ulm university universität ulm

Institut für Psychologie und Pädagogik, Abteilung Human Factors

Psychological Processes in the Formation and Calibration of Trust in Automation

Dissertation

zur Erlangung des Doktorgrades
doctor rerum naturalium (Dr. rer. nat.)

der Fakultät für Ingenieurwissenschaften, Informatik und Psychologie der Universität Ulm

vorgelegt von

Johannes Maria Kraus

aus Göttingen

2020

Amtierender Dekan: Prof. Dr.-Ing. Maurits Ortmanns, Universität Ulm

GutachterInnen:

1. Prof. Dr. Martin Baumann, Universität Ulm
2. Prof. Dr. Josef Krems, Technische Universität Chemnitz
3. Prof. Dr. Tina Seufert, Universität Ulm

Promoviert am 10.07.2020

Peer-Reviewed Articles and Book Chapters Published During the Course of the Dissertation

Kraus, J., Scholz, D., & Baumann, M. (2020). What's driving me? - Exploration and validation of a hierarchical personality model for trust in automated driving. *Human Factors*. Advance online publication. https://doi.org/10.1177/0018720820922653

Kraus, J., Scholz, D., Messner, E.-M., Messner, M., & Baumann, M. (2020). Scared to trust? – Predicting trust in highly automated driving by depressiveness, negative self-evaluations and state anxiety. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02917

Kraus, J., Forster, Y., Hergeth, S., & Baumann, M. (2019). Two routes to trust calibration: Effects of reliability and brand information on trust in automation. *International Journal of Mobile Human Computer Interaction, 11*(3), 1-17. https://doi.org/10.4018/IJMHCI.2019070101

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

Hock, P., Babel, F., **Kraus, J.**, Rukzio, E., & Baumann, M. (2019). Towards opt-out permission policies to maximize the use of automated driving. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 101-112. https://doi.org/10.1145/3342197.3344521

Hock, P., **Kraus, J.**, Babel, F., Walch, M., Rukzio, E., & Baumann, M. (2018). How to design valid simulator studies for investigating user experience in automated driving: Review and hands-on considerations. *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 105-117. https://doi.org/10.1145/3239060.3239066

Forster, Y., **Kraus, J.**, Feinauer, S., & Baumann, M. (2018). 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*, 118-128. https://doi.org/10.1145/3239060.3239070

Wintersberger, P., Noah, B. E., **Kraus, J.**, McCall, R., Mirnig, A. G., Kunze, A., Thakkar, S., & Walker, B. N. (2018). Second workshop on trust in the age of automated driving. *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 56-64. https://doi.org/10.1145/3239092.3239099

Kraus, M., **Kraus, J.**, Baumann, M., & Minker, W. (2018, May). Effects of gender stereotypes on trust and likability in spoken human-robot interaction. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. https://www.aclweb.org/anthology/L18-1018

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 G. Meixner & C. Müller (Eds.), *Automotive user interfaces: Creating interactive experiences in the car* (pp. 273-294). Springer International Publishing. https://doi.org/10.1007/978-3-319-49448-7_10

Noah, B. E., Wintersberger, P., Mirnig, A. G., Thakkar, S., Yan, F., Gable, T. M., **Kraus, J.**, & McCall, R. (2017). First workshop on trust in the age of automated driving. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct*, 15-21. https://doi.org/10.1145/3131726.3131733

Kraus, J. M., Nothdurft, F., Hock, P., Scholz, D., Minker, W., & Baumann, M. (2016). Human after all: Effects of mere presence and social interaction of a humanoid robot as a co-driver in automated driving. *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 129-134. https://doi.org/10.1145/3004323.3004338

Hock, P., **Kraus, J.**, Walch, M., Lang, N., & Baumann, M. (2016). Elaborating feedback strategies for maintaining automation in highly automated driving. *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 105-112. https://doi.org/10.1145/3003715.3005414

Kraus, J. M., Sturn, J., Reiser, J. E., & Baumann, M. (2015). Anthropomorphic agents, transparent automation and driver personality: Towards an integrative multi-level model of determinants for effective driver-vehicle cooperation in highly automated vehicles. *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 8-13. https://doi.org/10.1145/2809730.2809738

Danksagung

Zunächst möchte ich mich insbesondere herzlich bei meinem Doktorvater Prof. Dr. Martin Baumann bedanken. Du hast mich auf dem Prozess dieser Promotion zu jedem Zeitpunkt unterstützt, an mich geglaubt, mir die Freiheit gelassen, genau das zu erforschen, was mich interessiert hat und mich doch immer wieder auf den richtigen Weg gelenkt, falls ich von diesem einmal abkam. Weiterhin danke ich Prof. Dr. Josef Krems dafür, dass er sich meinem Thema angenommen hat. Auch möchte ich Prof. Dr. Morten Moshagen, Prof. Dr. Klaus Melchers und Prof. Dr. Tina Seufert für ihre Beteiligung in meiner Prüfungskommission danken. Dann gilt ein besonderer Dank natürlich meinen lieben Kolleginnen und Kollegen in der Abteilung, die mich in den letzten Jahren begleitet haben. Ich bin froh, dass ich immer auf Eure Ideen, Gedanken, Kreativität und Kritik bauen kann und dass mich mit vielen von Euch mittlerweile eine gute Freundschaft verbindet. Insbesondere möchte ich mich hier nochmal bei Kristin bedanken. Du bist immer da, wenn man Dich braucht und Du hast mich durch deine herzliche Art und deinem wissenschaftlichen Können immer unterstützt. Weiterhin möchte ich mich bei Philipp für die vielen Lachkrämpfe und die technische Unterstützung bedanken. Außerdem auch bei Linda, die mir besonders auf den letzten Metern der Arbeit noch viele wertvolle Ratschläge gegeben hat. Dann gilt mein Dank weiterhin insbesondere Tanja, Marcel, Jürgen, Franziska und aus dem erweiterten Kollegenkreis Dana und Jana. Ganz besonders danke ich außerdem Renate, die mir auf diesem Weg immer mit Rat und Tat zur Seite stand. Weiterhin möchte ich mich bei allen wissenschaftlichen Hilfskräften bedanken, die mich bei meiner Arbeit unterstützt haben. Allen voran natürlich David, der für mich im Laufe der Zeit zu einem wichtigen Sparringpartner und einem guten Freund geworden ist. Besonders möchte ich auch Michael herzlich für seine gute Arbeit danken. Dann gilt ein besonderer Dank meinen Koautorinnen und Koautoren. Hier möchte ich insbesondere nochmal Matthias erwähnen. Du warst immer für mich da, um mir mit Deiner Freundschaft und deinem fachlichen Wissen zur Seite zu stehen. Auch Yannick sei hier nochmal erwähnt, mit dem man insbesondere freitags sehr gut telefonieren kann.

Es ist mir weiterhin ein wichtiges Anliegen, mich bei allen Menschen zu bedanken, die mich im privaten Umfeld bei meiner Arbeit unterstützt haben. Mein Dank gilt zunächst meiner Familie – meiner Mutter Stephanie, meinem Vater Bernhard und meiner Schwester Catharina mit ihrem Mann Dennis. Ihr habt mich irgendwie dazu gebracht, das hier zu schaffen. Ihr habt immer hinter mir gestanden, Ihr habt mich mit Eurer Liebe und Eurem Glauben an mich immer unterstützt. Ich danke Euch von Herzen und ich bin wahnsinnig froh, dass ich Euch habe. Ich danke auch allen meinen Freunden. Die Mannheimer Freunde, die mir immer noch genauso wichtig und lieb sind wie damals (danke für diesen Trick mich loszuwerden). Insbesondere seien hier Christin, Sebastian und Mathi genannt. Danke auch an alle meine Ulmer Freunde. Ihr habt es geschafft, dass ich in Ulm eine neue Heimat gefunden habe. Ihr seid mir sehr wichtig und ich danke Euch für all die schönen Stunden und die moralische Unterstützung. Abschließend möchte ich mich besonders bei Catharina bedanken. Du hast mir die nötige Ruhe, Sicherheit und Kraft gegeben, dieses Projekt erfolgreich abzuschließen. Du machst mein Leben jeden Tag besser. Danke, dass Du in dieser Zeit mit viel Verständnis, Loyalität, Fürsorge und Liebe zu mir gestanden hast. Ich freue mich auf alles was jetzt kommt.

Table of Contents

List of Figures IX
List of Tables X
List of Abbreviations XI
Summary XII
Zusammenfassung XIV
1 Introduction 16
1.1 Trust in Automation 18
1.1.1 Calibrated Trust and Trust Calibration 19
1.2 Trust Calibration in Driver-Vehicle Interaction with Automated Vehicles 20
1.3 Theories of Trust and Trust Formation 22
1.3.1 Theories on Levels of Trust Attributions 23
1.3.2 Lee and See's Trust in Automation Model 24
1.3.3 Trust Layers 26
1.4 Review on Factors Influencing Trust in Automation 28
1.4.1 Characteristics of the User 28
1.4.2 Characteristics of the Automated System 29
1.4.3 Characteristics of the Situation, Task Environment 30
1.5 Problem Statement and Aims of this Dissertation 31
1.5.1 Aims of the Proposed Framework for Trust in Automation 32
1.5.2 Aims of the Empirical Studies 32
2 Research Framework: Three Stages of Trust in Automation 33
2.1 Propositions of the Three Stages of Trust Framework 33
3 Summary of Studies 43
3.1 Studies 1 and 2: What's Driving Me? - Exploration and Validation of a Hierarchical Personality Model for Trust in Automated Driving 44
3.1.1 Introduction and Study Hypotheses 44
3.1.2 Study 1 - Method 46
3.1.3 Study 1 - Results 47
3.1.4 Study 2 - Method 48
3.1.5 Study 2 - Results 48
3.1.6 Study 1 and Study 2 - Conclusion 49
3.2 Study 3: Scared to Trust? – Predicting Trust in Highly Automated Driving by Depressiveness, Negative Self-Evaluations and State Anxiety 51
3.2.1 Introduction and Study Hypotheses 51
3.2.2 Method 53
3.2.3 Results 53

3.2.4 Conclusion....55
3.3 Study 4: Two Routes to Trust Calibration: Effects of Reliability and Brand Information on Trust in Automation....56
3.3.1 Introduction and Study Hypotheses....56
3.3.2 Method....57
3.3.3 Results....58
3.3.4 Conclusion....60
3.4 Studies 5 and 6: The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction and System Transparency....61
3.4.1 Introduction and Study Hypotheses....61
3.4.2 Study 5 - Method....62
3.4.3 Study 5 - Results....63
3.4.4 Study 6 - Method....64
3.4.6 Study 6 - Results....64
3.4.7 Studies 5 and 6 - Conclusion....65
4 General Discussion....66
4.1 Summary of Findings and Evaluation of the Proposed Framework....66
4.2 Theoretical Integration and Directions for Future Research....71
4.2.1 Trust as a Dynamic and History-based Variable....71
4.2.2 Propensity to Trust....72
4.2.3 Initial Learned Trust....73
4.2.4 Dynamic Learned Trust....74
4.2.5 Influences of Person Characteristics, System Features, and Characteristics of the Situation....76
4.2.6 Recommendations for the Design of Research on Trust in Automation....77
4.3 Practical Implications....78
4.3.1 Designing for Calibrated Trust, Diverse Users, and Situational Variability....79
4.3.2 Trust Calibration Prior to Interaction....79
4.3.3 Trust Calibration During Interaction....81
4.4 Strengths and Limitations....82
4.4.1 Evaluation of the Proposed Framework....82
4.4.2 Evaluation of the Studies....83
4.5 Conclusion....85
References....87
Appendix....105

Appendix A: The Short Learned Trust in Automation Scale, German Version (LETRAS-G) ...106
Appendix B: Research Articles ...108
Article I: What's Driving Me? – Exploration and Validation of a Hierarchical Personality Model for Trust in Automated Driving ...109
Article II: Scared to Trust? – Predicting Trust in Highly Automated Driving by Depressiveness, Negative Self-Evaluations and State Anxiety ...140
Article III: Two Routes to Trust Calibration: Effects of Reliability and Brand Information on Trust in Automation ...156
Article IV: The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction and System Transparency ...174
Erklärung ...194

List of Figures

Figure 1: Model by Lee and See (2004)……..24
Figure 2: Modell by Hoff and Bashir (2015)……...26
Figure 3: Three Stages of Trust Framework……...34
Figure 4: Studies 1 and 2. Hypothesized hierarchical structure of personality traits……...45
Figure 5: Study 1. Model fit and path weights of the final model……47
Figure 6: Study 2. Model fit and path weights of the final model……49
Figure 7: Study 3. Results of the path models separate for each trait……..……..……...54
Figure 8: Study 4. Study procedure……57
Figure 9: Study 4. Pooled means for the interaction between reliability and need for cognition.....59
Figure 10: Study 5. Study procedure……..62
Figure 11: Studies 5 and 6. Trust development in the experimental groups …….…....….63
Figure 12: Summary of study findings and inclusion in the Three Stages of Trust Framework…..68

List of Tables

Table 1: Means and standard deviations of trust at t_1 and t_2 ……..……..……58

Appendix: Table A1: Item wording and item characteristics of the LETRAS-G ……..…..……107

List of Abbreviations

3M	Meta-Theoretic Model of Motivation
ACC	Automated Cruise Control
AI	Artificial Intelligence
ANOVA	Analysis of Variance
ATS	Automation Trust Scale
CI	Confidence Interval
e.g.	Exempli Gratia (for example)
ELM	Elaboration Likelihood Model
et al.	Et Alia (and others)
H	Hypothesis
HT/LT	High/Low Transparency
LETRAS-G	Short Learned Trust in Automation Scale, German Version
LOC	Locus of Control
MF+/MF-	Malfunction/No Malfunction
NFC	Need for Cognition
OEM	Original Equipment Manufacturer
PAS	Perfect Automation Scheme
PLABE	Theory of Planned Behavior
SAE	Society of Automotive Engineers
SEM	Structural Equation Model
t	Time of Measurement
TAM	Technology Acceptance Model
TOR	Take Over Request
TRA	Theory of Reasoned Action
UTAUT	Unified Theory of Acceptance and Use of Technology

In this thesis, for reasons of better readability alone, at times, only the masculine form is used, which is to be understood without regard to gender.

Summary

Automation is one of the key drivers of technical progress. Hereby, automated driving is an area in which this progress has been prominently discussed in recent years. In order for automated systems to contribute to any gain in efficiency, safety, or comfort at all, they must be used in accordance with their capabilities. Since trust was found to build an essential psychological basis for how users rely on automated systems, an appropriate level of trust in automation (calibrated trust) is a prerequisite for efficient and safe interaction with these systems. Numerous research projects have been carried out over the past 30 years that investigated the psychological mechanisms of the formation and development of trust in automation. At this point, there is no comprehensive empirically validated understanding of how trust in automated technology is established and how different variable groups affect it. Starting from the assumption of a dynamic and interactive process of trust development, in combining assumptions of the work of Lee and See (2004) and Hoff and Bashir (2015) and extending these with assumptions of theories from other psychological domains (e.g., theories on interpersonal trust and social cognition), this dissertation presents the Three Stages of Trust framework for the investigation of trust development. The basic assumption of the model is that automation trust is formed and calibrated at three successive trust stages. In the first stage the propensity to trust in automation is formed from a combination of more general personality, other dispositions, and the individual learning history with technology and automation. At stage 2 initial learned trust is formed from a combination of the propensity to trust and the interpretation of the information available before the actual use of a specific system. It is proposed that the formation of initial learned trust follows to a large extent the psychological mechanisms of attitude formation and is based on expectations towards the system. Both these expectations and the initial learned trust constitute a starting point for the first reliance decision and the calibration of dynamic learned trust during the interaction with the system at stage 3. At this stage, in a feedback process in which the system expectation and actual system behavior are compared, dynamic learned trust is continuously calibrated. Dynamic learned trust is proposed to influence reliance decisions in regard to the automation essentially. At stages 2 and 3, the framework additionally proposes an influence of moderating variables in both the processing of available information and the relation between trust and reliance behavior. In summary, the framework emphasizes the relevance of three successive trust stages and specifies the psychological processes at these stages in more detail. Building on this, this dissertation presents six studies from four published research articles, that used the framework as a basis for theorizing and research design.

In the first two studies, a hierarchical personality model for predicting both the propensity to trust and dynamic learned trust in a highly automated driving system was explored (study 1) and validated (study 2) in an online setting. Dynamic learned trust was found to be well predicted by the propensity to trust and the prior acceptability of automated driving. These technology-specific dispositions, in turn, were in part explained by technological affinity and dispositional interpersonal trust. Furthermore, the more general personality traits self-esteem, extraversion, and agreeableness predicted variables at the more specific model levels. **Study 3**, in a driving simulator paradigm, examined a mediation of the effect of personality (depressiveness and dispositional negative self-assessments) to dynamic learned trust by the experienced anxiety at the time of the introduction of an automated driver assistance system. In particular, a negative correlation between state anxiety and trust was found. The level of anxiety was predicted by negative trait self-evaluations, especially by low self-esteem and low self-efficacy. These effects of personality differences

on trust were mediated by state anxiety. **Study 4** examined the influence of different types of prior system information on initial learned trust in an online setting. Two types of information were manipulated: system reliability and the system manufacturer (brand reputation). While both types of information affected trust in automation, in line with the assumptions of the Elaboration Likelihood Model (ELM), it was found that the reliability information had a more stable effect on trust and that people with a high need for cognition changed their trust more in the direction of the presented information than people with low need for cognition. Study findings support the proposition of the framework that initial learned trust is built up, at least in part, analog to processes of attitude formation and change. **Studies 5 and 6** focused on the investigation of dynamic trust development by measuring trust repeatedly over time in a driving simulation experiment. First, study 5 examined the temporal development of trust in an automated driver assistance system in two study groups. While in one group, the system worked flawlessly, in the other group the system had a malfunction during the drive. Trust increased over time with error-free system function, decreased temporarily after a system malfunction, and recovered again in the further course of the drive. In study 6, additionally, the provided information on system malfunctions, participants received before using the system for the first time, was manipulated. If drivers were informed of the reasons for malfunctions in advance the temporary trust decrease after an experience of the malfunction during the drive did not occur. These findings support the proposed feedback process at stage 3 in which available information is used to steadily update trust in the system. Additionally, study 6 supports an effect of the information acquired at stage 2 on trust calibration at stage 3.

