. A quadratic

TABLE 1: Internal Consistency of the Scales at All Times of Measurement Indicated With Cronbach's α

Variable	t_{pre}	t_0	t_1	t_2	m_1	t_3	t_4	t_5	t_6	t_7	m_2	t_{Post}
Trust	0.88	0.94	0.92	0.92	0.89	0.86	0.88	0.83	0.83	0.84	0.90	0.86
Predictability	0.76	0.73	—	—	—	—	0.83	—	—	—	—	0.82
Competence	0.80	0.91	—	—	—	—	0.94	—	—	—	—	0.95
Reliability	0.75	0.80	—	—	—	—	0.70	—	—	—	—	0.88

Note. The row "Trust" shows Cronbach's α, with all items of the respective scale. For trust at t_{pre}, t_0, t_4, and t_{post} the long version was used, and for the remaining points of measurement the short version was used.

TABLE 2: Descriptive Values of Trust of Both Groups at Each Measurement Point

	M_{tn} (SD_{tn})											
	t_{pre}	t_0	t_1	t_2	m_1	t_3	t_4	t_5	t_6	t_7	m_2	t_{post}
MF–	4.61	4.88	5.09	5.51	5.21	5.33	5.22	5.49	5.64	5.72	5.12	5.02
	(1.09)	(1.37)	(1.37)	(1.26)	(1.34)	(1.17)	(1.35)	(1.14)	(1.07)	(1.06)	(1.49)	(1.35)
MF+	4.58	5.04	5.09	5.53	5.35	5.69	5.06	5.74	5.84	5.99	5.80	5.50
	(0.95)	(1.07)	(0.90)	(0.95)	(0.95)	(0.88)	(0.81)	(0.76)	(0.78)	(0.88)	(1.07)	(0.74)

Note. t's indicate the different times of measurement during automated driving, *m*'s indicate the times of measurement in manual driving after a take-over request occurred. MF– = no malfunction; MF+ = with malfunction.

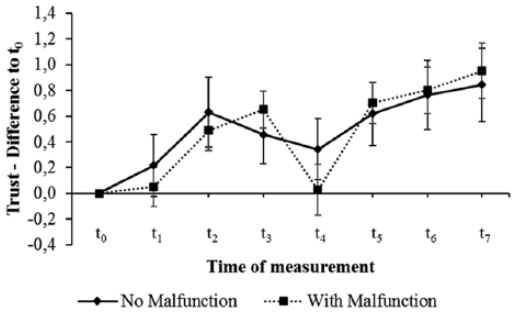

Figure 5. Trust development over the course of the experiment in the groups with and without malfunction.
Note. Before t_4 the malfunction occurred in the respective condition. t_0 serves as a baseline to facilitate comparison of the trust curves. Error bars indicate ±1 *SE*.

trend suggests that the means for the first and last measurement are equal, while means in between are lower (positive trend) or higher (negative trend). In case of a cubic trend, the curve first rises, drops down, and rises again (positive trend) or vice versa (negative trend). For H5 correlations between beliefs and trust at different times of measurement were investigated. In the MF– group at t_{pre} and t_2 trust was not normally distributed. As it is commonly argued that the overall *F*-test of the ANOVA (e.g., Schmider, Ziegler, Danay, Beyer, & Bühner, 2010) is robust against this violation, we reported these results with corrected test statistics.

Results

Trust. To test H1, we conducted a repeated-measures ANOVA with t_0-t_7 for the group that did not experience a malfunction (MF–). Degrees of freedom were Huynh-Feldt corrected as sphericity was not fulfilled. An *F*-test revealed a significant impact of time of measurement on trust, $F(4.27, 59.77) = 3.35$, $p = .013$, $\eta_p^2 = 0.19$. To evaluate, if the baseline (t_0) and the last measurement point (t_7) differed significantly, a *t*-test was used. It revealed a significant difference, $t(14) = -2.98$, $p = .005$, $d = 0.69$. In addition, polynomial contrasts in line with H1 revealed that trust development of the MF– group fitted a positive linear trend, $F(1,14) = 6.01$, $p = .028$, $\eta_p^2 = 0.30$.

For H2, a polynomial contrast analysis (t_2, m_1, and t_3) was conducted for the combined

TABLE 3: Polynomial Contrast Analysis for the TORs Prior and After the Malfunction

Condition	Trend	*df* (Error)	*F*	η_p^2	*p*
TOR 1	Linear	1 (29)	<0.01	0.00	.966
	Quadratic	1 (29)	5.60	0.16	.025
TOR 2 (MF–)	Linear	1 (14)	10.05	0.42	.007
	Quadratic	1 (14)	1.69	0.11	.215
TOR 2 (MF+)	Linear	1 (15)	9.71	0.39	.007
	Quadratic	1 (15)	0.15	0.01	.706

Note. TOR = take-over request; MF– = no malfunction; MF+ = with malfunction.