Taken together, the studies empirically support key propositions of the presented framework and extent the knowledge on the formation and calibration of trust in automation. Above this, the presented framework represents a theoretical advancement in the specification of trust processes before and during the interaction with automated technology. The integration of different models on the formation of trust in automation with fundamental psychological theories supports a broader perspective on psychological processes prior and during the interaction with automated systems in the investigation of trust in automation. On this basis, various research questions, for example, in regard to the interrelation of the different trust stages and the influence of moderating variables, can be derived to further specify and refine the framework in future research endeavors. In the discussion of this dissertation, the theoretical and empirical contribution of the presented framework and the study results are summarized. Furthermore, an agenda for future research to further enhance the understanding of the psychological process of trust development is proposed. In addition, recommendations for the practical implementation of the gained insights into the design of prior information, training, system functions, and user interfaces of automated systems and automated vehicles are discussed. To conclude, the strengths and limitations of this work are discussed.

Zusammenfassung

Automatisierung ist eine der Schlüsseltechnologien des gegenwärtigen technischen Fortschritts. Das automatisierte Fahren bildet einen der Bereiche, in denen dieser Fortschritt in den vergangenen Jahren prominent diskutiert wird. Damit automatisierte Systeme überhaupt zu einem Gewinn an Effizienz, Sicherheit oder Komfort beitragen können, müssen sie entsprechend ihrer Fähigkeiten genutzt werden. Da Vertrauen in automatisierte Systeme eine wichtige psychologische Grundlage für die Entscheidung sich auf ein automatisiertes System zu verlassen bildet, ist ein angemessenes Maß an Vertrauen in automatisierte Technologie (kalibriertes Vertrauen) eine wichtige Voraussetzung für eine effiziente und sichere Benutzung dieser Systeme. In den vergangenen 30 Jahren wurden zahlreiche Forschungsarbeiten durchgeführt, die die psychologischen Mechanismen der Entstehung und Entwicklung von Automationsvertrauen untersucht haben. Aktuell liegt kein empirisch validierter Forschungsrahmen vor, der die Entstehung von Automationsvertrauen und den Einfluss verschiedener Variablengruppen zusammenfassend erklärt. Ausgehend von der Annahme eines komplexen, dynamischen und interaktiven Prozesses der Vertrauensentwicklung, wird in dieser Dissertation ein Rahmenmodell, das Three Stages of Trust Framework, zur Erforschung dieses Entwicklungsprozesses vorgestellt. Das Rahmenmodell verknüpft die Annahmen der beiden Arbeiten von Lee und See (2004) sowie Hoff und Bashir (2015) und erweitert diese mit Annahmen aus psychologischen Theorien aus anderen Bereichen. Die Grundannahme des Modells ist, dass das Automationsvertrauen in drei aufeinander aufbauenden Vertrauensstufen ausgebildet und kalibriert wird. Auf Stufe 1 entsteht die dispositionelle Vertrauensneigung gegenüber Automation unter anderem aus einer Kombination von allgemeineren Persönlichkeitseigenschaften, Einstellungen und der individuellen Lerngeschichte mit Technik und Automatisierung. Auf Stufe 2 entwickelt sich das initial gelernte Vertrauen in ein spezifisches System aus einer Kombination dieser allgemeinen Vertrauensneigung und der Interpretation der vor der eigentlichen Systembenutzung verfügbaren Informationen. Es wird angenommen, dass dieser Prozess weitestgehend den psychologischen Mechanismen der Einstellungsbildung folgt. Das initial gelernte Vertrauen wird maßgeblich durch die initialen Erwartungen an das System beeinflusst. Diese Erwartung und das initial gelernte Vertrauen bilden den Ausgangspunkt für die erste Reliance-Entscheidung und die Kalibrierung von dynamisch gelerntem Vertrauen während der Interaktion mit dem System auf Stufe 3. Auf dieser Stufe wird durch einen Feedbackprozess, in dem die Systemerwartungen mit dem tatsächlichen Verhalten des Systems abgeglichen werden, das dynamisch gelernte Vertrauen kalibriert, welches dann die einzelnen Entscheidungen, sich auf die Automation zu verlassen, maßgeblich beeinflusst. Zudem wird im Rahmenmodell auf Stufe 2 und 3 der Einfluss moderierender Variablen angenommen. Zusammenfassend betont das Rahmenmodell die Bedeutung dreier aufeinander aufbauender Vertrauensstufen und spezifiziert die psychologischen Prozesse auf diesen Stufen näher. Hierdurch können Forschungsfragen und Designs sowie Hypothesen für künftige Studien zur schrittweisen Verfeinerung des Modells abgeleitet werden.

Auf dieser Grundlage werden in dieser Dissertation sechs Studien aus vier veröffentlichten Forschungsartikeln vorgestellt. Die Forschungsfragen der Studien zu Zusammenhängen von Variablen auf den verschiedenen Vertrauensstufen, sowie Einflüssen von personen- und systembezogenen Variablen, wurden aus den Annahmen des Rahmenmodells abgeleitet. **In den ersten beiden Studien** wurde ein hierarchisches Persönlichkeitsmodell zur Vorhersage der dispositionellen Vertrauensneigung in Automation und des dynamisch gelernten Vertrauens in ein hochautomatisiertes

Fahrassistenzsystem in einem Onlineparadigma exploriert (Studie 1) und validiert (Studie 2). Dynamisch gelerntes Vertrauen konnte durch die dispositionelle Vertrauensneigung gegenüber Automation und Voreinstellungen zum automatisierten Fahren substantiell vorhergesagt werden. Diese beiden technikspezifischen Dispositionen wurden wiederum durch Technikaffinität vorhergesagt. Zudem zeigte die Disposition zu zwischenmenschlichem Vertrauen einen Zusammenhang mit der Vertrauensneigung gegenüber Automation. Auf der Ebene allgemeinerer Persönlichkeitseigenschaften sagen vor allem das Selbstwertgefühl, Extraversion und Verträglichkeit die spezifischeren Dispositionen vorher. **In Studie 3** wurde eine Mediation der Zusammenhänge von Persönlichkeitseigenschaften (Depressivität und negative Selbstbewertungen) mit dem dynamisch gelernten Automationsvertrauen durch das bei der Vorstellung eines automatisierten Fahrassistenzsystems berichtete Ausmaß an Angst untersucht. Es fand sich insbesondere ein starker negativer Zusammenhang zwischen der State-Angst und Automationsvertrauen. Hierbei wurde das Ausmaß der Angst durch dispositionelle negative Selbstbewertungen, insbesondere durch ein niedriges Selbstwertgefühl und eine niedrige Selbstwirksamkeit erklärt. Die Effekte der Persönlichkeitseigenschaften auf Automationsvertrauen wurden durch die empfundene Angst mediiert**. In Studie 4** wurde der Einfluss verschiedener Vorabinformationen auf das initial gelernte Automationsvertrauen in einem Onlineparadigma untersucht. Zwei Arten von Informationen wurden variiert: die Systemzuverlässigkeit und das Markenimage des Systems. Entsprechend der Annahmen des Elaboration-Likelihood-Modells (ELM) zeigte sich, dass die Information zur Reliabilität einen zeitlich stabileren Effekt auf das Vertrauen hatte und dass diese Information stärker von Menschen mit einem hohen Need for Cognition genutzt wurde. Die Ergebnisse dieser Studie unterstützen die Annahme des Rahmenmodells, dass initial gelerntes Vertrauen auch durch Mechanismen von Einstellungsbildungsprozessen ausgebildet wird. In den **Studien 5 und 6** wurde die dynamischen Vertrauensentwicklung auf Stufe 3 in einem Fahrsimulationsparadigma untersucht. Zunächst wurde in Studie 5 die zeitliche Entwicklung von Vertrauen in ein automatisiertes Fahrassistenzsystem in einer Bedingung ohne und einer Bedingung mit Systemfehler während der Fahrt untersucht. Das Vertrauen stieg über die Zeit bei fehlerfreier Systemfunktion und sank nach einem Systemfehler temporär ab, erholte sich jedoch im weiteren Fahrtverlauf. In Studie 5 wurde zusätzlich die Vorabinformation zu Systemfehlfunktionen manipuliert. In einer Bedingung, in der die Fahrer im Vorfeld über die Gründe des Auftretens der Fehler informiert waren, blieb der temporäre Vertrauensabfall aus. Diese Ergebnisse unterstützen eine zeitliche Abfolge von Vertrauensstufen und dass erhaltene Informationen auf Stufe 2 die Vertrauenskalibrierung auf Stufe 3 beeinflussen.

Zusammenfassend konnten in den Studien zentrale Annahmen des Rahmenmodells empirisch gestützt werden. Das Rahmenmodell stellt eine theoretische Weiterentwicklung und Spezifikation des Verständnisses der Vertrauensprozesse vor und während der Interaktion mit automatisierter Technik dar. Die Integration verschiedener Forschungszweige zu Vertrauen mit grundlegenden psychologischen Theorien erlaubt die Ableitung einer Vielzahl von Forschungsfragen zum Zusammenspiel der verschiedenen Modellstufen und des Einflusses von Moderatoren in diesem Prozess. In der Diskussion werden die Ergebnisse im Zusammenhang mit dem Rahmenmodell diskutiert, der theoretische und empirische Beitrag der Arbeit eruiert und eine Agenda für zukünftige Forschungsarbeiten vorgeschlagen. Zudem werden Ableitungen zur praktischen Implementierung der gewonnenen Erkenntnisse in der Gestaltung von Vorinformationen, Trainings, Systemfunktionen und Nutzerschnittstellen von automatisierten Systemen, insbesondere in der Anwendung beim automatisierten Fahren, diskutiert. Abschließend werden die Stärken und Schwächen dieser Arbeit diskutiert.

"Trust is the glue of life. It's the most essential ingredient in effective communication. It's the foundational principle that holds all relationships." – Stephen R. Cove

"All the world is made of faith, and trust, and pixie dust." – J.M. Barrie, Peter Pan

1 Introduction

Technology is becoming increasingly sophisticated in recent years and enters more and more areas of private, public, and work-life. Current technology trends that govern this development are besides others artificial intelligence (AI), fast and mobile internet connectivity, and automation (e.g., Marr, 2019). This dissertation addresses the latter topic and contributes to the investigation of the psychological process in which trust in automated technical systems is established and adapted.

In the following, the term *automated system* is used to subsume all systems performing tasks traditionally conducted by a human (e.g., Parasuraman & Riley, 1997), including active data selection, transformation of information, decision-making, and controlling processes (e.g., Lee & See, 2004). This includes "older" automated systems like industrial robots, but also technology like AI-enabled software (e.g., translation software or spell checkers), sophisticated robots for public and private use, and automated vehicles. Automated functioning can be distinguished by *the level of automation,* indicating the extent to which responsibilities in task fulfillment are automated (e.g., Parasuraman et al., 2000). In the early days of automation, tasks were mostly dichotomously assigned either to the automation or the human operator (e.g., "Men Are Better At-Machines Are Better At"-lists, e.g., Fitts, 1951). This rather static task allocation has been discussed in relation to different problems like task decomposition, out-of-the-loop problem, loss of expertise, and complacency (e.g., Hoc, 2000; Dekker & Woods, 2002). In the face of these problems and the increasing capabilities of technology, a substitution of the human by automated systems or a mere coexistence seems dated. Two concepts paying respect to these challenges and allowing for new directions in human-automation interaction are adaptive automation and human-machine cooperation.

As opposed to static automation, *adaptive automation* has the capability to adapt its functioning and automation level dynamically to the task environment or user needs (Parasuraman et al., 1992). Accordingly, the interaction with such systems is substantially more complex than with static systems. Instead of switching an automation on and off, in line with Woods (2001), interaction concepts must account for „intermediate, coordinative modes of interaction, which allow human operators to focus the power of the automation on particular sub-problems, or to specify solution methods that account for unique aspects of the situation" (p. 3). As such complex tasks are increasingly cooperatively accomplished, they call for an extended perspective on the role of the automated system as a teammate (e.g., Klein et al., 2004; Atkinson et al., 2012). Accordingly, the concept of *human-machine cooperation* has gained substantial interest in the last 20 years (e.g., Hoc, 2000; Dekker & Woods, 2002; Young et al., 2007). It describes a mode of interaction in which two agents (the human and the technical system) strive towards a shared goal, interfere with each other in terms of resources and procedures and manage these interferences in order to facilitate their activities towards task completion (Hoc, 2001). The perspective of cooperation between the human and the automated system changes the role distribution, character, and goals of the interaction essentially. From being a mere substitute of the human, automated systems become team players working together with their human partners to find dynamic solutions for complex problems under consideration of the respective skill set (Christofferson & Woods, 2002; Klein et

al., 2004; Dekker & Woods, 2002; Hoc et al., 2009). In line with this change, also the perspective in system design changes from mere reliance and compliance to the consideration of social exchange relationships between the human and the automated system (e.g., Madhavan & Wiegmann, 2007). Building on previous work (e.g., Dekker & Woods, 2002; Christoffersen & Woods, 2002; Hoc et al., 2009; Klein et al., 2004), Walch and colleges (2017) name four requirements for successful human-machine cooperation: mutual predictability, directability, shared situation representation, and calibrated trust.

This dissertation takes a closer look at the nature of trust in automation and aims at a deeper understanding of the psychological processes leading to trust formation and trust calibration in regard to a specific automated system. In line with current human factors models (e.g., Lee & See, 2004; Hoff & Bashir, 2015), trust in automation is conceptualized as a subjective attitude, dynamically evolving over time on the basis of available information. Trust has been found to influence user decisions in regard to automated systems essentially and is under certain circumstances highly predictive for reliance in and use of automated systems (e.g., Muir & Moray, 1996; Lee & Moray, 1992; Biros et al., 2004). A *calibrated level of trust*, accurately reflecting the actual capabilities and performance levels of a system (Muir, 1987), builds an important precondition for safe and efficient system use. Accordingly, the understanding of the psychological processes in trust formation is crucial for the design of systems that are safe, efficient, and enjoyable. At this point, essential questions like how trust in automation is built up, which characteristics of the user, the automated system, and the situation influence the formation of trust in automation, how it is adapted in the course of interaction, and how system design can enhance appropriate trust have not yet been fully answered (e.g., Lee & See, 2004; Walker et al., 2016; Hoff & Bashir, 2015). This dissertation proposes a psychological framework for the investigation of trust in automated systems integrating theorizing from the domains of interpersonal trust and trust in automation. On this basis, six studies are presented that investigated psychological processes of trust formation and calibration before and during the initial interaction with automated systems. All of these six studies were conducted in the domain of automated driving. Nevertheless, the investigated processes can likely be extended to the interaction with automated systems from other domains.

The first two studies investigated the role of several personality traits and technology attitudes for individual differences in trust in automation by exploring a hierarchical personality model in the first and validating it in the second study. In a third study, a mediation of the effects of personality traits on trust in automation by users' state anxiety at the time of being introduced to an unfamiliar automated driving system was investigated. In a fourth study, the utilization of different kinds of information about an automated system presented before the first interaction was investigated based on theoretical assumptions of the Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1986). In studies 5 and 6, the dynamic development of trust over the course of a first interaction with an automated driving system was investigated. Thereby, both studies investigated trust in flawless and erroneous system functioning. In study 6, additionally, the effects of prior information about system malfunctions on trust changes after the experience of such a malfunction during the drive was examined.

Regarding the organization of this dissertation, in the first chapter, interpersonal trust and trust in automation are defined (chapter 1.1), the relevance of considering trust in the design of automated vehicles is derived along with an overview of basic concepts (chapter 1.2), essential theories on trust are introduced (chapter 1.3), and a brief review of variables found to be predictive for trust

in automation is provided (chapter 1.4). Based on this, the aims of this dissertation are further specified (chapter 1.5). In chapter 2, the propositions of the introduced framework are derived from theory. Following on from this, in chapter 3, the empirical studies of this dissertation are presented (chapters 3.1 to 3.4). In the discussion (chapter 4) the study findings are summarized (chapter 4.1), and theoretically integrated along with directions for future research (chapter 4.2). Furthermore, practical implications (chapter 4.3), strengths, and limitations are discussed (chapter 4.4), and a conclusion is drawn (chapter 4.5).

1.1 Trust in Automation

Trust can be generally described as a psychological variable that influences a *trustor's* (the trust-giving agent) willingness to accept risk in regard to a *trustee* (the trust-receiving agent; e.g., Hoff & Bashir, 2015; Mayer et al., 1995; Thielmann & Hilbig, 2015). In early trust research by Deutsch (1960) the theoretical foundation on the role of trust for human cooperation had been established. Later, the focus of interpersonal trust research shifted to an investigation of trust as a dispositional variable reflecting people's general tendency to trust others (e.g., Rotter 1967, 1980) and as a key variable in the development of dyadic relationships between humans (e.g., Rempel et al., 1985). A review of the history of research and empirical findings in the domain of interpersonal trust can exemplarily be found in Simpson (2007).

In the interpersonal context, trust has been defined by Rotter (1967) as an expectation of an individual that he or she can rely on another human being's promise or written communication. Mayer et al. (1995) provide a somewhat more extensive definition of trust as a party's general willingness "to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (p. 712). In their interdisciplinary review on trust Thielmann and Hilbig (2015) identified three common features entailed in most trust definitions: a) *uncertainty* and *risk*, b) an *expectation* in regard to the actions of the interaction partner, and c) acceptance of *vulnerability*.

The basic concepts of trust between humans can be transferred to a broad array of decisions in the interaction with automated systems. Accordingly, the psychological concept of trust was transferred from the domain of cooperation between humans to the interaction of humans with automated systems (e.g., Muir, 1987, 1994, Lee & Moray, 1992; Lewandowsky et al., 2000; Madhavan & Wiegmann, 2007, see chapter 1.3). In the case of human-automation interaction, the trustor is the human user (e.g., driver) and the trustee is the autonomous system (e.g., automated driving system). Uncertainty and subjective risk result, e.g., from the user's incomplete picture of a) the system's technological components, b) their functioning, and c) the inherent mechanisms of the relation between system input and output. Based on the available information, an expectation is built of how the system "is" and how it is likely to behave. Furthermore, the early interaction with such systems is likely to be characterized by a subjective or objective vulnerability of the user as inappropriate reliance might lead to non-optimal task execution or accidents. This overlap of defining features between interpersonal trust and trust in automation makes trust an essential concept for the understanding and prediction of decision making in the interaction with automated systems.

Trust in automation has been investigated for over three decades now, which is reflected in various definitions (see Lee & See, 2004 for an overview). This work builds on the well-recognized definition for trust in automation by Lee and See (2004) "as the attitude that an agent will help achieve

an individual´s goal in a situation characterized by uncertainty and vulnerability“ (p. 51). The underlying understanding of trust as a *specific dynamic attitude* that develops and changes in the course of the familiarization with an automated system puts trust in automation in the role of an intermediate subjective variable that affects decisions in the interaction with such systems.

Early investigations of the relationship between trust in automation and reliance behavior were conducted in process control scenarios in which participants supervised, controlled, and delegated processes in simulated plant micro worlds (e.g., Muir & Moray, 1996; Lee & Moray, 1992, 1994). Muir and Moray (1996), in such studies, for example, found that trust in automation could be meaningfully assessed with subjective scales, that these assessments varied with changes in system reliability and were predictive for reliance behavior (e.g., Lewandowsky et al., 2000). It was further shown that trust in automation is not merely a reflection of automation behavior, but is also affected by task-related self-perceptions (e.g., self-confidence, Lee & Moray, 1994) and learning about the system (e.g., Khastgir et al., 2018). Taken together, trust in automation is conceived as a distinguishable psychological construct influencing the use of automated systems (e.g., Lewandowsky et al., 2000). In support of this claim, since the earlier days of research on trust in automation, there has been a huge deal of research underpinning a relationship of subjective trust in automation with reliance and use of various automated systems (e.g., Biros et al., 2004; Sanchez et al., 2004; Daly, 2002; Hergeth et al., 2016; Rajaonah et al., 2006).

1.1.1 Calibrated Trust and Trust Calibration

As for the described theoretical relevance of trust in automation for safe and efficient decision making with automated systems, it is a goal of both research and practice to investigate and provide directions for facilitating a *calibrated level of trust* in users – a condition in which a user's level of trust is precisely congruent with the system's actual capabilities and performance (Muir, 1987). In such a situation of calibrated trust, an automated system is likely to be used in an adaptive manner that maximizes positive and minimizes adverse outcomes. On the other hand, the more the individual trust level deviates from a calibrated level, the more likely suboptimal reliance patterns in terms of disuse (using an automated system not to its fullest capability) and misuse (using an automated system beyond its capabilities, e.g., Parasuraman & Riley, 1997) emerge. In this regard, Lee and See (2004) linked distrust (trusting a system too little) to system disuse and overtrust (trusting a system over its capabilities) to system misuse. In order to favor the emergence of calibrated trust in the design of automated systems and the corresponding user interfaces, the psychological processes in which trust levels are adapted to the anticipated or experienced capabilities of an automated system need to be considered (e.g., Lee & See, 2004; Khastgir et al., 2018).

Trust is commonly conceptualized as a dynamic variable evolving and adjusting along with available information and accumulated knowledge about a system to correspond to objective trustworthiness of a system, a process that is referred to as *trust calibration* (e.g., Kraus, Scholz, et al., 2019; Khastgir et al., 2018; Muir, 1994). In this regard, calibrated trust is the optimal result of the dynamic process of trust calibration prior to and during the interaction with an automated system. While at the beginning of system interaction, trust was found to change relatively fast with new information and experiences (e.g., Muir & Moray, 1996; Hergeth 2016; Khastgir et al., 2017), over the course of the interaction trust has been found to stabilize if system behavior meets the expectations of the users (e.g., Beggiato et al., 2015). In the process of trust calibration, besides others, system knowledge and the experience of system reliability were proposed to play an important role (e.g., Khastgir et al., 2017, 2018).

Although the psychological processes involved in the formation and calibration of trust in automation have been both discussed on a theoretical level and empirically investigated in the last 30 years, they are not yet sufficiently understood at this point. In the face of the depicted role of calibrated trust in automation, the questions of how it is built up, which information is used to adapt one's level of trust, and how trust in automation dynamically develops over time are essential to be answered for facilitating a safe and from a user's perspective positive interaction with automated systems in their design and introduction.

While a good part of early research on trust in automation was conducted in professional settings in which very specific tasks are to be performed by highly specialized experts (e.g., plant simulations), in recent years, automated systems are more and more introduced into public and private settings. This is especially the case for household robots and automated vehicles, which are used by lay users on a broad basis without systematic training. The question of how to facilitate trust calibration in such systems has significant implications at both the individual (e.g., user well-being) and societal level (e.g., road safety). As we are contemporary witnesses of the transition from manual to automated driving, in which intermediate systems allowing for automated support in a part of the driving task are more and more introduced in current vehicles, the scientific analysis of trust processes in this domain can influence system design and introduction from an early phase of technology dissemination on. Accordingly, the presented studies of this dissertation were set in the domain of highly automated driving. In the following, the essential role of trust calibration and calibrated trust for safe and efficient use of automated driving systems is discussed before theories of trust formation and calibration are introduced (chapter 1.3).

1.2 Trust Calibration in Driver-Vehicle Interaction with Automated Vehicles

Commonly, the advent of automated vehicles is discussed in relation to positive outcomes like efficient, safe, and comfortable driving (Maurer et al., 2016). Before fully autonomous driving can become a reality and these projected positive effects come fully into play, several technological and legal obstacles have to be overcome. Therefore, it will likely take another 10 to 20 years before fully autonomous cars are used on a daily basis (e.g., Beiker, 2016). During this period of transition, intermediate modes of automation are introduced, in which the automation takes over full or partial control temporarily for a specific road type or driving situation it is designed for (e.g., driving on a two-lane rural road). These intermediate modes refer to level 1 (driver assistance), level 2 (partial driving automation), level 3 (conditional driving automation), and level 4 (high driving automation) of the taxonomy of the Society of Automotive Engineers (SAE, 2018), in between level 0 (no driving automation), and level 5 (full driving automation). Automated driving at the SAE levels 3 and 4 provides the driver, depending on the automation level, with the possibility to more or less disengage for a certain time from the driving task. On the other hand, at levels 1 and 2 the driver has to supervise system functioning at all times. While all four intermediate levels are referred to as *automated driving* in the following, levels 3 and 4 are referred to as *highly automated driving*. The functional scope of automated driving systems is restricted by system limitations and malfunctions.

A *system limitation* occurs when a limit of an automated driving system's operational design domain (Czarnecki, 2018; SAE, 2018) is reached, e.g., when the vehicle leaves an area the automation is designed to manage (e.g., change of road type) or the environmental conditions change into a constellation under which the system is not able to drive automatically (e.g., restricted sight). In

the case of a system limitation, the system initiates a *Take Over Request* (TOR; e.g., Gold et al., 2013; Walch et al., 2017), after which control is handed back to the human driver in a transition area. Accordingly, system limitations are intended restrictions to the scope of automated systems. On the other hand, *system malfunctions* refer to "sudden, unpredicted errors related to a system's reliability within its area of application" (Kraus, Scholz, et al., 2019, p. 1) leading to disruption of normal driving and resulting for example in a fallback strategy like a safe halt (e.g., Emzivat et al., 2017). As in these situations the control of the complete driving task or subtasks has to be handed over, the intermediate modes of automated driving reflect a cooperative task-sharing between the driver and the automation during the drive (Woods, 2001). Accordingly, as the drivers are commonly assumed to be still in charge of function allocation in driving automation at intermediate automation levels (e.g., Beiker, 2016; Hoc et al., 2009), drivers are facing complex decisions during the drive. These include if and to what extent they use the automation, continue to monitor it, and if, when, and to what extent they regain control. Above this, the interaction between the driver and the driving automation might include training of the automation, collaborative maneuver planning, and cooperative resolution of uncertainties.

In the face of this complexity and the associated dangers of inappropriate system use (as discussed in chapter 1.1.1), it is essential to minimize both the negative effects of misuse and disuse of automated driving systems in their design and introduction. Disuse may, for example, arise if drivers do not feel comfortable with the technology and decide not to use it. In this case, the positive effects of advantages commonly associated with automated driving, like reduced CO2 emissions and increased road safety, do not fully come into play. On the other hand, misuse might result if drivers are too prone to the technology and use it excessively, which is dramatically illustrated by recent accidents with automated vehicles that have not been used in an appropriate manner (e.g., Wang & Li, 2019). Accordingly, if one aims at pursuing the goal of a safe and sustainable introduction of automated driving systems (e.g., by user education), it seems meaningful to direct one's attention to psychological processes leading to the drivers' decisions if, when and how to use the driving automation.

Like in other domains, also in automated driving trust in automation has been found to be an essential predictor for the reliance and use of automated driving systems (e.g., Payre et al., 2016; Hergeth et al., 2016; Rajaonah et al., 2006). The importance of trust for decision making in the interaction with automated driving systems is further reflected by an ever-growing body of literature on trust in the domain of automated driving systems in recent years (e.g., Hergeth et al., 2017; Beggiato et al., 2013, 2015; Hartwich et al., 2019; Frison et al., 2019; Körber et al., 2018; Ekman et al., 2018; Waytz et al. 2014). In this regard, it is an essential research focus of how trust calibration in drivers before and during the drive can be facilitated. For example, in regard to trust processes prior to the drive, the role of prior information about a system's capabilities has been investigated (e.g., Körber et al., 2018; Beggiato et al., 2013). In the same manner, the role of prior experience and practice with automated driving systems have been addressed (e.g., Gold et al., 2015). In regard to processes during the drive, the dynamic trust development over time has been investigated for short (e.g., Hergeth et al., 2015, 2016) and longer periods of interaction with automated driving systems (Beggiato et al., 2013, 2015). Along the same line, the effects of TORs and system limitations on trust calibration have been investigated (e.g., Hergeth et al., 2015, 2016).

Additionally, some studies investigated how provided information during the drive (e.g., Wintersberger et al., 2017; Helldin et al., 2013) and different design characteristics of the respective user interface affect trust during the drive with automated vehicles (e.g., Frison et al., 2019).

At this point, like in general in the investigation of trust in automation, essential questions like how trust in automated vehicles is dynamically built up, which variables affect this process and how trust calibration and calibrated trust can be facilitated in the introduction, design, driver training, and associated user interfaces of such systems remain not sufficiently answered. Against this background, this dissertation aims at enhancing the theoretical foundation for investigating trust in automation by presenting a psychological framework for the investigation of the formation and calibration of trust in automation. On this basis, six studies are presented that investigated several of the framework's propositions in the setting of automated driving. To gain an understanding of the psychological processes and influencing variables hypothesized to be involved in the formation and calibration of trust in automation, in the following, essential theories of trust and trust formation from the domains of interpersonal trust and trust in automation are introduced.

1.3 Theories of Trust and Trust Formation

The conceptual proximity between trust in automation and interpersonal trust has been the subject of scientific discussion since the beginning of research on trust in automation. Most researchers see similarities between these two forms of trust but differ in regard to the extent of the attested comparability (Hoff & Bashir 2015; Lee & Moray, 1992; Jian et al., 2000; Lee & See, 2004). In favor of the similarity hypothesis, people were found to treat computers and technological systems as social objects and apply social rules from human interaction to the interaction with these systems (e.g., computers are social actors; e.g., Nass et al., 1994; Nass & Moon, 2000). On the other hand, essentially, the character of available information about the interaction partner (e.g., language structures in human partners) as well as the applied cognitive schemas were discussed to be different in human-human and human-automation trust (Madhavan & Wiegmann, 2007). In the same manner, also individual differences that affect information processing in the two trust domains were discussed to vary (e.g., in terms of self-evaluations; Lewandowsky, 2000). In particular, it was also discussed from early on that the temporal order of available information to inform one's trust over the course of the individual phases of trust formation differs in the two domains (Lee & Moray, 1992). Furthermore, in human-machine interaction, characteristics and trust-relevant processes on the side of the trustee are easier to objectivize (e.g., restricted subjectivism on the side of the machine). For example, in the interaction with automation, social projection (e.g., Thielmann & Hilbig, 2015) and the intentionality of the trustee (Lee & See, 2004) are restricted. In conclusion, while the basic psychological processes of trust in other humans and trust in automated systems are assumed to share essential underlying mechanisms, the considered information, attributional processes and the process of trust formation over time are likely to differ. However, while specific theories and models of trust in automation are essential to account for these and other differences, theorizing on trust in automation originated in the theoretical work on human-human trust. Thereby, especially, the models proposing different levels of trust attributions from the interpersonal domain (e.g., Rempel et al., 1985) and the models by Lee and See (2004) and Hoff and Bashir (2015) have been influential and build an important basis for this dissertation. In

the following, these three models are depicted in more detail and open research questions are identified. For an additional overview of theoretical work on trust in automation, see French et al. (2018) and Walker et al. (2016).

1.3.1 Theories on Levels of Trust Attributions

Early theories on the formation of trust in automation built on the model by Rempel et al. (1985) from the domain of dyadic relationships between humans. An underlying assumption of this model is that trust as a general expectancy of certain events evolves in a dynamic process of uncertainty reduction over the course of relationship development (Holmes & Rempel, 1989). In the model by Rempel et al. (1985) it is proposed that in the course of this process, the information basis of trust involves three increasing levels of attributional abstraction.

At level 1, *predictability*, the extent to which the behavior of the partner can be anticipated, provides the basis for trust (Rempel et al., 1985). This anticipation is influenced by knowledge on reward contingencies, stability, and control over the behavior. At level 2, the reasons behind the partner's behavior come into focus, which are subsumed under the term *dependability*. At this level, rather than from discrete behavior, trust is inferred from an attribution of a partner's personality in regard to cooperativeness, reliability, honesty, and benevolence. A partner's dependability is evaluated over time on the basis of accumulated experiences in repeated encounters with the partner in situations involving personal risk and uncertainty. At level 3, especially in the later progress of a relationship, trust is additionally proposed to be built on *faith*. This level entails an anticipation of future behavior of the interaction partner and was defined as the general "belief that one's partner will act in loving and caring ways whatever the future holds" (Rempel et al., 1985, p. 109). Taken together, according to the model by Rempel et al. (1985), the sources for inferences about trust in a partner evolve and broaden in the course of a relationship from behavioral observation via a derivation from perceived dispositions to the projection of future behavior. The model entails three essential characteristics of trust processes: a) trust develops dynamically over time on the basis of b) available information about the interaction partner, whereby c) the characteristics of the utilized information tend to change over the course of relationship building with the partner.

The model by Rempel et al. (1985) was transferred to the area of human-machine trust by Muir (1987, 1994). She proposed that the assessment of predictability towards automated systems builds on the objective behavior of the automated system, the user's ability to predict it, and the stability of the environment. For the second level, Muir argued that dependability is built through experience with the automation and testing the limits of the system in uncertain and risky situations leading to less need for monitoring of the automation. Third, faith in the domain of human-automation interaction was proposed to refer to the incomplete understanding users still have after they have been using the system for a longer period of time as the underlying technological processes are in most cases too complex to be fully understood. In this regard, faith builds on a perception of appropriateness and flexibility of the programming of the system in a way that the system's dependability is projected to unforeseeable future interactions.