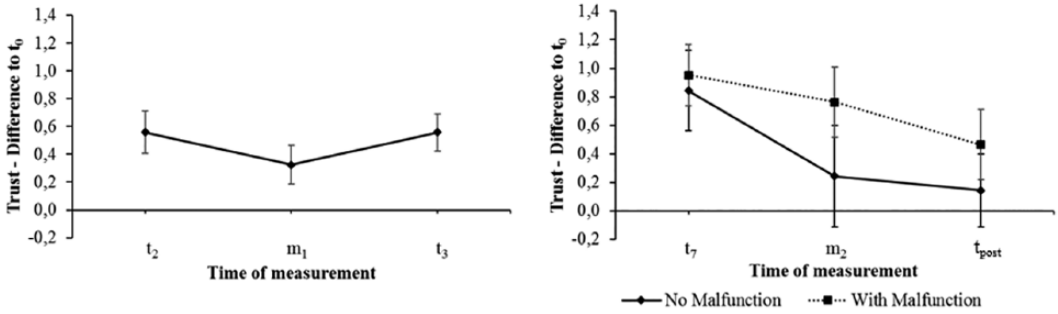

Figure 6. Trust development prior to, during, and after the first TOR (left) and the second TOR (right).
Note. t_0 serves as a baseline to facilitate comparison of the trust curves. Error bars indicate ±1 *SE*.
TOR = take-over request.

sample of both study groups (TOR 1; Table 3, Figure 6, left). In line with H2, trust development associated with the TOR showed a significant quadratic trend, while the linear one was not significant. It can be concluded that the first TOR influenced trust as hypothesized and led to a temporary decrease and a subsequent recovery. For the second TOR, both study groups showed a negative linear trend (Figure 6, right), indicating a more permanent trust impairment.

H3.1 predicted that trust temporarily decreases after a malfunction, and H3.2 predicted that trust reestablishes to the level prior to the malfunction (quadratic trend). To test this, we used polynomial contrast analysis comparing trust prior to (t_3) and directly after the malfunction (t_4) and after some further interaction with the system (t_5; see Table 4). The analysis showed that trust development of the MF– group could neither be described as linear nor as quadratic.

However, the MF+ showed a significant quadratic relationship, while a linear trend was not significant. It can be concluded that in the MF+ group trust decreased temporarily after the malfunction and was then quickly reestablished in the course of interaction with the system. Furthermore, a two-tailed *t*-test revealed no significant difference for trust at t_7 between the MF+ and MF– groups, $t(29) = -0.77$, $p = .449$, $d = 0.28$, indicating that the malfunction did not have a long-term effect on trust.

H5 predicted a significant correlation of trust and the preceding beliefs. Table 5 shows the correlations of reliability, predictability, and competence with trust at the different points of measurement. Except for the correlation between trust and competence at t_{pre} all correlations were significant, underlining a close relationship between beliefs and trust.

TABLE 4: Polynomial Contrast Analysis for Trust at $t_{3\text{-}5}$ for the Different Groups With and Without Malfunction

Condition	Trend	*df* (Error)	*F*	η_p^2	*p*
No malfunction	Linear	1 (14)	1.71	0.11	.212
	Quadratic	1 (14)	1.91	0.12	.189
With malfunction	Linear	1 (15)	0.35	0.02	.564
	Quadratic	1 (15)	15.32	0.51	.001

TABLE 5: Correlations of Trust With Reliability, Predictability, and Competence

	Study 1				Study 2			
Variable	t_{pre}	t_0	t_4	t_{post}	t_{pre}	t_0	t_4	$t_{overall}$
Reliability	0.543**	0.671**	0.594**	0.707**	0.646**	0.293*	0.466**	0.548**
Predictability	0.591**	0.484**	0.467**	0.626**	0.376**	0.256	0.480**	0.499**
Competence	0.302	0.636**	0.454*	0.654**	0.591**	0.552**	0.418**	0.467**

*$p < .05$. **$p < .01$.

STUDY 2

This study served in part as a confirmation of the findings of Study 1. As a second independent variable, system transparency was manipulated.

Method

While this study followed a very similar design in general, all TORs were removed. Thus, after the initial hand-over, participants remained in the automated mode. Furthermore, the secondary task was changed as an attention assist was included. Also, a final overall evaluation after the experimental drive ($t_{overall}$) was added to the design to assess the relationship between the beliefs and trust on a more general level. Finally, two further study groups were introduced to be able to investigate transparency as a second independent variable, which resulted in a 2 (malfunction: yes vs. no) x 2 (transparency: high vs. low) experimental design with four study groups: No malfunction (MF–)/low transparency (LT), Malfunction (MF+)/LT, MF–/high transparency (HT), and MF+/HT.