Also, in the tradition of Rempel's theory, Lee and Moray (1992) and based on Muir's work defined four dimensions of attributional abstraction in trust processes specific for the interaction with automated systems. The authors adapted the naming of these dimensions to represent the conditions of human-machine interaction. The first dimension of trust is proposed to be formed by a general assumption that natural and social laws apply. The second dimension of trust, *performance*

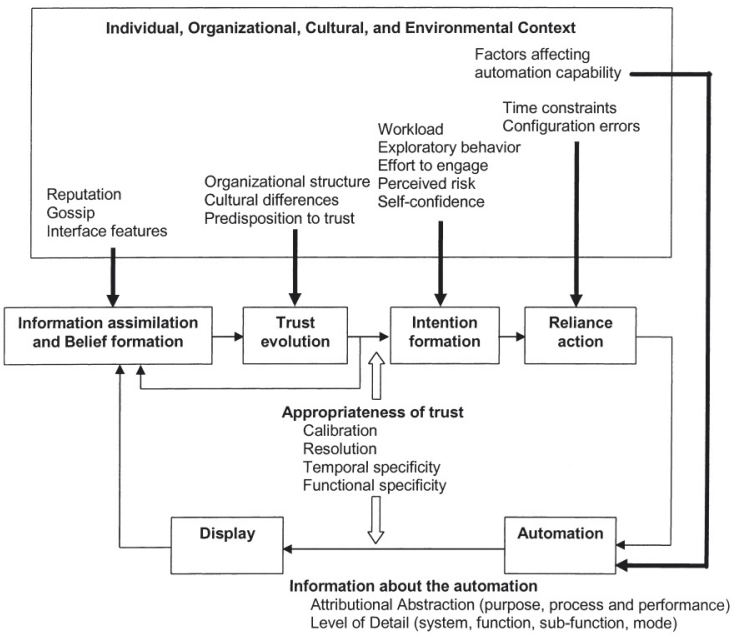
(adapted from predictability), is related to what the system does and an expectation of its "consistent, stable, and desirable performance or behaviour" (Lee & Moray, 1992, p. 1246). Third, *process* (from dependability) refers to an understanding of the mode of operation of the system involving, for example, sensors and algorithms and their appropriateness for the situation and goal attainment. Fourth, *purpose*, adapted from the dimension faith, covers the motives and intentions of the system designer and how the system is intended to be used. While the first level is not mentioned again in further discussion of the trust levels in Lee and Moray's paper, the remaining three trust levels provide a good foundation for the various attributions from system behavior to the evaluation of trustworthiness. For example, Lee and See (2004) used these three levels to categorize a broad array of different proposed trust bases from trust theories from different domains. While Lee and Moray (1992) do not further specify how their proposed trust dimensions are interrelated, they state that they do not reflect a sequential order. This is important as the label "levels" of the attributional trust models suggests a timely order of these processes, although the levels by Rempel et al. (1985) were not proposed to be necessarily run through in hierarchical order nor to be mutually exclusive. Rather, they reflect "a developmental progression in terms of the time and emotional investment required to establish each component and in terms of the level of attributional abstraction each demands" (Rempel et al., 1985, p. 98). Therefore, the levels proposed by Muir (1989, 1994) and Lee and Moray (1992) do not per se reflect a process over time in the relationship building between users and automated systems, but rather describe different attributional abstractions in the process in which available information is used to build up trust. In the following, the two models by Lee and See (2004) and Hoff and Bashir (2015) are introduced. These models entail further assumptions on the dynamic character of the development of trust over the course of getting to know and familiarization with automated systems.

1.3.2 Lee and See's Trust in Automation Model

Lee and See (2004) integrate definitions, concepts, and research on trust from different (interpersonal, psychological, sociological, and organizational) domains into an influential conceptual

Figure 1

Lee and See's (2004) conceptual model for trust in automation, illustrating the proposed dynamic feedback process of trust in automation. (Figure from Lee & See, 2004).



model for explaining the process of dynamic trust formation in the interaction with automated technology (see Figure 1). Based on their definition of trust as an attitude, the authors transfer the *theory of reasoned action* (TRA; Ajzen & Fishbein, 1980) to the domain of automation trust. The TRA builds on the assumption that people exert some kind of control over their behavior and that behavior is essentially influenced by the intention to perform this behavior, which in turn is formed on the basis of an integration of available information (e.g., Ajzen, 1991). Thereby, a behavioral intention refers to "the motivational factors that influence a behavior; […] indicators of how hard people are willing to try, of how much of an effort they are planning to exert, in order to perform the behavior" (Ajzen, 1991, p. 181). Accordingly, it is proposed that the higher the intention to engage in a certain behavior, the more likely it is to result in the actual performance of this behavior (Ajzen, 1991). Based on the assumption of behavioral control, the TRA further proposes that behavior is influenced through the combined influence of norms, beliefs, attitudes, and intentions. Beliefs can be defined as subjective probabilities of the link between particular attributes and a given object (e.g., a system), which are derived from salient information and, in turn, influence attitudes (Albarracín & Chan 2018; Ajzen, 1991). Attitudes can be defined as "a summary evaluation of a psychological object captured in such attribute dimensions as good-bad, harmful-beneficial, pleasant-unpleasant, and likable-dislikable" (Ajzen, 2001, p. 28). Accordingly, in comparison to beliefs, which are specific for the information available in the situation, attitudes are rather integrative evaluations and compared to beliefs less fluctuating and more robust across situations. Lee and See (2004) simplify and adapt the proposed interrelations of the TRA in their model to explain trust-based behavioral decisions in a closed-loop dynamic decision-making process in the interaction with an automated system (see Figure 1), in which "the dynamic interaction with the automation influences trust and trust influences the interaction with the automation" (Lee & See, p. 68). This process is hypothesized to be affected by contextual factors in the environment, characteristics of the automation, and user characteristics. A further specific role is attributed to the content and format of the information that is displayed in the user interface during the interaction.

Lee and See's model represents an important step toward a foundation of the theorizing about psychological processes in the formation of trust in automation. Thereby, the model's notions that trust evolves dynamically and that in this process, various factors play together in their effects on trust formation are of specific importance for a more complex approach to the phenomenon. The article by Lee and See (2004) in which the model was first introduced is often cited in the field (especially for the attitude-based trust definition) and has been influential in the scientific discussion on trust in automation since its publication. However, very few studies have been published that empirically tested the main assumptions of the model. Furthermore, the model awaits further theoretically advancement and specification. Also, it inherited some of the criticisms of the TRA and the theory of planned behavior (PLABE theory; Ajzen, 1988, 1991) to which the TRA was extended and which shares the basic assumptions Lee and See entailed in their model. Both the TRA and the PLABE theory have been criticized, for example for their overly restrictive rational perspective on decision making not sufficiently accounting for affective influences (e.g., Conner et al., 2013), unconscious or implicit motivations and direct effects of variables on behavior when controlling for the included predictors (e.g., Sniehotta et al., 2014). While the model of Lee and See (2004) accounts for some of these influences (e.g., affective processes), it still yields empirical confirmation and might benefit from further specification. First, the included variables and the

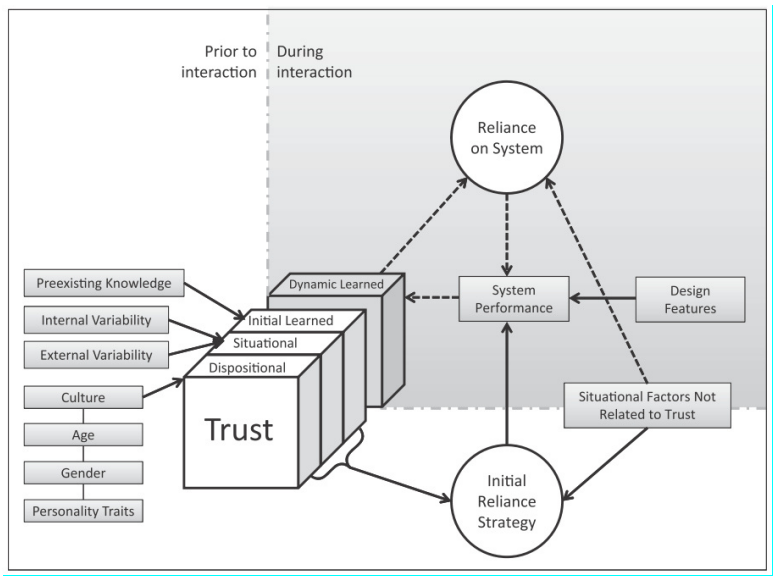
psychological processes, in which they relate to trust, are in part not sufficiently defined and specified. In this regard, the underlying definition of trust as an attitude could be built on to further provide specifications. Second, the authors conceptualize trust as a history-based construct but do not provide further details on the trust processes taking place before the interaction with an automated system. Third, while the model proposes a general influence of individual differences, there is no further specification of how user dispositions affect trust processes. Fourth, also an integration of different trust layers (Marsh & Dibben, 2003; Hoff & Bashir, 2015) into the model seems fruitful.

1.3.3 Trust Layers

From the theories of trust attributions, it can be concluded that the information and processes on which trust in automation is based change over time during the familiarization with a system. To understand the character of trust at different points in the trust process, it is meaningful to differentiate between several trust layers. Based on a review of 127 studies on factors influencing human-automation interaction, Hoff and Bashir (2015) propose four interdependent "trust layers" and different categories of variables affecting trust at these layers (see Figure 2). First, in line with Marsh and Dibben (2003) as well as Merritt and Ilgen (2008), dispositional and learned trust can be differentiated. Unlike *learned trust* which is built up on the basis of the experience with a specific trustee (e.g., person or automated system), *dispositional trust* was conceptualized as "a diffuse expectation of others' trustworthiness […] based on early trust-related experiences" (Merritt & Ilgen, 2008, p. 195). While Hoff and Bashir (2015) do not further differentiate, a distinction between the more general tendency to trust in other people (e.g., Rotter, 1967, 1980; Gurtman, 1992), and the tendency to trust in automated systems (e.g., Merritt & Ilgen, 2008) seems essential. Both variables are conceptualized as stable personality traits referring to the individual disposition to trust in other entities. As there is some disagreement about the naming of these constructs in the literature and to prevent confusion, the tendency to trust other people is referred to as *dispositional interpersonal trust* in the following (propensity to trust, e.g., in Frazier et al., 2013 and Mayer et al., 1995). Dispositional interpersonal trust has been defined as "a generalized expectancy held by an individual that the word, promise, or written statement of another individual or group can be

Figure 2

Model by Hoff and Bashir (2015) differentiating four trust layers. (Figure from Hoff & Bashir, 2015).



relied on" (Rotter, 1980, p. 1) and more simply as "a general willingness to trust others" (Mayer et al., 1995, p. 715). On the other hand, the more specific dispositional trust in automated systems is referred to as the *propensity to trust in automation* (in line with Merritt & Ilgen, 2008; dispositional trust in Hoff & Bashir, 2015), which was defined as "an individual's overall tendency to trust automation, independent of context or a specific system" (Hoff & Bashir, 2015, p. 413). While learned trust is built up over the course of experiences with a specific automated system, dispositional trust in automation is a disposition of the user general for automated technology, which he or she brings into the situation of getting to know a specific system. The authors state that dispositional trust in automation is built up by "both biological and environmental influences" (Hoff & Bashir, 2015, p. 413) and name general personality traits, culture, age, and gender as possible influencing variables. The nature of the propensity to trust in automation is discussed in more detail in chapter 2.

Learned trust is specific for a system, based on the individual learning history with this system (Merritt & Ilgen, 2008), and "represents an operator's evaluations of a system drawn from past experience or the current interaction […] directly influenced by the operator's preexisting knowledge and the automated system's performance" (Hoff & Bashir, 2015, p. 420). Learned trust can be conceptualized as a specific, history-dependent, dynamic attitude regarding a system's trustworthiness as defined in chapter 1.1 (Lee & See, 2004). Hoff and Bashir (2015) further differentiate between initial learned trust and dynamic learned trust. *Initial learned trust* represents the user's evaluation of the system drawn from knowledge about the system available prior to the interaction, including experience with similar technology, the reputation of the system, and subjective expectations (Hoff & Bashir, 2015). Thus, initial learned trust is assumed to determine the initial reliance strategy. *Dynamic learned trust,* on the other hand, results from the sum of experiences with a particular system during its use up to the specific moment of measurement and covers the fluctuations of trust in correspondence with system performance. Also, initial and dynamic learned trust are further discussed in chapter 2.

As a fourth trust layer, Hoff and Bashir (2015) name *situational trust* which covers situation-specific trust differences evoked by user states (internal variability), characteristics of the automation and the task (external variability) as well as the variances of the relation between trust and behavior induced by such variables. Situational trust, thus, is specific for a particular situational context of the interaction and varies along with these situational influences.

Taken together, the work by Hoff and Bashir (2015) constitutes an essential integration of research on factors influencing trust in automation and provides a detailed research agenda for the different trust layers. Up to now, similar to the Lee and See's (2004) model, there were not yet many attempts to provide empirical support for the proposed model's assumptions. While the authors propose a basic structure of the processes relating the four layers of trust (see Figure 2), their work focuses on providing an overview of the variables influencing trust at the different stages. In this regard, the psychological processes along which trust evolves along the different trust layers might be further specified in the future. First, the different trust layers appear somewhat underspecified with regard to their definitions and psychological causation. Second, the interrelationship of the different trust layers is not substantially specified (e.g., the influence of dispositional trust to learned trust). Third, some of the depicted relations between variables seem restrictive (Figure 2, e.g., personality might directly affect dynamic learned trust). Fourth, the process in which context factors influence the trust layers is not explained in detail and requires further specification. Fifth,

it is unclear whether the influencing factors summarized under the term situational trust constitute an own trust layer or if this is a subsuming term for situational variables influencing the formation of dynamic learned trust (external variability) that might moderate the relationship between dynamic learned trust and reliance (internal variability and task environment). Sixth, an integration of the proposed trust layers with the model of Lee and See (2004) might constitute a substantial step forward in theorizing on trust in automation. To provide an empirical basis of which factors and which information plays a role in the psychological process of trust calibration, in the following, several variables that have been investigated in their effect on trust in automation are reviewed.

1.4 Review on Factors Influencing Trust in Automation

Trust is dynamic and affected by factors on the side of the system, the person, and the situation. This complexity can be understood by taking a look at the meta-analyses by Schaefer et al. (2016), Schaefer et al., (2014) and Hancock et al. (2011) as well as the reviews by Lee and See (2004) and Hoff and Bashir (2015), subsuming factors influencing trust in automation. To provide a picture of the complexity of the processes involved in trust formation and calibration, in the following, a brief review of some of the variables that have been studied in their relation to trust processes in human-machine interaction along the three categories *user-related, system-related*, and *situation-related* is provided. Thereby, under situational variables both environmental factors (e.g., task characteristics, physical variations) and *user states* as situation-specific variations in the user's adaptation to situational circumstances (Hamaker et al., 2007) are subsumed. User-related variables, on the other hand, refer to relatively stable differences between users that are assumed to influence trust in automation (besides others personality traits and general technology-related attitudes).

1.4.1 Characteristics of the User

In regard to attributes of the user investigated in relation to trust in automation, exemplarily, findings on demographic variables, working memory, and user domain expertise are briefly discussed. As for the emphasis of the presented studies on personality traits and technology-related attitudes, these are discussed in more detail. First, demographic variables like age and gender have been repeatedly linked to trust in automation. At this point, while some age differences were reported (e.g., Donmez et al., 2006), these seem to variate between contexts (Hoff & Bashir, 2015). Recently, Hartwich et al. (2019) investigated age group effects for trust in an automated driving system and did not find any differences in the considered groups. Also for gender, no general effects on trust in automation can be reported at this time (Hoff & Bashir, 2015). Second, Rovira et al. (2017) found higher working memory to be associated with more appropriate trust calibration. This initial finding points into the direction that trust calibration might be cognitive demanding under certain circumstances, which entails implications for the design of both user education and user interfaces. Third, higher domain expertise (e.g., automated driving in general) was repeatedly found to result in less reliance on automated systems (e.g., Sanchez et al., 2011). For such a relationship trust might be a mediating variable, but up to now, this has not yet been conclusively investigated. Fourth, several general personality traits and more specific individual dispositions have been repeatedly investigated in relation to differences in trust in automation. Within the group of more general personality traits, the traits of the Five-Factor Model of Personality (e.g., McCrae & John, 1992; Costa & McCrae, 1992) have been repeatedly investigated in relation to trust. Extraversion has been well-established to be positively related to dispositional interpersonal trust

(e.g., Gaines et al., 1997; Evans & Revelle, 2008). This positive relationship was found to generalize to the propensity to trust in machines (Merritt & Ilgen, 2008). For neuroticism, a negative relation to interpersonal trust was found (Evans & Revelle, 2008). While at this time, there is no direct evidence for a link of neuroticism to trust in automation, a reported relationship between neuroticism and the agreement with an automated advice system (Szalma & Taylor, 2011) points into the direction of such a relationship. Third, agreeableness and consciousness were found to be positively related to the propensity to trust in automation (Chien et al., 2016). Regarding further general traits, on a theoretical level, also dispositional interpersonal trust has been related to trust in automation (e.g., Madhavan & Wiegman, 2007), but up to now empirical evidence for such a relationship has not been reported. Within the group of more specific personality traits, the propensity to trust in automated systems (Merritt & Ilgen, 2008) has been found to be positively related to trust in automation. Similarly, also general automation-specific attitudes have been investigated in regard to trust in automation. In this respect, implicit positive attitudes towards automation have been found to predict trust in automation (Merritt et al., 2013). Furthermore, recently several studies investigated the role of a *perfect automation scheme* (PAS) for trust in automation (Merritt et al., 2019; Merritt, Unnerstall, et al., 2015; Lyons & Guznov, 2019), which was first proposed by Dzindolet et al. (2002) and later conceptualized as "a tendency to view automation as perfectly performing or infallible" (Merritt et al., 2019, p. 4). The PAS consists of the two separate factors *high expectations* for the performance of automated systems and *all-or-none thinking*, a belief that automated systems either work flawlessly or not at all (Merritt, Unnerstall, et al., 2015). For the factor high expectations Lyons and Guznov (2019) reported in three studies significant positive relationships between the factor high expectations and trust in the investigated automated systems (robots and an automated flight assistance system). On the other hand, as for a considerable theoretical overlap between the high expectations factor and the propensity to trust in automation, the incremental validity of this factor has to be further investigated. In Merritt, Unnerstall, et al. (2015) this theoretical overlap was supported in a hierarchical approach indicating that the factor high expectations did not account for additional variance explanation in trust in automation over the propensity to trust in machines. For the factor all-or-none thinking findings are inconclusive (Merritt, Unnerstall, et al., 2015; Lyons & Guznov, 2019) and rather point into the direction that it moderates the effect of system malfunctions on trust in automation.

The example of the PAS illustrates that dispositional variables are related to trust in a complex manner, and besides direct relationships, individual dispositions are likely to moderate some of the relationships between characteristics of the automation or situational variables and trust in automation. Also, at this point, only a limited number of studies have been conducted that relate personality variables and other individual characteristics to trust in automation. Furthermore, in some cases, there are inconsistencies in regard to the effect sizes and direction of effects of single personality traits and attitudes on trust in automation. These inconsistencies might be based on differences in the studied systems, the context of interaction, and the task.

1.4.2 Characteristics of the Automated System

The variables in the design of automated systems, which have been studied to affect trust in automation, can be divided into capability-related and appearance-related variables. Capability-related variables include the type of the system, its features, and functionality, as well as the reliability of task performance. Appearance-related variables include all features of the user interface and of the design of the technological device itself. In terms of capability-related variables, system reliability

is one of the most frequently investigated in relation to trust in automation (e.g., Parasuraman & Riley, 1997). In this regard, malfunctions of automated systems were found to result in declines in trust in the systems (e.g., Lee & Moray, 1992; Muir & Moray, 1996; Dzindolet et al., 2003), which, presumably, to some extent, results from the perception of lower system reliability (e.g., Madhavan & Wiegmann, 2007). Reliable behavior and communication of automated systems, on the other hand, were found to be positively related to trust across a broad array of tasks and system types, while at the same time, also possible moderations by task characteristics (e.g., task difficulty) have been discussed (see Schaefer et al., 2016 for an overview). Furthermore, there seems to be no simple relationship between automation level and trust (Rani et al., 2000; French et al., 2018). Irrespective of the automation level, users were found to trust systems more that provide some extent of control (e.g., Hoff & Bashir, 2015) and lead to predictable outcomes (Schaefer et al., 2016). For a further review of capability-related variables, see Hoff and Bashir (2015) and Schaefer et al. (2016).

On the side of system appearance, variables related to system aesthetics like the degree of anthropomorphism (Waytz et al., 2014) and communication style of the automated system (e.g., voice type; Large & Burnett, 2014) have been found to influence trust in automation. Only recently, measuring trust and distrust as two dimensions, Frison et al. (2019) found that a system with higher usability was trusted more while a more aesthetic system was less distrusted. Apart from that, as the investigation of specific design elements (regarding surface characteristics like aesthetics) is not likely to lead to generalizable design guidelines for trustworthy interfaces for automated systems, the investigation of broader psychological functions of the user interface might be more promising. Two related examples of user interfaces' psychological functions that have been prominently discussed in relation to trust in automation are *transparency* and *feedback*. Transparency refers to the extent to which a system's processes and decisions are accessible by the user (e.g., Seong & Bisantz, 2008) in a way that enhances system comprehensibility and predictability, in turn supporting trust calibration (e.g., Kraus, Scholz, et al., 2019). System transparency might either be enhanced by providing prior information about system functioning or by an implementation in the user interface. Exemplarily, Körber et al. (2018) found prior information with regard to the trustworthiness of an automated driving system to influence trust ratings in the interaction with the system. Furthermore, Beggiato et al. (2013) showed that prior information about the limitations of an automated driving system affected trust prior to the interaction and during repeated interaction with the system over several weeks. Also, in regard to transparent user interfaces, initial support for an association of higher transparency with increased trust levels has been presented (Ekman et al., 2016). Above this, the character of feedback - e.g., feedback accuracy, amount, and type - has been found to influence trust development (Schaefer et al., 2016). Variables like system transparency or system feedback might be implemented in a manifold of different ways within a system (e.g., transparency through graphical or speech interfaces), which affect trust in automation in similar ways. Furthermore, it is likely, that design variables interact in their effect on trust in automation with other system- or person-related factors and, thus, to better understand the interactive dynamics of trust calibration a theoretical approach might account for interactions between these variable groups (e.g., different reactions of different user groups to the same interface).

1.4.3 Characteristics of the Situation, Task Environment

Besides the discussed variables on the side of the user and the system, also the role of situational variables should be considered under a dynamic perspective of trust formation. In this regard, Hoff

and Bashir (2015) name task difficulty, workload, attentional capacity, self-confidence, perceived risks and benefits, organizational setting, task framing, and mood as factors that could affect trust. Here, exemplarily, workload, self-confidence, and emotional states are discussed in more detail.

First, some studies report a role of workload for the formation of trust in automation. For example, Biros et al. (2004) found task load to moderate the relationship between trust and automation use. Phillips and Madhavan (2011) found that different types of distractor tasks affected trust in an automated x-ray aid. Second, self-confidence, the situational anticipated self-assessment of one's own ability to successfully perform a task (e.g., Briggs et al., 1998), has been repeatedly investigated in relation to trust in automation. Several authors found the relative ratio of trust in automation and self-confidence (as reflected in difference scores) to predict task allocation (e.g., Lee & Moray, 1994; Lewandowsky et al., 2000; de Vries et al., 2003). For example, Lee and Moray (1994) found in a semi-automated pasteurization plant simulation that users prefer automated control if trust exceeded self-confidence and in the opposite case prefer manual control. Third, only few studies have investigated the influence of emotional user states on trust in automation. Merritt (2011) found a positive relationship between happiness and trust in an automated x-ray system. Similarly, Stokes et al. (2010) reported higher trust in participants with positive than with negative affect while this effect was strongest at the beginning and disappeared in the course of the interaction. In line with this, Frison et al. (2019) found a negative relationship between negative affect and trust in an automated driving system.

In summary of the review of empirical findings on factors influencing trust in automation, it can be stated that while many studies have been conducted and various effects have been reported, at this time, no clear picture of generalized effects over different automated systems, tasks, and situations can be provided. This is presumably a result of a complex interaction of the different variable groups in the formation of trust in a specific automated system. In this regard, enhancing the understanding of the underlying psychological trust processes seems fruitful to increase insight into this interplay of the different variable groups in their effect on trust in automation.

1.5 Problem Statement and Aims of this Dissertation

Trust in automation has been discussed and empirically supported to play an essential role in the safe and efficient interaction with automated systems in general and automated driving systems in specific. Early theories on trust levels (Rempel et al., 1985; Muir, 1994; Lee & Moray, 1992) and the models by Lee and See (2004) and Hoff and Bashir (2015) constituted essentially to an understanding of the nature of formation and calibration of trust in automation. Up to now, these streams of theorizing have mainly been discussed separately, and no integrative research framework reflecting the complex, history-based, and dynamic nature of trust has yet been introduced. Such an integration and further specification of psychological processes in trust formation might help to cope with the complexity of the factors contributing to trust in automation in research (as illustrated in chapter 1.4). At this point, several open questions on how different trust layers are interrelated, which processes are involved in their formation and calibration, and how the identified variable clusters feed into these processes are still pending answers. First, the role of dispositional trust (both in its generalized and automation-specific form) in the perception of unfamiliar automation systems has not yet been conclusively understood. Second, in this regard, it is an open question how dispositional interpersonal trust relates to dispositional trust in automation. Third,

the interplay of different trust layers, as proposed by Hoff and Bashir (2015), has not been investigated in detail. Fourth, it is unclear how predispositions resulting from personality differences and individual experiences in learning history influence the use and processing of information provided before users interact with an automated system for the first time and in the course of system interaction. Fifth, the question of how initial learned trust informs dynamic learned trust and decision-making with automated systems is yet unanswered. Sixth, at this point, it is an open question how dynamic learned trust evolves along with the observed system behavior over time and how the perception of automation reliability is interpreted against information the users received prior to the interaction with a system. Seventh, the understanding of the role of communicated or observed system characteristics and situational variables in the establishment of learned trust remains restricted. While these research questions only subsume some of the open questions in the formation and calibration of trust in automation, they constitute the motivation for this work.

Against this background, this dissertation aims at presenting: 1) an integrative psychological research framework for investigating the dynamic formation and calibration of trust in automation, and 2) a series of studies in the domain of highly automated driving to investigate some of the relationships and specific variables proposed to affect trust in automation.

1.5.1 Aims of the Proposed Framework for Trust in Automation

There is a general consensus in the field that trust is built in a dynamic process over time (Lee & See, 2004) in which different variables interact with each other. While two influential papers in recent years have essentially contributed to theoretically describe the process of trust development in automated systems (Lee & See, 2004; Hoff & Bashir, 2015), their assumptions have not yet been integrated into a conclusive model nor sufficiently empirically validated. Moreover, these models still require further specification of the psychological processes of trust development. Starting from the assumption of a dynamic learning process of trust development in which different groups of variables interact, this dissertation presents a framework for the investigation of the process of trust development integrating the trust layers by Hoff and Bashir (2015) and the trust processes proposed by Lee and See (2004). The proposed framework specifies three stages at which trust is formed along three trust layers (propensity to trust in automation, initial, and dynamic learned trust) that sequentially build on each other. Furthermore, the complexity of influencing factors, as depicted in chapter 1.4, is integrated into the framework. Additionally, the psychological processes in which the trust layers are proposed to be formed at the three stages are further specified on the basis of concepts and models from interpersonal trust, personality psychology, and social-cognitive theories on attitude formation. The presented framework is specified along ten propositions, integrating and extending the state of theorizing on trust in automation. The framework aims at providing a theoretical basis for the integration of research on trust in automation and serves as a heuristic for the derivation of hypotheses and research designs in order to test, falsify, and further specify the entailed propositions on the psychological processes playing a role in the formation of trust in automation.

1.5.2 Aims of the Empirical Studies

On the basis of the proposed framework, a series of six studies from four published research articles in the area of automated driving is presented. To contribute to closing the described research gaps, these studies focused on an investigation of trust formation processes as specified in the propositions of the framework. More specifically, they addressed interrelations between the dif-

ferent trust layers, the impact of available information (prior to system use and during the interaction) on trust formation and calibration and the role of several dispositional and state variables in trust development. These studies are depicted in more detail after a presentation of the research framework.

2 Research Framework: Three Stages of Trust in Automation

The *Three Stages of Trust framework* (see Figure 3) integrates variables and processes that have been proposed to be involved in the formation of trust in the areas of interpersonal trust and human-technology interaction. The framework consolidates the main assumptions of the two models by Lee and See (2004) and Hoff and Bashir (2015) and further specifies the psychological processes involved in trust formation along ten general propositions integrating assumptions of the two models with theorizing from the domains of interpersonal trust, personality psychology, and social cognition. The ten propositions of the framework entail general theoretical assumptions about the interrelations of variable groups acting together in forming and calibrating trust on the basis of available information. Altogether, the framework proposes three sequential stages of trust formation as depicted in Figure 3. The Three Stages of Trust provide a framework for integrating existing research and orienting hypothesizing and design of future research endeavors for investigating trust in automation (see chapters 1.5.1, 4.2, and 4.4.1 for a more detailed discussion of the advances by the framework). In the following, ten propositions, theoretically underpinning the framework, are derived and specified.

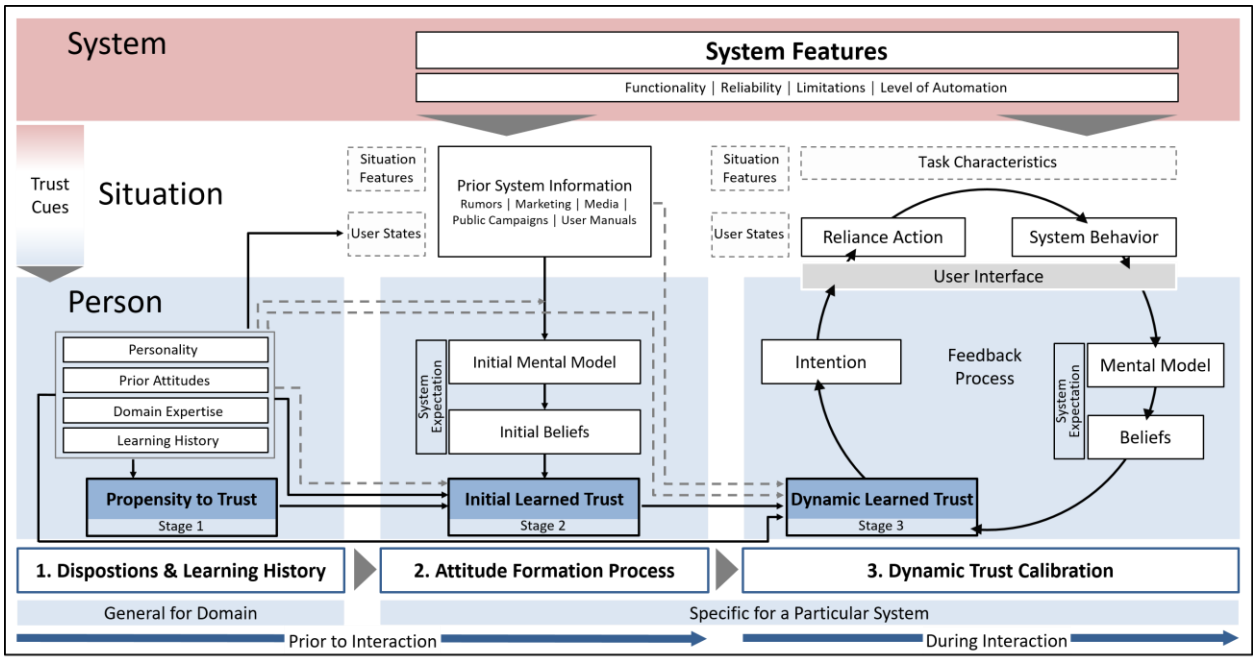
2.1 Propositions of the Three Stages of Trust Framework

An essential defining psychological momentum of trust is an inherent uncertainty about the trustworthiness of the trustee (see chapter 1.1). Thus, trust formation can be described as a psychological process in which sources of available information are used to reduce uncertainty about the trustee and to build up an expectation about his actual trustworthiness over time. In a similar manner as interpersonal trust can be understood as a result of a learning process of other humans' trustworthiness (e.g., Rempel et al., 1985), also trust in automation is hypothesized to be built up in a learning process in which the basis of trust changes over the course of the relationship with an automated system (e.g., Lee & Moray, 1992; Lee & See, 2004). As relevant information for the interaction with a system is already available before the interaction begins, the trust processes are assumed to start before the interaction with an automated system. Accordingly, and in line with Hoff and Bashir (2015), the proposed framework distinguishes and relates trust processes prior to and during the interaction with an automated system. Based on the work by Rempel et al. (1985), Lee and See (2004), and Hoff and Bashir (2015), it is proposed that trust in an automated system is dynamically built up over the course of the individual learning history with technology and the usage history with a particular system (as depicted in the x-axis of Figure 3). Thereby, trust in a specific system at a given time is built on the basis of the individual learning history and the aggregate of received information and experiences over all instances of familiarization with a system and its subsequent use up to this point. From this, the first proposition of the framework is derived:

Proposition 1: Trust in automation is formed in a dynamic learning process over time, which begins before the actual interaction with a specific automated system.

Figure 3

Illustration of the Three Stages of Trust Framework for the investigation of the dynamic formation and calibration of trust in automation. Trust in automation is assumed to be built up through three sequential trust layers at three trust stages: 1. the propensity to trust in automation, 2. initial learned trust, and 3. dynamic learned trust. The framework provides an integration and specification of psychological processes and relationships of variable groups that affect trust formation. Solid black arrows illustrate hypothesized direct effects and (partial or full) mediations. The dashed grey boxes and arrows represent hypothesized moderations in the processing of available trust-relevant information. Note: not every theoretically meaningful relationship is specifically illustrated to increase clarity of the illustration.



While this temporal sequence of trust formation is depicted in the x-axis of Figure 3, on the y-axis of Figure 3, the framework differentiates three categories of variables – person-, system-, and situation-related – which are assumed to affect trust development over the course of the familiarization with an automated system (see chapter 1.4). In Figure 3, the components and properties of the technical *system* are located in the upper area and dispositional *person*-related influencing variables are depicted in the lower area. The relationship between a person and a system is proposed to be built up in a series of *situations* in which information about the system is available or experiences with the system are made. On the basis of the collection of such information and experiences, users dynamically build their trust in the system. These situations are characterized by environmental conditions (e.g., the task or the constellation of present stimuli) and situational variations on the side of the user (e.g., states, motivation) that define the situational circumstances in which users receive information about the system and experience its behavior. As for the differentiation of trust processes general for automated systems and those specific for a particular system, as reflected by the arrows in the lower area of Figure 3, the properties of a specific system and the situation only affect trust formation at the later, system-specific stages. Accordingly, it is proposed:

> *Proposition 2: Trust in automation is essentially influenced by three variable groups: characteristics of the person, the system, and the situation.*

To summarize the first two propositions of the framework, trust in automation is assumed to be built up in a learning process, in which an expectation about the actual trustworthiness of an automated system is formed from the available information before and during the interaction with a

system. This process is proposed to be influenced by factors on the side of the system, the person and the situation, in which the trust-relevant information is presented or respective experiences are made. Thus, the process of the formation and calibration of trust in an automated system can be characterized as a learning process over time, in building up an expectation about a system on the basis of prior information and learning about it during the interaction from a sequence of observations of its behavior and the output of its user interface. This history-dependent process of trust formation is accounted for in the depicted conceptions of different levels of attributional abstraction by Rempel et al. (1985) in the interpersonal domain and by Lee and Moray (1992) in the domain of interaction with automated technology (chapter 1.3.1).