Sample. Participants were recruited like in Study 1 and received 10€ or study credit for their participation. Altogether, 50 participants completed the study. One participant had to be excluded, as the study protocol revealed an unplanned error in the automation. Two additional participants had to be excluded as they did not comply with study instructions. This resulted in a final study sample of $N = 47$ (27 female, 20 male) with $n_{MF+/LT} = 11$, $n_{MF+/HT} = 12$, $n_{MF-/LT} = 13$, $n_{MF-/HT} = 11$. The average age of the participants was $M = 27.45$ ($SD = 9.75$). They held a driving license for $M = 9.38$ ($SD = 9.16$) years and rated their experience with driving assistance with $M = 2.15$ ($SD = 1.66$). No significant group differences in these demographic variables between the study groups were found.

Procedure. In general, Study 2 followed the same procedure as Study 1. Prior to the depicted procedure, participants answered personality questionnaires (not in the scope of this research). To guarantee the participant's attention during the drive, we implemented an attention assistant system: a beep sound repeated at a 1-minute interval to which participants had to react (gaze on the road and touching the steering wheel). Instead of Tetris (Electronic Arts, 2013), Angry Birds (Rovio Entertainment Corporation, 2010) was used as the secondary task, as it allows for interruption. In the low transparency (LT)

TABLE 6: Internal Consistency of the Scales at All Times of Measurement Indicated With Cronbach's α

Variable	t_{pre}	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	$t_{overall}$
Trust	0.82	0.88	0.77	0.78	0.74	0.84	0.80	0.80	0.80	0.79	0.83
Predictability	0.73	0.77	—	—	—	0.75	—	—	—	—	0.67
Competence	0.86	0.92	—	—	—	0.89	—	—	—	—	0.86
Reliability	0.77	0.88	—	—	—	0.70	—	—	—	—	0.73

Note. The row "Trust" shows Cronbach's α, with all items of the respective scale. For trust at t_{pre}, t_0, t_4, and $t_{overall}$ the long version and for the remaining points of measurement the short version was used.

TABLE 7: Descriptive Values for Trust of All Groups at Each Measurement Point

	M_{tn} (SD_{tn})										
Condition	t_{pre}	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	$t_{overall}$
MF–/LT	5.03	5.66	5.72	5.91	5.98	6.09	6.11	6.17	6.22	6.22	6.03
	(0.57)	(0.73)	(0.94)	(0.62)	(0.73)	(0.79)	(0.76)	(0.71)	(0.66)	(0.67)	(0.66)
MF–/HT	4.47	5.13	5.60	5.67	5.73	5.61	5.93	5.85	6.04	5.98	5.81
	(0.69)	(1.20)	(0.83)	(0.92)	(1.11)	(0.97)	(0.91)	(1.00)	(0.78)	(0.91)	(1.09)
MF+/LT	4.64	4.95	5.27	5.42	5.60	5.13	5.55	5.75	5.84	5.96	5.45
	(1.23)	(1.14)	(1.13)	(1.06)	(0.98)	(1.22)	(1.09)	(1.06)	(1.04)	(0.92)	(1.12)
MF+/HT	5.20	5.25	5.53	5.75	5.75	5.67	5.83	5.82	5.90	5.90	5.81
	(0.87)	(0.84)	(0.88)	(0.85)	(0.64)	(0.61)	(0.58)	(0.59)	(0.59)	(0.56)	(0.66)

Note. MF– = no malfunction; LT = low transparency; MF+ = with malfunction; HT = high transparency.

groups, the malfunction and safe mode was removed from the practice trial. Participants in the high transparency groups (HT) received information about malfunctions prior to the practice trials. The malfunctions were explained to be resulting from ambiguous sensor information caused by reflections and glares in the driving environment by large reflecting white surfaces (i.e., the side panel of a truck), which may lead to an emergency safe mode. In addition, the HT group were instructed in detail about the safe mode and experienced a malfunction and a safe mode during the practice trial. The study groups with LT only received the information that the system is constructed in such a way that in case of a malfunction, a safe mode is activated. These groups did not receive any information on potential reasons for malfunctions, nor did they experience a system malfunction with subsequent safe mode in the practice trial.

Material. Apparatus and questionnaires. In Study 2, the same study setup and instruments were used as in Study 1. Again, the utilized scales showed satisfying levels of Cronbach's α (with the exception of predictability at $t_{overall}$; see Table 6).

Analysis. For descriptive values of the dependent variables, see Table 7. Overall, participants showed a considerably high level of trust descriptively increasing over the course of the study (see Figure 7). The groups reported an initial trust of $M = 4.85$ ($SD = 0.89$) at t_{pre} and $M = 5.26$ ($SD = 0.99$) at t_0. For both, t_{pre}, $F(3,43) = 1.80$, $p = .162$, and t_0, $F(3,43) = 1.15$, $p = .339$, no significant differences for the four study groups were found. Hypotheses were tested with the same procedures as in Study 1. Again, for some measurement points, the assumption of a normal distribution was violated (t_4 and $t_{overall}$ in the MF+/HT group). However, as stated, *F*-tests have been shown to be robust against this

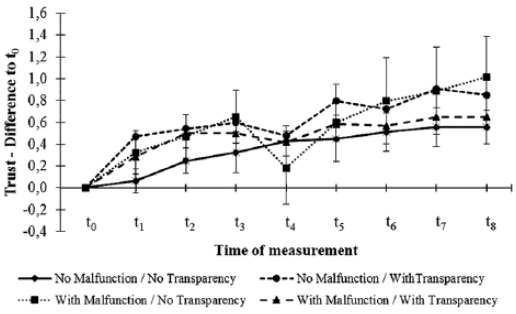

Figure 7. Trust development over the course of the experiment in the different groups with and without malfunction and transparency.
Note. Before t_4 the malfunction occurred in the respective condition. t_0 serves as a baseline to facilitate comparison of the trust curves. Error bars indicate ±1 *SE*.

violation (Schmider et al., 2010) and thus are reported with corrected statistics.

Results

H1 predicted that trust increases over time for the MF–/LT group (this group resembles the MF– group of Study 1). An ANOVA (Huynh-Feldt corrected) with all points of measurement from t_0 to t_8 revealed a significant effect of time of measurement on trust, $F(2.46, 29.51) = 3.22$, $p = .045$, $\eta_p^2 = 0.21$. Furthermore, in line with H1, a *t*-test between t_0 and t_8 revealed a significant difference in trust, $t(12) = -3.64$, $p = .002$, $d = 0.79$, and polynomial contrasts showed that trust development of the MF–/LT group fitted a positive linear trend, $F(1,12) = 6.31$, $p = .027$, $\eta_p^2 = 0.35$.

H3.1 stated that trust will temporarily decrease after a malfunction in a non-transparent system. For the MF+/LT group, the same pattern as in Study 1 was expected. Polynomial contrasts confirm the findings of Study 1 and showed a good approximation to a quadratic, but not to a linear trend for the MF+ group, while in the MF– group neither trend was significant (Table 8). Again, it can be concluded that a malfunction leads to a trust decrease, but that trust was rebuilt in the subsequent interaction interval (see Figure 7). A two-tailed *t*-test revealed no significant difference in the last measurement point (t_8) between the two LT groups, $t(22) = 0.78$, $p = .446$, $d = 0.32$, again indicating that the malfunction did not have any long-term effects on trust (H3.2).

H4 proposed that system transparency can prevent a trust decline after a malfunction. First, the success of the transparency manipulation was inspected. As a manipulation check, a set of three items on the expectation of a malfunction were combined to a scale mean ($\alpha = .684$). These items were only presented in the MF+ groups ($n = 22$), as they directly referred to the experienced malfunction participants. In a *t*-test, a significantly higher transparency for MF+/HT ($M = 3.82$; $SD = 1.12$) than for MF+/LT ($M = 2.39$; $SD = 1.03$) was found, $t(20) = 3.10$, $p = .006$, $d = 1.32$, indicating a successful manipulation of transparency. Interestingly, neither the information about reasons for system malfunctions nor the introduction of the safe mode led to a significant decrease in trust at t_0 as indicated by a nonsignificant *t*-test comparing the two LT and HT groups, $M_{LT} = 5.33$ ($SD = 0.98$), $M_{HT} = 5.19$ ($SD = 1.01$; $t(45) = .485$; $p = .630$; $d = -0.14$). This implies that transparent information on system shortcomings does not necessarily lead to an initial trust reduction. To test H4, polynomial contrasts were calculated with t_3, t_4, and t_5 for the two study groups that received transparent information (MF+/HT vs. MF–/HT). For both groups neither the linear nor the quadratic trend were significant (see Table 8). It follows that if users interacted with a transparent system, a malfunction did not result in a significant decline in trust, as opposed to a situation in which they did not receive any transparency information in advance.

As in Study 1 and in line with H5, beliefs showed significant correlations with trust at t_{pre}, t_0, t_4, and $t_{overall}$ (except for the correlation of predictability and trust at t_0; see Table 5). These correlations strongly support the hypothesis that beliefs play a crucial role for trust establishment.

Overall discussion and implications for the design of automated vehicles. Taken together, all study hypotheses gained substantial support in the two simulator experiments. In line with H1, results of both studies show that trust in an automated vehicle increases during the early phase of interaction. This is in line with earlier studies (e.g., Beggiato et al., 2015; Dzindolet et al., 2003; Hergeth et al., 2016; Lee & Moray,

TABLE 8: Polynomial Contrast Analysis for Trust at $t_{3\text{-}5}$ for the Different Groups

	Trend	*df* (Error)	*F*	η_p^2	*p*
No malfunction/low transparency	Linear	1 (12)	0.35	0.03	.568
	Quadratic	1 (12)	0.15	0.01	.709
With malfunction/low transparency	Linear	1 (10)	0.09	0.01	.772
	Quadratic	1 (10)	5.65	0.36	.039
No malfunction/high transparency	Linear	1 (10)	1.83	0.16	.206
	Quadratic	1 (10)	2.95	0.23	.117
With malfunction/high transparency	Linear	1 (11)	1.36	0.11	.269
	Quadratic	1 (11)	1.87	0.15	.199

1992). In Study 1, it was found that trust decreased temporarily after the experience of a TOR. In line with H2, trust was recovered quickly in the course of the drive. This supports the findings of Hergeth and colleagues (2015). Surprisingly, for a later second TOR (m_2), no subsequent trust recovery could be observed. Also, a previous experience of a system malfunction did not make a difference in the subsequent trust reduction after a second TOR. It remains an open question for future research if repeated TORs lead to more permanent consequences for trust or if trust recovers over a longer period of time.