In line with these conceptions, it is proposed that in the process of the formation of trust in automation three trust layers are built up at three stages of trust formation, which are run through sequentially prior to and during the process of getting to know and using a system (in Figure 3 in the lower part). Thereby, similarly to the models of trust levels, at these three stages, the psychological processes of trust formation build on available information with different character. The emerging trust layers at each stage are conceptualized on the basis of the trust layers discussed in chapter 1.4 (Marsh & Dibben, 2003; Hoff & Bashir, 2015): the propensity to trust (stage 1), initial learned trust (stage 2), and dynamic learned trust (stage 3).

As for the underlying definition of trust as an attitude (chapter 1.1), before the explication of the trust processes at each of the proposed stages, a conceptual differentiation between the propensity to trust and learned trust in automation seems indicated. Attitudes as dispositions to more or less favorably respond to an object (Ajzen, 2005) are hypothesized to vary in their generality vs. specificity according to the object to which they refer (e.g., Sherman & Fazio, 1983; Albaraccín & Chan, 2018). Accordingly, the two forms of trust can be differentiated on the basis of this dimension of specificity. While the propensity to trust is more globally related to all kinds of automated systems, learned trust is specific for the system under consideration. Besides this, attitudes share many resemblances with personality traits in their explanation of the relation between individual differences and variations in behavior (e.g., Ajzen, 2005; Sherman & Fazio, 1983; Albaraccín & Chan, 2018). In this respect, personality was found to essential influence the formation of attitudes. This relation was nicely summarized by Albaraccín and Chan (2018): “Attitudes are seen as a product of experience leading to individual differences and stemming from general, stable personal dispositions. […] [P]ersonal factors, including personality traits, are likely to pose a ceiling to the degree of situational social influence” (p. 4). Based on this, the propensity to trust in automation is conceptualized as a technology-specific personality trait in line with theoretical conceptions of the propensity to trust in other people (e.g., Frazier et al., 2013; Mayer et al., 1995). Personality traits are “relatively stable interindividual differences in the proneness […], tendency […], style […], or disposition […] to behave […], feel […], or think in certain ways […]” (Hamaker, 2007, p. 297) associated to tendencies of a person to interact with and adapt to social and physical environments (e.g., Larsen & Buss, 2005). Based on the notion that personality traits at different levels of specificity exist, the propensity to trust in automation can be understood as a surface trait resulting from the interplay of more general traits (e.g., dispositional interpersonal trust), attitudes, the individual learning history, and the domain involvement (e.g., hierarchical personality models; Mowen, 2000). Learned trust in automation, on the other hand, reflects a specific momentary attitude towards a certain system in line with the provided definition (Lee & See 2004, chapter 1.1).

In the following, a short wrap-up of the proposed processes at each stage is provided. At the *Trust Stage 1*, before any specific system is in the focus of the user*, the propensity to trust in automation,* a specific personality trait regarding the tendency to perceive automated systems as trustworthy, is formed on the basis of broader personality variables, the individual learning history, and emerging dispositional technology-related attitudes. At the *Trust Stage 2*, *initial learned trust* is formed on the basis of the propensity to trust in automation and available information about a specific system before any interaction (e.g., marketing, word-of-mouth). As for the underlying definition of trust as an attitude, it is proposed that the available information is used, at least in part, analogous to information processing in attitude formation. At the *Trust Stage 3*, on the basis of the former trust stages, *dynamic learned trust* is calibrated along the experienced system behavior during the interaction with the system. This trust layer is proposed to be calibrated in a dynamic feedback loop (e.g., Lee & See, 2004), in which the current system expectation is compared against the observed system behavior and user interface output. Taken together, the framework proposes:

Proposition 3: Trust in automation is formed in three consecutive trust stages, at which three trust layers build on each other:

1) Stage 1: Dispositions and learning history - the propensity to trust in automation,
2) Stage 2: Attitude formation process - initial learned trust in automation,
3) Stage 3: Dynamic trust calibration - dynamic learned trust in automation.

At the first stage of the framework, the propensity to trust in automation – a general individual predisposition to trust in an automated system irrespective of a specific context, system, or task – is established (Hoff & Bashir, 2015; see chapter 1.3.3). In the same manner, as prior trust experiences are assumed to direct expectations of trustworthiness in specific other people in the interpersonal domain (Thielmann & Hilbig, 2015), the propensity to trust can be described as a generalized form of trustworthiness expectation towards all kinds of automated systems reflecting an individual's learning history with automated systems.

Based on this conceptualization, it is proposed that the propensity to trust in automation results from an interplay of broader personality variables, the sum of learning experiences with automated systems, and general attitudes towards technology and automation. Thereby, the formation of the propensity to trust in automation is grounded in the more general personality of a user, which accentuates the experiences with technology and automated systems throughout the life span. In the course of the individual learning history, this accentuation and the sum of experiences lead to individual differences in the propensity to trust in automation that, in turn, is proposed to influence the expectations and trust processes at the more specific stages of the framework. Accordingly, people with a higher propensity to trust in automation will likely show higher initial trust levels and tend to engage in more trusting behavior in the first encounter with a new automated system. In the same personality-accentuated learning process, besides the propensity to trust also other specific technology-related personality traits and general technology-related attitudes are established that are proposed to influence the propensity to trust and the trust processes at the later stages. An exemplary technology-specific personality trait is the affinity for technology (Karrer et al., 2009). General technology attitudes reflect, for example, the general evaluation of automated technology or of a more specific domain like automated driving. Taken together, for the trust processes at the first stage of the proposed framework, it is assumed:

Proposition 4: The propensity to trust in automation is formed on the basis of user personality, the individual learning history with technology and automated systems (e.g., experience and domain expertise), as well as general attitudes towards technology and automation.

Based on the general propensity to trust, at stages 2 and 3, during system familiarization, learned trust in a specific system is proposed to be established in a process that forms trust from an interpretation of available information allowing for conclusions about the trustworthiness of the automated system. For the inherent uncertainty in trust situations, in line with Thielmann and Hilbig (2015) "trustors can never be completely sure about the trustee's trustworthiness but have to estimate the probability of trust appreciation" (p. 12). To account for this, an essential element of most trust definitions is the trustor's expectation of the trustee's trustworthiness and the perception of risk and vulnerability associated with reliance decisions (e.g., Mayer et al., 1995; Lee & See, 2004; Thielmann & Hilbig, 2015; see chapter 1.1). In the case of trust in automation, the *system expectation* entails the assessment of the trustworthiness of the automated system under consideration. This system expectation, in turn, is assumed to build an essential basis for trust and the intention to rely on a system. It is further assumed that in the process of establishing both layers of learned trust, this expectation of the system's trustworthiness is informed along with available diagnostic information about the system, both prior and during system interaction. Thereby, from the available information *trust cues* are extracted, which can be characterized as "in some way observable or given pieces of evidence a trustor might use to draw inferences about a trustee's trustworthiness in a specific situation" (Thielmann & Hilbig, 2015, p. 21). These trust cues are evaluated and interpreted by the user on the basis of individual dispositions and learning experiences (e.g., propensity to trust, stage 1). It can further be hypothesized that the character of trust cues changes over the course of the two stages of learned trust, e.g., in terms of their attributional abstraction (Lee and Moray, 1992). While before the interaction, more less specific purpose information, e.g., from marketing or information campaigns, might be available, during the interaction, users can directly assess the system performance.

As a cognitive basis of system expectations and antecedents of learned trust, the framework further integrates mental models at stages 2 and 3 (see Figure 3). Mental models are defined as "a rich and elaborate structure, reflecting the user's understanding of what the system contains, how it works, and why it works that way" (Carroll & Olson, 1987, p. 12). Following Durso and Gronlund (1999), it is assumed that this knowledge structure consists of representations of if-then relations about system functionality under certain conditions. Mental models were found to affect the characteristics that people attributed to automated systems (Beggiato & Krems, 2013) and accordingly have been discussed as an essential basis for trust calibration (Beggiato et al., 2015). It is proposed that the mental model of a system is established from the sum of processed information in an individual learning history with a system. As a second, more evaluative basis for the establishment of system expectations (in line with Lee & See, 2004), additionally, beliefs about the system are integrated into the framework. These are proposed to be established from available information, the content of the mental model, and other variables (e.g., affective processes). Based on the provided definition (chapter 1.1), beliefs reflect specific attributions of characteristics to a given system and include, for example, the subjective, expected predictability, dependability, competence, and reliability of a system (Muir, 1994; Muir & Moray, 1996; Chancey et al., 2017). In combination with the available information and entailed trust cues, the mental model and the beliefs are proposed to form the system expectation, which in turn informs the formation and calibration of

learned trust. This relationship of mental models, beliefs, system expectation, and learned trust is represented in Figure 3 at the stages 2 and 3. Accordingly, it is assumed:

Proposition 5: Learned trust in automation is built up and calibrated in two phases - initial learned trust and dynamic learned trust - on the basis of available information (trust cues) and the system expectation (informed by the mental model and beliefs).

At the two stages of learned trust the framework proposes two processes in which information is used to build up and calibrate trust. Thereby, it has to be clearly stated that these processes are not separated and are likely combined at both stages. As for the different character of available trust cues and the different phases in relationship development, the relative importance of these processes is proposed to change between the two stages.

Initial learned trust (at stage 2) is proposed to be formed on the basis of trust cues provided with *prior system information*, which encompasses any kind of communication about a system before a user actually interacts with it (similar to the diagnostic information Madhavan & Wiegmann, 2007 discuss). This includes, e.g., marketing instruments like advertisement, public information campaigns, word-of-mouth, technical documentation, and training for the automated system under consideration (see Figure 3). In regard to the content of the provided information that is relevant for the formation of trust in automation, again, an orientation on similar processes in the interpersonal domain provides starting points. In the interpersonal setting trust before an actual interaction with another human is being inferred, exemplarily, from trust cues related to the appearance, reputation, and social category of the other person (e.g., Thielmann & Hilbig, 2015). In the case of automated technological systems, appearance and reputational information can also be used as trust cues. In terms of appearance, future users might, for example, be addressed with an advertisement previewing the design and the associated user interface of a technological system, from which an initial trustworthiness expectation in regard to the system might be derived. In the case of reputational information, for example, the brand of a system might carry information about the quality and image of the system relevant to the expected trustworthiness of the system. Additionally, as mentioned before, also a classification of trust cues based on the levels of Lee and Moray (1992; see chapter 1.3.1) seems useful. From this, an integrated, preliminary classification of trust cues for the context of human-automation trust into five categories can be derived: *reputation*-, *purpose*-, *process*-, *performance*-, and *appearance*-related trust cues.

In line with the definition of trust as an attitude (see chapter 1.1), the psychological process at stage 2, in which available pieces of information (trust cues) are selected, processed, and weighted is proposed to be characterized, at least to a substantial extent, as an attitude formation process. A transfer of theories and models of attitude formation, attitude change, persuasion, and on the relationship between attitudes and behavior can be helpful in understanding the interplay of different kinds of information, processes, and individual characteristics of the trustor in the formation and calibration of initial learned trust. In this regard, basic theories and models from social psychology and social cognition (see, e.g., Maio et al., 2018 for an overview) provide promising additions to the theoretical understanding of how trust in automation is formed and changed. This view is compatible with the three proposed routes of information processing through which information influences trust in automation by Lee and See (2004): *analytic*, *analogical*, and *affective*. Information about the automated system that is processed by the user before and during system use can be

transformed along these three routes. These routes are not proposed to be separated, but information can be processed in varying relative proportions by several routes and the routes are assumed to influence each other (Lee & See, 2004). First, through the analytical route, trust is formed comparable to rational decision-making under uncertainty in which the ratio of expected gains and losses are evaluated. In the second route, analogical processes derive trust from characteristics of the interface or the category membership and, third, through the affective route, emotions towards the trustee or the situation inform trust formation. This view of different routes of information processing in attitude formation and change is compatible with *dual-process theories* from social psychology (Chaiken & Trope, 1999; Chen et al., 1996), which also correspond well with the attributional levels proposed by Lee and Moray (1992). In this regard, exemplarily, the Elaboration Likelihood Model (ELM, Petty & Cacioppo, 1986), a well-established dual-process theory, is discussed here in more detail. The ELM proposes two routes through which messages can influence attitude change. While the central route is characterized by an in-depth elaboration of the meaning and the arguments of the message resulting in stable attitude changes, attitude change through the peripheral route is a result of attributional processes in relation to surface characteristics of the message or the sender of the message. The attitude changes affected by the latter are superficial, unstable, and prone to change. The likelihood of elaboration, thus, the likelihood that a message is processed with cognitive effort through the central route, depends on the motivation and the ability of the message receiver to engage in effortful elaboration. The ELM allows for the derivation of a number of study hypotheses to investigate the formation of initial learned trust and builds the theoretical foundation for study 4 of this dissertation (see chapter 3.3). Taken together, initial learned trust is proposed to be formed to a substantial extent in a process that corresponds to the social cognitive perspective of attitude formation and change on the basis of available prior system information and a system expectation:

Proposition 6: Initial learned trust in automation is formed and calibrated on the basis of prior system information in a process that follows to a substantial extent the mechanisms of attitude formation and change.

Regarding the trust processes, in which dynamic learned trust is calibrated and related to reliance behavior during the interaction with an automated system, in orientation on the model of Lee and See (2004), the proposed framework assumes a psychological mediation process in which dynamic learned trust plays a crucial role in translating the interpretation of available trust cues to actual reliance decisions. Such trust cues during the system interaction are, for example, entailed in observations of system behavior and user interface content. In line with the underlying definition of trust as an attitude by Lee and See (2004) and in orientation on their model, the framework entails the core elements of the PLABE theory (Ajzen, 1991, chapter 1.3.2), namely, beliefs, attitudes, and intentions, as a basis to inform the calibration of dynamic learned trust at stage 3. Additionally, as detailed in proposition 5, the mental model is incorporated in this cascade as a memory-related knowledge basis for beliefs and system expectations. In sum, this mediation cascade is proposed to constitute to establish a psychological process in which during the interaction trust is dynamically calibrated and translated into behavioral intentions and reliance decisions:

Proposition 7.1: Dynamic learned trust in automation influences reliance decisions as part of a mediation process (mental models, beliefs, attitudes, intention).

Additionally, at stage 3, dynamic learned trust is proposed to be initially informed by the system expectation built up prior to system use. On this basis, during the interaction with the system,

dynamic learned trust is proposed to be continuously updated and adapted in a dynamic learning process, in which the expectation of system behavior is compared to its actual behavior and the received information about system functioning in the user interface. To account for this continuous update process of dynamic learned trust, the framework incorporates the general assumption of a feedback loop by Lee and See (2004), in which dynamic learned trust is dynamically calibrated in the proposed mediation cascade (proposition 7.1) along the observed system behavior. This dynamic feedback loop is illustrated in Figure 3 at stage 3. As an addition to Lee and See's (2004) model, the feedback process in which trust is calibrated is proposed to function comparable to the feedback mechanisms entailed in established theories of goal-directed behavior and self-regulation (e.g., Carver, 2004; Miller et al., 1960; Hacker, 2006). In these feedback processes, a stimulus is repeatedly tested for its congruence with a desired state until a specific goal is satisfied (e.g., test-operate-test-exit in Miller et al., 1960). For clarification the feedback processes in which dynamic learned trust is proposed to be calibrated can be illustrated, for example, along the four elements of the behavioral feedback loop from the self-regulation theory (Carver, 2004): 1) input function (perception of the status in the environment), 2) reference value (internal goal of the person), 3) comparator (input vs. reference), and 4) output (behavior or non-behavior), from which the effect in the environment is fed back to the input function. The initial system expectation (initial mental model and beliefs) and the associated initial learned trust are proposed to build the initial reference value in this feedback loop. Based on this reference, as an output, the automation is relied on or not in different subtasks. The resulting system behavior is observed as input, compared to the expectation, and on this basis, the system expectation, and dynamic learned trust are adapted, updated, and used to inform a subsequent behavioral intention and reliance decision. The adoption of the mechanism from feedback loops of other basic psychological phenomenon helps to derive more specific predictions of how trust is calibrated at stage 3. In line with this reasoning, it is further proposed for the calibration of dynamic learned trust in automation:

Proposition 7.2: Dynamic learned trust in automation is calibrated in a feedback process during the interaction with an automated system along a comparison of the current system expectation and the observed system behavior and the associated output of the user interface.