In both studies, the hypothesis that trust decreases after a malfunction could be supported. Furthermore, there was strong evidence suggesting that trust recovers very quickly after such an experience. Moreover, there was no significant difference for trust between the MF– and MF+ groups (or MF–/LT and MF+/LT, respectively) at the last point of measurement, supporting the notion that there is no permanent trust reduction by a single malfunction of an automated driving system.

In line with H4, it was found in Study 2 that high transparency about malfunctions and the character of a safe mode led to an absence of a trust reduction when such a malfunction occurred. While the low transparency group that experienced a malfunction showed a significant decrease in trust at t_4 (H3.1), the group that received information about system limitations and the character of the safe mode beforehand (high transparency group) did not show this decrease (H4). System transparency seems to diminish trust reduction in face of a system malfunction. With sufficient information, users seem to better anticipate system behavior and are able to adapt their expectations early in the process and thus are not negatively surprised when the malfunction occurs. This is in line with earlier studies showing that appropriate information about system functioning may lead to a facilitated trust calibration when system malfunctions occur (e.g., Dzindolet et al., 2003; Wang et al., 2009).

Regarding the hypothesized relationship between beliefs and trust (H5), in both studies, 22 of the investigated 24 correlations between beliefs and trust were significant (prior to, during, and after system interaction). This provides initial support for the notion of the introduced trust calibration model that trust fluctuations in trust calibration are corresponding with changes in beliefs about the automation. For a better understanding of the role of beliefs in trust calibration, further research should be conducted.

Implications for trust-reliability calibration and automation design. The results of the presented studies provide insights into the nature of trust calibration prior to and during the early phase of interaction with highly automated vehicles and support some major implications of the presented trust calibration model (see Figure 1).

First, the reported findings show that trust dynamically changes in the early phase of interaction and that trust calibration follows the mechanics of an updating feedback loop, as proposed in the model (Figure 1). In line with our reasoning and the frameworks by Lee and See

(2004) and Hoff and Bashir (2015), the reported findings support the idea that trust is established in a dynamic process in which new information is used to calibrate expectations (beliefs and attitudes) of system capabilities and functioning. In this feedback loop, beliefs and attitudes (e.g., dynamic learned trust) are updated as experience with a system increases. With new information, this evaluation is either stabilized (affirmative information) or adjusted. In the depicted studies, in case of a flawless interaction, trust development followed a linear increase. In line with the "perfect automation schema" (Dzindolet et al., 2003) in the face of any unforeseen events related to system reliability, trust declined for an instance, but recovered with continued error-free interaction. Taken together, the early phase of interacting with an unfamiliar automated system is of major importance for the development of trust and thus should be carefully taken into consideration in the design of automated vehicles.

Second, findings for H1-3 support the notion that information provided prior to and during initial system use interact in trust calibration during the drive (e.g., Hoff & Bashir, 2015). Moreover, the support for H4, in which transparent information inhibited trust reduction subsequent to a malfunction, underlines the importance of a priori information during trust calibration. In fact, combined evidence of H3 and H4 indicates that transparent information about a malfunction prior to actual system interaction may serve as a safeguard against a trust decrease after system malfunction, as the expectation of system functioning is no longer violated. In this sense, expectation and experiences seem to interact in trust calibration.

Third, as supported by the findings of Study 2, a priori information about system deficiencies (e.g., malfunctions) as well as the experience of system limitations (TORs), system errors, and a safe mode in the early phase of system interaction do not necessarily lead to initial trust reductions or any negative long-term consequences for trust development. This allows for a positive view of a priori information to foster trust calibration even with seemingly negative information (see also Payre, Cestac, & Delhomme, 2014). Knowledge about situations leading to a TOR or entailing a higher risk for system malfunctions may reduce uncertainties and therefore help users to calibrate their trust more efficiently (e.g., Helldin et al., 2013) and use the system safely. In fact, the design of prior information may be a key determinant for how information during system interaction is interpreted and thus for the establishment of appropriate trust. It may even be a perspective for automation design to favor trust calibration by letting drivers experience low-consequence system limits (TORs) and malfunctions during the early phase of driving, to facilitate the construction of a valid mental model.

Taken together, the reported findings argue for an implementation of driver training, user guides, and tutorials for automated vehicles in favor of trust calibration (e.g., Chavaillaz, Wastell, & Sauer, 2016; Hoc, Young, & Blosseville, 2009; Muir, 1994; Muir & Moray, 1996). Valid a priori information may help to establish a more realistic picture about which situations a system can handle and under which conditions it faces problems. This in turn should diminish distrust and overtrust from the beginning of system use (e.g., Kazi et al., 2007; Muir, 1994). In addition, car interfaces for highly automated driving should be designed to foster trust calibration during driving by providing real-time information on system behavior, status, and functioning as well as maneuver planning (e.g., Endsley, 2017). This is especially important in the context of TORs and malfunctions as their plausibility seems to be a key for trust recovery. Furthermore, information about system components and functionality could accumulate drivers' comprehension and mental models and hereby enable them to realistically assess a system's trustworthiness. Taken together, an implementation of these kinds of calibration information should prevent distrust and overtrust at the same time (in specific in SAE level 3) and thereby provide means to enhance safety in highly automated driving.

Limitations and future research. Several limitations in the reported studies need consideration and should be addressed in future research. First, in simulator studies, the associated risks and consequences of malfunctions, errors, or accidents are considerably lower as compared to

real driving (De Winter, Van Leuween, & Happee, 2012). On the other hand, at this point driving simulators are the most valid experimental means to investigate the psychological mechanisms in highly automated driving. The psychological realism of the reported studies was increased by clearly communicated consequences in cases of accidents or traffic violations. Future research should investigate if the same characteristics for early trust calibration processes hold true for real-world automated vehicles. Second, while the findings of Study 1 were further supported in Study 2, the findings on transparency should be replicated and extended in larger samples. Third, these studies only investigate the effects of single malfunctions with rather mild consequences. Future studies should also include repeated, high-risk malfunctions and varying degrees of perceived control. In case of the TORs, Study 1 provided first evidence that repeated TORs may be associated with more permanent trust reduction while in other research no permanent decrease was found (e.g., Hergeth et al., 2015). Thus, an investigation of moderating variables seems fruitful. Fourth, the transparency manipulation of Study 2 is only one of manifold possibilities to increase transparency. It is fair to assume that timing and character of transparent information has differential effects on trust development to be explored in future studies (e.g., a dynamic display of system performance could lead to general higher trust levels). Fifth, it would be interesting to include behavioral measures of trust and examine their relation to self-reported trust (e.g., Miller et al., 2016). Sixth, some groups showed a non-significant decrease in trust at t_4. This might be a consequence of a more critical system evaluation due to a higher number of investigated variables at this point. It is advisable to use the same item context at all points of measurement in future studies.

CONCLUSION

Walker and colleagues (2016) characterize trust as "a dynamic phenomenon, moving along a continuum, spiraling upwards or downwards based on perceptions of how the vehicle system operates, beliefs about what those perceptions mean, and the positive or negative attitudinal attribution that arises" (p. 4). By providing an enhanced research model of trust calibration and empirical findings of two simulator studies, this paper contributes to an understanding of trust as a dynamic attitude that is calibrated prior and during the interaction with an automated vehicle along the available information (e.g., TORs, system malfunctions, and system transparency). To optimize an automated driving system, this dynamic psychological process should be addressed in system design and the communication about a system.

ACKNOWLEDGMENTS

Thanks to all interviewers, who helped conducting the reported studies. Thanks to all colleagues from the Human Factors Department—especially Kristin Mühl and Philipp Hock—as well as Morten Moshagen and Matthias Messner for their helpful comments and discussions.

KEY POINTS

- A model of trust calibration based on Lee and See (2004) is presented.
- Two simulator studies on the dynamics of trust development in the early phase of interaction with highly automated vehicles were presented. Study 1 showed a steady increase of trust in the case of an error-free automation and a temporal trust reduction in case of take-overs and system malfunction. In both cases, trust was recovered in subsequent error-free interaction with the system.
- In Study 2, these findings for error-free functioning and trust recovery after a system malfunction could be supported in a second independent sample. As an additional finding, in Study 2 it could be shown that a priori information about the causes and the characteristic of a malfunction eliminated the decrease in trust in case of a system malfunction.
- The study findings provide new insights into the psychological processes involved in trust calibration prior to and during the interaction with automated systems.

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179–211. doi:10.1016/0749-5978(91)90020-T

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood-Cliffs, NJ: Prentice Hall.

Beggiato, M., & Krems, J. F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation Research Part F: Traffic Psychology and Behaviour*, *18*, 47–57. doi:10.1016/j.trf.2012.12.006

Beggiato, M., Pereira, M., Petzoldt, T., & Krems, J. (2015). Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *35*, 75–84. doi:10.1016/j.trf.2015.10.005

Bühner, M., & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler* [Statistics for psychologists and social scientists]. Hallbergmoos, Germany: Pearson Deutsch-land GmbH.