This dynamic feedback process in which trust is calibrated is further assumed to be directed at an equilibrium between the expected trustworthiness of the system and the actual reliability of the system. Accordingly, and following the assumptions on the character of trust formation in the interpersonal domain by Rempel et al. (1985), it can be assumed that dynamic learned trust stabilizes in the course of the interaction relationship with an automation. At the beginning of system use, the difference between the expected and the actual system behavior should, as for the incomplete picture of the user about the system, tend to be relatively large. In the course of interaction with the system, if the system behavior corresponds with the user's established system expectation, this difference should be iteratively reduced with repeated calibration of trust and system reliance. The findings of Beggiato et al. (2015) support such a stabilization of trust in the course of growing experience with an automated system. In this sense, especially, the first encounters with a system are crucial for trust calibration. In line with this reasoning, it is assumed:

Proposition 7.3: Dynamic learned trust in automation is stabilized in the course of the interaction relationship with an automated system if the system continuously functions in accordance with the user's system expectations.

Based on the proposition that the preceding trust stages influence the subsequent ones (proposition 3), initial learned trust is assumed to be essentially influenced by the propensity to trust in automation (arrow between stages 1 and 2 in Figure 3). In line with this, the propensity to trust in automation has been considered as a variable affecting learned trust (e.g., Lee & See, 2004; Hoff & Bashir, 2015). As empirical evidence for this relationship is scarce at this time (Merritt & Ilgen, 2008; Merritt et al., 2013), in studies 1 and 2, it was investigated in more detail (chapter 3.1). Furthermore, it is assumed that besides the propensity to trust in automation also additional user characteristics (e.g., user personality or expertise) influence initial learned trust. Thereby, these influences might either be direct (e.g., people who are low in trait A trust automated systems less in general) or indirectly mediated by the propensity to trust in automation. Also, a combination of these situations is possible (a part of the influence is direct, a part of the influence is mediated by the propensity to trust). In a statistical interpretation, these assumed possible effects from dispositions to initial learned trust can be framed as direct effects, partial, and full mediations. The direct path from user dispositions to initial learned trust is illustrated in Figure 3. Similarly, user states are proposed to affect initial learned trust directly (e.g., a sad person might trust the system less irrespective of provided information). Direct effects of user states are not illustrated in Figure 3.

In a similar manner, the trust processes at stage 3 are proposed to be substantially influenced by user characteristics, user states, and the trust processes at the two preceding stages (as illustrated in Figure 3). In this sense, initial learned trust builds the starting point for the calibration of dynamic learned trust and the initial decision to use or not use a system. In the same manner, direct effects, as well as full and partial mediations of the effects of user dispositions through the propensity to trust and initial learned trust, are proposed (as illustrated in Figure 3). On this basis, it is further assumed:

Proposition 8.1: User dispositions, user states, and the preceding trust stages directly influence the formation of both layers of learned trust. Thereby the effects of the more general variables can be fully and partially mediated by the more specific variables.

Additionally, the assumption that trust processes at stage 3 are based on those of the preceding stages, includes, that prior system information guides the interpretation of system behavior and output in the system's user interface. For example, when a system has to execute a calibration of its sensors, such a behavior might be very differently interpreted if this calibration was explained to a user before as compared to a situation in which the user does not expect such a system behavior. The same effects are proposed to play a role in the interpretation of user interface output. The moderation of the interpretation of observed system behavior and user interface output is illustrated in Figure 3 by a dashed grey arrow from prior system information to dynamic learned trust.

Proposition 8.2: Prior system information influences the interpretation of observed system behavior and user interface output at stage 3.

Furthermore, the perception, use, processing, and interpretation of information at stages 2 and 3 are proposed to be moderated by user characteristics and situational variances at the time of their presentation (illustrated by grey dashed arrows and boxes in Figure 3).

Regarding initial learned trust at stage 2, individual differences were found to essentially guide the search, selection, and interpretation of available information in the process of attitude formation (e.g., Cacioppo et al., 1983). In the same manner, in building up initial learned trust, prior system

information is proposed to be processed differently by different persons, besides others, based on differences in personality, technology attitudes, and the propensity to trust in automation. In this regard, the formation of initial learned trust is proposed to be not only directly influenced by user characteristics and the propensity to trust in automation, but user dispositions are proposed to moderate the use and interpretation of available information in the formation of initial learned trust (dashed grey arrow in Figure 3). In the same manner, the situational conditions, in which trust cues are extracted from prior system information, are proposed to influence the attitude formation processes at this stage, which is, for example, in line with propositions of dual-process theories (e.g., Chaiken & Trope, 1999). For example, differences in terms of distraction by other presented information or time constraint should influence the visibility and perceptibility of information and, thus, influence the inclusion of this information in the trust formation process. Similarly, fluctuating user states are proposed to influence how provided information is used and interpreted. In this regard, for example, users' momentary motivation, emotional states, and cognitive resources should guide information search and interpretation (dashed grey boxes at stage 2 in Figure 3).

In a similar process, user characteristics and situational factors are assumed to affect information processing in trust calibration during the interaction with an automated system at stage 3. In this regard, the perception and interpretation of the system behavior and associated information in the user interface of the system are proposed to be affected by fluctuating individual and situational factors (illustrated by dashed grey arrows and boxes in Figure 3). Exemplarily, the interpretation of an output of a graphical or auditory interface might vary over users based on differences in their domain expertise. Similarly, the same ambiguous user interface output might be interpreted differently in two situations based on the constellation of stimuli in the environment. Also, if a user is distracted at the moment when the system performs a task or when the user interface presents a specific output, the user cannot adequately use this information to calibrate his or her trust level. Taken together, it is assumed:

Proposition 9: Information perception, use, evaluation, and processing in the formation of both layers of learned trust is moderated by user characteristics (dispositions and traits), user states (e.g., emotional states, motivation) and properties of the environment in which the information is presented (situation features and task characteristics).

In line with the common finding that trust is not always predictive for system use and reliance decisions (e.g., Wiegmann et al., 2001) it is proposed that the link between trust in automation and reliance is moderated by user states (e.g., emotional states, motivation), user characteristics (e.g., risk awareness), situational variables (e.g., environmental factors like lighting conditions), and task characteristics (e.g., risk-benefit relation), at the time of the single reliance decisions. As trust is essential in situations of risk and uncertainty (see chapter 1.1), its relevance for reliance decisions should be especially high in the early phase of system interaction. Thus in this phase, the relationship between the level of trust and the behavioral decision should be relatively strong. In the course of interaction with a system, the level of uncertainty is likely to be reduced and trust-directed reliance decisions might be replaced more and more by knowledge-based decisions. At the same time, for the associated uncertainty in complex and sophisticated automation, trust in automation presumably predicts reliance decisions over all phases of system interaction to some extent. Regarding individual differences that moderate the relationship between trust in automation and behavioral decisions in the interaction with automation, it was shown that self-confidence influences the relationship between trust and reliance behavior (e.g., Lee & Moray, 1994). In the

same manner, other dispositional or situational differences in the user's self-assessment might moderate the link between trust and reliance. In the same manner, characteristics of the task are proposed to moderate the association between trust and reliance decisions (e.g., the involved degree of risk). Taken together, it is assumed:

Proposition 10: The strength of the association between dynamic learned trust and reliance behavior is moderated by user characteristics (e.g., traits and states), situational variables, and task characteristics.

Taken together, the Three Stages of Trust integrates and extends the theorizing on the formation and calibration of trust in automation by including assumptions from other psychological areas on the nature of information processing, learning, and attitude formation. In specific, the framework emphasizes the interrelation of three subsequent trust stages in trust formation. Furthermore, essentially, the influence of individual differences and situational characteristics at the two stages of learned trust are emphasized. The Three Stages of Trust serve as a framework for the investigation of trust in automation and supports the derivation of research questions and hypotheses for future research to empirically investigate, extend and refine the framework propositions. Thereby, the framework cannot be supported or falsified in full by single studies. Rather like any theory, it might be falsified and improved in a cumulative research process and thereby fostering discussions and research ideas. From the framework propositions, also, a number of design requirements for studies on trust in automation can be derived. These are discussed in detail in chapter 4.2.6 and include, besides others, implications for the measurement of trust in automation. In the appendix (Appendix A), on this basis, the German Version of the Short Learned Trust in Automation Scale (LETRAS-G) is presented. This scale was used in five of the six studies, that will be presented in the following. The investigated study hypotheses aimed at an empirical investigation of some of the proposed relationships of the presented framework.

3 Summary of Studies

On the basis of the propositions of the Three Stages of Trust framework, a series of six studies was conducted investigating the formation and calibration of trust in automation at the different trust stages in the domain of highly automated driving. The studies take a history-based approach to trust formation and either measure one of the proposed trust layers repeatedly over time or several of the proposed trust layers. The presented studies were conducted in the domain of highly automated driving (SAE level 3) and were published in four peer-reviewed articles.

Studies 1 and 2 focused on the role of user dispositions in trust formation at stages 1 to 3. In these studies, a hierarchical personality model for the prediction of dynamic learned trust in an automated driving system was explored in study 1 and validated in study 2 in an independent sample. The investigated model included hypotheses on the relation of personality traits to the propensity to trust in automation (stage 1) and dynamic learned trust (stage 3). Also, in these studies the relation of dispositional interpersonal trust, the propensity to trust in automation and dynamic learned trust was investigated. To gain a better understanding of the processes, in which personality affects trust formation, **study 3** investigated how the level of state anxiety at the time of introduction of an automated driving system mediates the effects of personality (depressiveness and negative self-evaluations) on dynamic learned trust. **Study 4** investigated the proposition that ini-

tial learned trust is formed analog to attitude formation. On the basis of the ELM, it was investigated how different types of prior system information about an automated driving system are used to build up initial learned trust at stage 2 and how individual differences in need for cognition moderate this process. Additionally, a link from personality to the emergence of initial learned trust was investigated (stage 1 to stage 2). In **studies 5 and 6**, the calibration processes of dynamic learned trust (stage 3) and the impact of both prior system information (stage 2) and observed system behavior (stage 3) was investigated in a driving simulator study with a highly automated driving system. More specifically, in study 5, the development of dynamic learned trust over time was compared between a situation of flawless system functioning and a situation in which a malfunction occurred. In study 6, the effect of prior system information about system malfunctions on the calibration of dynamic learned trust in the same situations as in study 5 was investigated. In this regard, study 6 provided an investigation of the use of prior information for the interpretation of observed system behavior at stage 3.

In the following, for each article, the research questions, hypotheses, study designs, and results of the statistical analyses are summarized. Additionally, for each study the findings are discussed in relation to the propositions of the presented framework. The articles are included in Appendix B.

3.1 Studies 1 and 2: What's Driving Me? – Exploration and Validation of a Hierarchical Personality Model for Trust in Automated Driving

Kraus, J., Scholz, D., & Baumann, M. (2020). What's driving me? – Exploration and validation of a hierarchical personality model for trust in automated driving. *Human Factors*. Advance online publication. https://doi.org/10.1177/0018720820922653

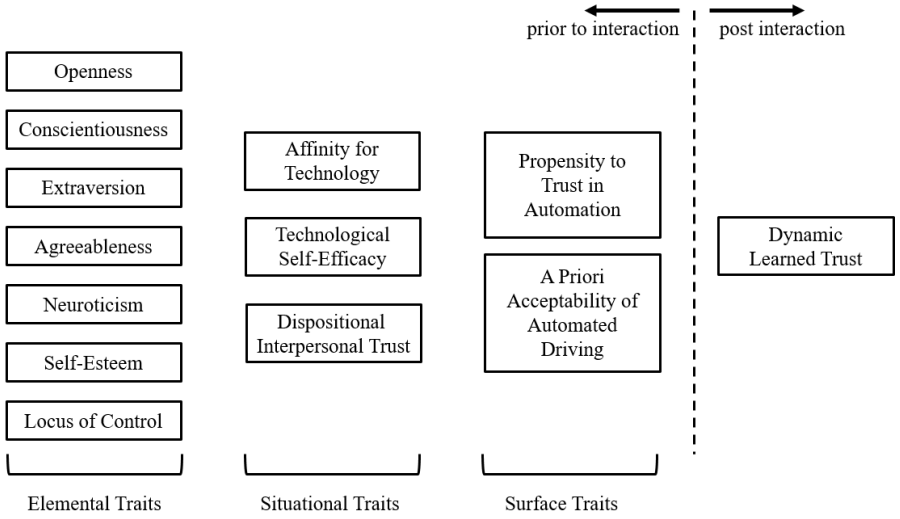
3.1.1 Introduction and Study Hypotheses

On the basis of prior research, the proposed framework identifies user personality and general technology-related attitudes as an essential source for individual differences in trust in automation (e.g., Schaefer et al., 2016; Hoff & Bashir, 2015; Lee & See, 2004). Essentially user dispositions are proposed to influence the formation of the propensity to trust in automation (stage 1, proposition 4). Furthermore, personality is proposed to influence learned trust at both stages, either directly, through the propensity to trust (proposition 8.1), or by affecting the use of available information at these stages (proposition 9). Against this background and on the basis of the inconclusive state of knowledge on which personality differences affect trust formation (subsumed in chapter 1.4.1), studies 1 and 2 aimed at an integrative investigation of personality variables theoretically associated with differences in dynamic learned trust in an automated driving system. Thereby, the studies followed a hierarchical approach to concurrently investigate the personality basis of dispositional interpersonal trust, the propensity to trust in automation, and dynamic learned trust.

On the theoretical foundation of Mowen's 3M model (Mowen, 2000), a hierarchical personality model that orders different personality and attitude constructs according to their specificity for dynamic learned trust in automation, was explored (study 1) and validated (study 2). Hierarchical personality models are based on the assumption of a hierarchical organization of personality in which cross-situational, general traits influence the emergence of more specific, narrower traits, which in turn affect behavior (Mowen et al., 2007) and thereby provide a solution for the bandwidth-fidelity dilemma in personality research (e.g., Cronbach & Gleser, 1965). This dilemma

Figure 4

Hypothesized hierarchical structure of the investigated user dispositions to predict dynamic learned trust in automation. (Figure from Kraus, Scholz, & Baumann, 2020, ©2020 Human Factors and Ergonomics Society).



describes the conflicting advantages and disadvantages of general and specific traits: while specific traits provide a good explanation of variance in specific behavior, their theoretical foundation is superficial. On the other hand, broader and theoretically well-established personality constructs like the Big Five (e.g., McCrae & John, 1992; Costa & McCrae, 1992) often provide only a limited variance explanation. Hierarchical personality models resolve this dilemma by integrating traits at different levels of abstraction for the explanation of specific phenomena (e.g., Allport, 1961; Mowen, 2000). At the same time, hierarchical modeling of personality traits allows for an integrative analysis of traits within the nomological network of a particular outcome. For the hierarchical personality model of dynamic learned trust in automated driving (see Figure 4), the included traits and attitudes were selected on the basis of a thorough review of available studies investigating the role of one or more dispositional variables for trust in automation and related areas. The proposed model orders user dispositions at three levels. The decision where to order the included variables was informed by the definitions of three trait levels by Mowen (2000): elemental, situational, and surface traits, the labels of which were kept although a general technology-related attitude was included in the model (the acceptability of automated driving). The model, as depicted in Figure 4, includes ten general hypotheses on the role of the included dispositional variables for the prediction of dynamic learned trust in automation (except for openness and consciousness). The hypotheses were carefully underpinned on the basis of psychological theory and empirical findings on trust in automation and related phenomena (see Appendix B for more details). It is further hypothesized that the variables at preordered model levels directly affect the variables ordered at the next succeeding model levels and that the effects of the elemental and situational traits on dynamic learned trust are mediated by the variables at the intermediate model levels. As for the explorative approach of this research, hypotheses were not further specified in regard to the single mediation paths. Rather more, the exploration of the mediation paths within the hypothesized hierarchical structure and the relative predictive power of the included traits for differences trust in automation were essential research questions of these studies.

To answer these research questions and investigate the hypothesized hierarchical personality model, two consecutive online studies were conducted in which participants were presented with

a series of videos of an automated driving system. In study 1, the hypothesized model was explored and reduced to allow an economic prediction of dynamic learned trust in automation. In study 2, the reduced final model was validated in an independent study sample.