Chavaillaz, A., Wastell, D., & Sauer, J. (2016). System reliability, performance and trust in adaptable automation. *Applied Ergonomics*, *52*, 333–342. doi:10.1016/j.apergo.2015.07.012

Chen, Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency*. MD: Aberdeen Proving Ground. Retrieved from https://www.arl.army.mil/arlreports/2014/technical-report.cfm?id=7066

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013, March 3–6). Impact of robot failures and feedback on real-time trust. In *8th ACM/IEEE international conference on human-robot interaction* (pp. 251–258). IEEE. doi:10.1109/HRI.2013.6483596

De Winter, J., Van Leuween, P., & Happee, P. (2012, August 28–31). Advantages and disadvantages of driving simulators: A discussion. In A. J. Spink, F. Grieco, O. E. Krips, L. W. S. Loijens, L. P. J. J. Noldus, & P. H. Zimmermann (Eds.), *Proceedings of measuring behavior 2012* (pp. 47–50). Utrecht, The Netherlands.

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*, 697–718. doi:10.1016/S1071-5819(03)00038-7

Ekman, F., Johansson, M., & Sochor, J. L. (2016, October 23–27). *To see or not to see: The effect of object recognition on users' trust in "automated vehicles."* Proceedings of the 9th Nordic Conference on Human-Computer Interaction (Article No. 42). doi:10.1145/2971485.2971551

Electronic Arts. (2013). TETRIS (Version 3.0.10) [Mobile application software]. Retrieved from https://play.google.com/store?hl=de

Emzivat, Y., Ibanez-Guzman, J., Martinet, P., & Roux, O. H. (2017, June). *Dynamic driving task fallback for an automated driving system whose ability to monitor the driving environment has been compromised*. IEEE Intelligent Vehicles Symposium, Redondo Beach, CA. Retrieved from https://hal.archives-ouvertes.fr/hal-01724931

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors*, *59*, 5–27. doi:10.1177/0018720816681350

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

Forster, Y., Kraus, J., Feinauer, S., & Baumann, M. (2018, September 23–25). *Calibration of trust expectancies in conditionally automated driving by brand, reliability information and introductionary videos: An online study*. Proceedings of the 10th international conference on Automotive User Interfaces and Interactive Vehicular Applications, Association for Computing Machinery, New York, NY.

George, D., & Mallery, M. (2003). *Using SPSS for Windows step by step: A simple guide and reference*. Boston, MA: Allyn & Bacon.

Gold, C., Damböck, D., Lorenz, L., & Bengler, K. (2013). "Take over!" How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*, 1938–1942. doi:10.1177/1541931213571433

Gold, C., Körber, M., Lechner, D., & Bengler, K. (2016). Taking over control from highly automated vehicles in complex traffic situations. *Human Factors*, *58*, 642–652. doi:10.1177/0018720816634226

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, *24*, 1494–1509.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*, 517–527. doi:10.1177/0018720811417254

Hartwich, F., Witzlack, C., Beggiato, M., & Krems, J. (2018). The first impression counts—A combined driving simulator and test track study on the development of trust and acceptance of highly automated driving. *Transportation Research Part F: Traffic Psychology and Behavior*. Advance online publication. doi:10.1016/j.trf.2018.05.012

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In J. Terken (Ed.), *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications* (pp. 210–217). New York, NY: Association for Computing Machinery. doi:10.1145/2516540.2516554

Hergeth, S., Lorenz, L., & Krems, J. F. (2017). Prior familiarization with takeover requests affects drivers' takeover performance and automation trust. *Human Factors*, *59*, 457–470. doi:10.1177/0018720816678714

Hergeth, S., Lorenz, L., Krems, J. F., & Toenert, L. (2015, June 22–25). Effects of take-over requests and cultural background on automation trust in highly automated driving. In *Proceedings of the eighth international driving symposium on human factors in driver assessment, training and vehicle design* (pp. 331–337). Iowa City: Public Policy Center, University of Iowa. doi:10.17077/drivingassessment.1591

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors*, *58*, 509–519. doi:10.1177/0018720815625744

Hoc, J.-M., Young, M. S., & Blosseville, J.-M. (2009). Cooperation between drivers and automation: Implications for safety. *Theoretical Issues in Ergonomics Science*, *10*, 135–160. doi:10.1080/14639220802368856

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*, 407–434. doi:10.1177/0018720814547570

IBM Corp. (2016). *IBM SPSS statistics* (Computer software). Ehningen, Germany: Author.

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*, 53–71. doi:10.1207/S15327566IJCE0401_04

Kazi, T. A., Stanton, N. A., Walker, G. H., & Young, M. S. (2007). Designer driving: Drivers' conceptual models and level of trust

in adaptive cruise control. *International Journal of Vehicle Design*, *45*, 339–360. doi:10.1504/IJVD.2007.014909

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, *66*, 18–31. doi:10.1016/j.apergo.2017.07.006

Lacson, F. C., Wiegmann, D. A., & Madhavan, P. (2005). Effects of attribute and goal framing on automation reliance and compliance. In *Human factors and ergonomics society 49th annual meeting proceedings* (pp. 482–486). Thousand Oaks, CA: SAGE. doi:10.1177/154193120504900357

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

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80. doi:10.1518/hfes.46.1.50_30392

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*, 277–301. doi:10.1080/14639220500337708

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In G. Gable & M. Vitale (Eds.), *Proceedings of the 11th Australasian Conference on Information Systems* (pp. 53–64). Melbourne, Australia: AIS.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*, 194–210. doi:10.1518/001872008X288574

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors*, *57*, 740–753. doi:10.1177/0018720815581247

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., & Ju, W. (2016). Behavioral measurement of trust in automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*, 1849–1853. doi:10.1177/1541931213601422

Molina, C. B. S. T., Almeida, J. R., de Vismari, L. F., González, R. I. R., Naufal, J. K., & Camargo, J. B. (2017, June 26–29). *Assuring fully autonomous vehicles safety by design: The autonomous vehicle control (AVC) module strategy*. 47th annual IEEE/IFIP international conference on Dependable Systems and Networks Workshops (DSN-W). doi:10.1109/DSN-W.2017.14

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

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

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*, 1905–1922. doi:10.1080/00140139408964957

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*, 429–460. doi:10.1080/00140139608964474

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51–55. doi:10.1145/975817.975844

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

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans*, *30*, 286–297. doi:10.1109/3468.844354

Payre, W., Cestac, J., & Delhomme, P. (2014). Intention to use a fully automated car: Attitudes and a priori acceptability. *Transportation Research Part F: Traffic Psychology and Behaviour*, *27*(B), 252–263. doi:10.1016/j.trf.2014.04.009

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge, UK: Cambridge University Press.

Rovio Entertainment Corporation. (2010). Angry birds (Version 7.9.3) [Mobile application software]. Retrieved from https://play.google.com/store?hl=de

SAE International. (2014). *Automated driving-levels of driving automation are defined in new SAE international standard J3016*. Retrieved from https://www.oemoffhighway.com/electronics/smart-systems/automated-systems/document/12183694/automated-driving-levels-of-driving-automation-are-defined-in-new-sae-international-standard-j3016

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? *Methodology*, *6*, 147–151. doi:10.1027/1614-2241/a000016

Seong, Y., & Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics*, *38*, 608–625. doi:10.1016/j.ergon.2008.01.007

Van den Beukel, A. P., van der Voort, M. C., & Eger, A. O. (2016). Supporting the changing driver's task: Exploration of interface designs for supervision and intervention in automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *43*, 279–301. doi:10.1016/j.trf.2016.09.009

Walch, M., Mühl, K., Kraus, J., Stoll, T., Baumann, M., & Weber, M. (2017). From car-driver-handovers to cooperative interfaces: Visions for driver–vehicle interaction in automated driving. In: Meixner G., Müller C. (Eds.), *Automotive user interfaces. Human–Computer Interaction Series* (pp. 273–294). Cham, Switzerland: Springer.

Walker, G. H., Stanton, N. A., & Salmon, P. (2016). Trust in vehicle technology. *International Journal of Vehicle Design*, *70*, 157–182. doi:10.1504/IJVD.2016.074419

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*, 281–291. doi:10.1177/0018720809338842

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, *2*, 352–367. doi:10.1080/14639220110110306

Wintersberger, P., von Sawitzky, T., Frison, A.-K., & Riener, A. (2017). Traffic augmentation as a means to increase trust in automated driving systems. In *Proceedings of the 12th biannual conference on Italian SIGCHI chapter* (pp. 1–17). New York, NY: Association for Computing Machinery. doi:10.1145/3125571.3125600

Woods, D. (2001). *Human-centered design of automated agents and human-automation team play*. Retreived from http://www.erogersphd.com/EMorePages/HRI/HRI-ARCHIVE/woods1.pdf

Würzburger Institute for Traffic Science GmbH. (2014). Driving simulation and SILAB 5.1 (Computer software). Veitshöchheim: Author. Retrieved from https://wivw.de/de/silab/

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction* (pp. 408–416). New York, NY: Association for Computing Machinery. doi:10.1145/2909824.3020230

Johannes Kraus is a research assistant at the Human Factors Department at Ulm University in Germany. He received his MS from the University of Mannheim in 2013. His research interests lie in the decision processes related to the interaction with automated systems, especially automated vehicles and humanoid robots.

David Scholz works at the Department of Human Factors at Ulm University. He received his BS in psychology from the University of Ulm in 2017. His research interest lies in trust in automation and intercultural studies on the use of technology.

Dina Stiegemeier works at the Human Factors Department at Ulm University. She received her BS in psychology from Ulm University in 2017. Her research interest lies in trust in automation, focusing especially on HMI aspects.

Martin Baumann is a professor in Human Factors at Ulm University and received his PhD from Chemnitz University of Technology in 2001. His main research interests are trust, comprehension of dynamic situations and intelligent systems, persuasive technologies as well as cooperative human-machine interaction in different domains, mainly traffic, and human-robot interaction.

Date received: September 26, 2018
Date accepted: May 1, 2019