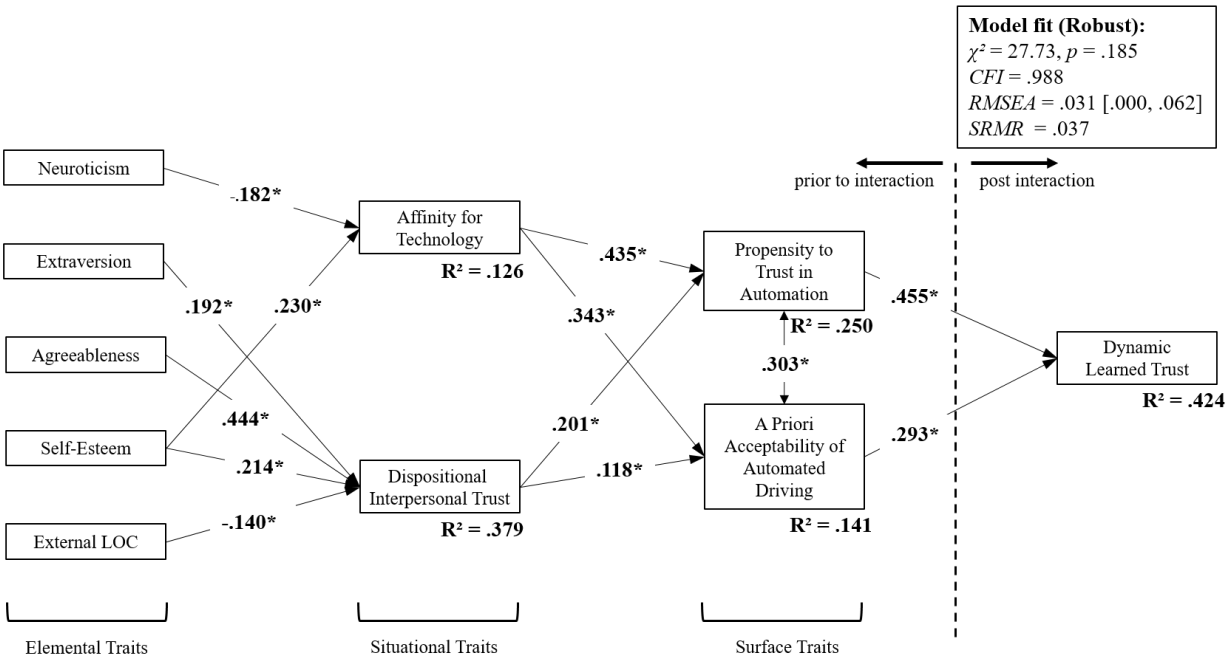
3.1.2 Study 1 - Method

The final sample of the study consisted of $N = 274$ participants (159 women, 58%, 113 men, one other and one "not specified") with an average age of $M_{age} = 33.88$ years ($SD = 14.75$). The study took about 47 minutes on average to be completed. First, after being welcomed, participants answered a series of personality and attitude scales. If available, validated short scales were used. For the remaining constructs, existing scales were shortened on the basis of item characteristics, factor loadings, and representativeness of content. In some cases, scales were translated from English to German and rephrased to be suitable for the domain of automated driving. After the personality questionnaires, participants were provided with general information about automated driving and were introduced to the "motorway pilot", the automated driving system (SAE level 3) under consideration in these two studies. The motorway pilot was described as a prototype of an automated driving system that is currently undergoing road testing and that allows for partially automated driving by taking over both longitudinal and lateral control for a certain interval of the drive. The system was further explained to be able to automatically conduct driving maneuvers in compliance with traffic rules. Additionally, participants were informed that system malfunctions lead to the activation of a safe mode in which the system drives to the service lane, completely stops, and activates the warning lights. In this case, the system could only be re-activated after the driver manually drives back on the motorway. After this, participants were presented with seven animated videos showing different scenes of driving with the system on the motorway from the driver's ego perspective. Video stimuli were created with the VICOM editor (TÜV | DEKRA arge tp 21, 2011) and selected on the basis of a pilot study ($N = 36$) to enhance the comparability of the videos. Each video showed a 30s drive on the motorway with a minimum of two lane changes and one overtaking maneuver. After each video, besides others, the LETRAS-G (see Appendix A) was presented. After the experimental set-up, a socio-demographic questionnaire was filled out. In the study, two variables not in the scope of this research question were manipulated. First, as a manipulation for system transparency, information about system malfunctions and its causes were provided in the initial system information in three levels of increasing transparency. Second, as a manipulation of system reliability, the fourth video either showed a non-critical situation like in the remaining videos or a system malfunction with subsequent activation of the safe mode.

To test the hierarchical personality model, path models with scale means were calculated with the R package lavaan (Rosseel, 2012; version 0.6-2; R version 3.4.3). At each mediation level, covariations between the residuals of the modeled variables were allowed to safeguard against biased standard errors and misspecification of the models (see Preacher & Hayes, 2008). Study 1 aimed at exploring the relevance of the included dispositional variables for the prediction of dynamic learned trust in automation. In line with this goal, model fit was optimized in an iterative procedure in which a series of path models were calculated and non-significant paths and variables with low predictive power for the subsequent model levels were excluded. For the evaluation of the model fit the χ^2 test, the comparative fit index (CFI; Bentler, 1990), the root mean square error of approximation (RMSEA, Steiger & Lind, 1980; Steiger, 2016) and the standardized root mean square residual (SRMR, Bentler, 1995) were inspected. The following cut-offs were used as indicators of a good model fit: CFI > .95, RMSEA < .06, and SRMR < .08 (e.g., West et al., 2012; Hu & Bentler,

Figure 5

Study 1: Model fit, standardized path weights, and explained variances for the reduced, final personality model. (Figure from Kraus, Scholz, & Baumann, 2020, ©2020 Human Factors and Ergonomics Society).



Note. * indicates significance (α = .05).

1999). Furthermore, direct and specific indirect effects were inspected (e.g., MacKinnon, 2008). Thereby, specific indirect effects constituted a test for mediations and were examined with the percentile bootstrap procedure (Hayes, 2018). Mediation hypotheses were supported if the resulting 95% confidence interval (CI) of the indirect effects did not include zero.

3.1.3 Study 1 - Results

As the outcome variable in the path models, the measurement of dynamic learned trust after the last video was used. As the assumption of multivariate normality was violated, robust Satorra-Bentler estimates of standard errors and test statistics were reported (Satorra & Bentler, 1994). The hypothesized model was tested in the complete sample. Therefore, to control for a bias of the experimental manipulations, a series of general linear models inspecting interaction effects of each assessed user disposition and the independent variables on the outcome variable was conducted. Three of the personality variables showed significant interactions (neuroticism, self-esteem, and internal LOC). Thus, their relationships in the investigated model might be moderated by the experimental manipulations. A 2x3 ANOVA with the trust scores (after the last video) as dependent variable did not reveal any significant differences between the experimental groups (malfunction: $F(1,268) = 0.16$, $p = .691$, $\eta_p^2 = .001$; transparency: $F(2,268) = 0.18$, $p = .836$, $\eta_p^2 = .001$; interaction: $F(2,268) = 0.46$; $p = .629$, $\eta_p^2 = .003$). As the first step in model exploration, a full parallel-serial-mediation model (Hayes, 2018), estimating all paths between the model levels, was fitted. As a second step, because some variables at the same model levels were considerably correlated, to rule out underestimations of relationships by multicollinearity (Hayes, 2018), a series of mediation models was tested for each trait without the inclusion of the remaining traits at the respective model level. In a third step, the reduced model was repeatedly fitted in an iterative process and the relationships with the lowest path weights were omitted until a satisfying model fit was reached.

On the basis of steps 2 and 3, technological self-efficacy and internal LOC were omitted from the model. This procedure resulted in the model depicted in Figure 5. The model test for the reduced final model showed a good fit to the data, $\chi^2(22, N = 274) = 27.73$, $p = .185$, CFI = .988, RMSEA = .031, SRMR = .037. The pre-ordered variables explained 42.4% of the variance in dynamic learned trust corresponding to a large effect, $f^2 = 0.74$. Between 12.6%, $f^2 = 0.14$, and 37.9%, $f^2 = 0.61$, of the variances of the surface and situational traits were explained, accounting for effect sizes of between weak to medium and large effects. In support of nine of the ten hypotheses, all included direct and indirect effects were found to be significant.

3.1.4 Study 2 - Method

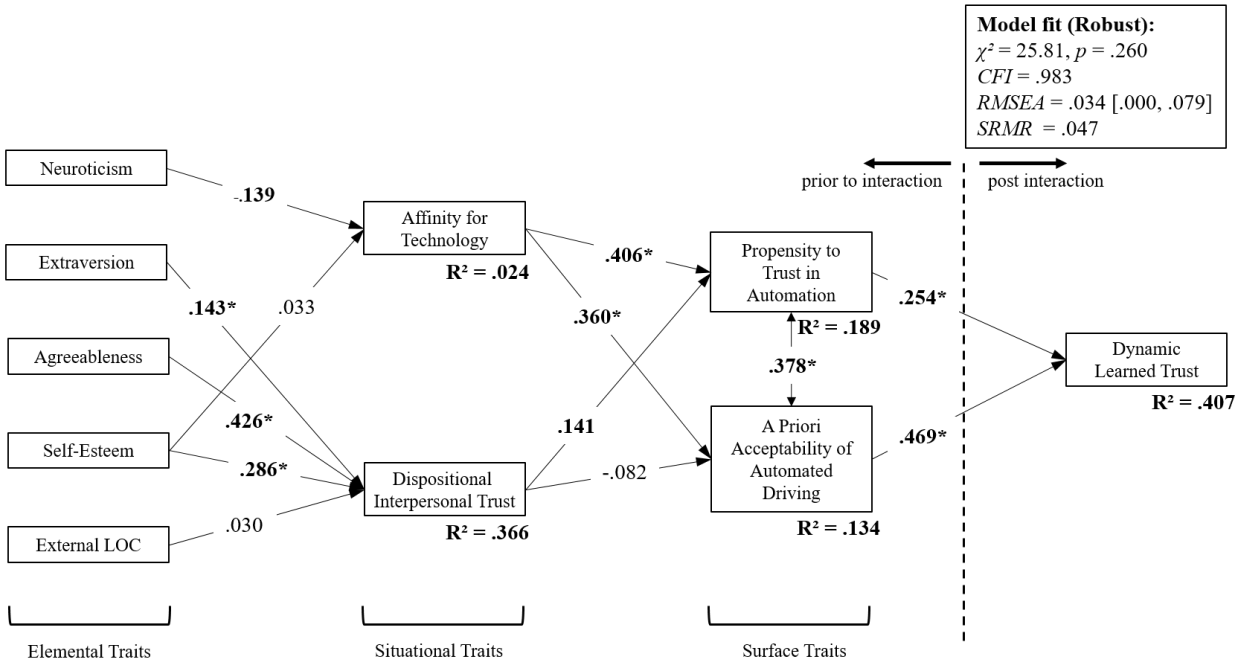
To provide a validation of the personality model explored in study 1, the final model was validated in study 2. The design of study 2 was similar to study 1. Participants rated their trust in the "motorway pilot" repeatedly after the presentation of each one of a series of seven animated videos of an automated driving system in an online study. As in study 1 the trust measure directly after the last video was used as the predicted outcome variable. Again, two independent variables were manipulated: system reliability and transparency. While the reliability manipulation was the same as in study 1, transparency was manipulated in two levels by the absence or presence of a head-up display with system and maneuver status. The final sample of the study consisted of $N = 149$ participants (93 females, 62.4%, 56 males) with an average age of $M_{age} = 29.96$ years ($SD = 11.33$). On average, participants completed the study in 40 minutes. As in study 1, in a series of general linear models, the interactions between the independent variables and the included dispositions were investigated. For two traits (neuroticism and the propensity to trust in automation) significant interactions were found. Thus, for this study, the relationships between these traits and trust in automation might be moderated by the manipulated system characteristics. Again, the experimental groups did not show any significant differences in the trust rating after the last video (malfunction: $F(1,144) = 1.38$, $p = .242$, $\eta_p^2 = .009$; transparency: $F(1,144) = 0.87$, $p = .420$, $\eta p^2 = .012$; interaction: $F(1,144) = 0.84$, $p = .361$, $\eta_p^2 = .006$).

3.1.5 Study 2 - Results

The final model showed a good fit to the data, $\chi^2(22, N = 149) = 25.81$, $p = .260$, CFI = .983, RMSEA = .034, SRMR = .047, replicating the modeled structure of interrelationships between the dispositional variables and dynamic learned trust of the final model of study 1 (see Figure 6). The model explained 40.7% of the variance in dynamic learned trust, corresponding to a large effect ($f^2 = 0.69$). While in this sample, the pre-ordered traits explained only 2.4% ($f^2 = 0.02$, small effect) of the variance in affinity for technology, for the remaining surface and situational traits between 13.4% ($f^2 = 0.15$) and 36.6% ($f^2 = 0.58$) of the variances were explained, corresponding to medium to large effect sizes. Seven of the twelve included path weights were found to be significant in this model. As for the smaller sample size and associated lower power, the point estimates of the standardized path weights were compared to the 95% CIs of those of study 1 (e.g., in adaption of the recommendations by the Open Science Collaboration, 2015, and Cumming, 2013). Three of the non-significant path weights were in the same range as in study 1 (bold with no asterisks in Figure 6), providing some support for their relevance. The estimated standardized weights for the paths from self-esteem to the affinity for technology, from external LOC to dispositional interpersonal trust and from dispositional interpersonal trust to the a priori acceptability were neither significant nor did a CI overlap indicate a comparable effect to the estimates in study 1. For all pre-ordered traits except for agreeableness, at least one specific indirect effect was in the

Figure 6

Study 2: Model fit, standardized path weights, and explained variances of the validation for the reduced, final personality model. (Figure from Kraus, Scholz, & Baumann, 2020, ©2020 Human Factors and Ergonomics Society).



Note. * indicates significance (α = .05); boldly printed path weights indicate that these paths were found to be significant in study 1 and that the point estimate of study 2 lies within the respective 95% CI of study 1.

same range of the significant effect from study 1. Nevertheless, a relation between agreeableness and interpersonal trust was also found in this study. Taken together, the combined evidence of study 2 provides at least some support for 8 of the 10 study hypotheses, while the integration of LOC and technological self-efficacy in the combined consideration with the remaining variables into the hierarchical personality model was not supported and the indirect effect of agreeableness on trust in automation through dispositional interpersonal trust needs further investigation.

3.1.6 Study 1 and Study 2 - Conclusion

This research was the first of its kind in combining personality traits and technology-related general attitudes at different levels of specificity into one hierarchical personality model for the prediction of dynamic learned trust in automation. In both studies, the final model showed a good fit and the included pre-ordered variables explained substantial variance proportions in both dynamic learned trust and the propensity to trust in automation. Especially the propensity to trust in automation and the general attitude towards the technological domain of automated driving showed strong relationships with dynamic learned trust in the investigated automated driving system. These, in turn, were found to be related to the more general situational trait the affinity for technology. Furthermore, the general disposition to trust in other people predicted the propensity to trust in automation at the surface level. In terms of the most basic included variables, neuroticism was found to be negatively related to the affinity for technology and extraversion, agreeableness and self-esteem were positively related to dispositional interpersonal trust. Taken together, the inclusion of eight of the investigated dispositions at the different levels of the model hierarchy was supported. Their role for trust formation and calibration is discussed in more detail in chapters 4.1 and 4.2. The relationship between LOC and trust in automation was not conclusively supported in

these studies and might be addressed in future studies (see chapter 4.2). In the same manner, the role of agreeableness for trust in automation needs further investigation.

The findings of these studies support some of the basic assumptions of the proposed framework. First, they underline the role of personality and general attitudes for the formation of the propensity to trust in automation (stage 1, proposition 4), which, in turn, was found to influence dynamic learned trust at stage 3 (proposition 8.1). Accordingly, second, the findings provide some support for the general notion of the framework that trust formation begins before the interaction with an automated system and that different trust layers are built up in a sequential manner (propositions 1, 3, and 8). This is indicated by the considerable correlation between the propensity to trust in automation measured at the beginning of the study before the introduction of the system and dynamic learned trust measured at the end of the study after the stimulus presentation. Third, the relationship between interpersonal trust and the propensity to trust supports the idea that interpersonal trust and trust in automation build on similar psychological processes (e.g., Lee & Moray, 1992; see chapter 1.3). At the same time, the reported proportions of shared variance also suggest that while there seem to be some conceptual overlaps between these traits, trust in automation has a large proportion of unique variance seemingly resulting from processes different from those in interpersonal trust. In this regard, the considerable relationships between the affinity for technology, general technology attitudes, and dynamic learned trust point into the direction that the personal learning history with automated technology might be one of these specific processes essential for the formation of trust in automation (see chapter 4.2 for a more detailed discussion).

It can be subsumed, that personality differences are important to be considered at every stage of trust formation both in research and in the practical implementation of automated systems. In this regard, the supported mediation paths from more general traits via specific dispositions to dynamic learned trust extent the understanding of trust processes indicating that besides the individual learning history as reflected in technology attitudes and the disposition to trust in interpersonal contexts, also the dispositional self-evaluation plays an important role for the formation of trust in automation. In this regard, it might be that people who hold a positive view of themselves generalize this self-assessment to the ability to functionally and effectively use automated systems. This relationship is further investigated in study 3. To summarize, it seems essential to address individual user dispositions in both future research on trust in automation and in the design and dissemination of automated driving systems (see chapter 4.3 for a more detailed discussion).

3.2 Study 3: Scared to Trust? – Predicting Trust in Highly Automated Driving by Depressiveness, Negative Self-Evaluations and State Anxiety

Kraus, J., Scholz, D., Messner, E.-M., Messner, M., & Baumann, M. (2020). Scared to trust? – Predicting trust in highly automated driving by depressiveness, negative self-evaluations and state anxiety. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02917

3.2.1 Introduction and Study Hypotheses

Building on the results of studies 1 and 2 in support of the role of personality in the formation of trust in automation and the proposition that user states influence the formation of learned trust, study 3 investigated the mediation of effects of personality traits to dynamic learned trust in automation through the experience of emotional states. This hypothesized mediation is theoretically founded in studies supporting a fundamental impact of personality for the experience of emotional states (e.g., Larsen & Ketelaar, 1991; Letzring & Adamcik, 2015) as well as the affect-as-information model (e.g., Schwarz & Clore, 1988) and the affect infusion model (e.g., Forgas, 1995), which propose emotions as an essential source for judgments. Evidence for such a mediation process would, again, support a sequential history-based process of trust formation (proposition 1 and 3), in which user personality affects emotional experiences in the process of familiarization with automation and, eventually, the formation of learned trust (proposition 8.1).

Against this background, in this study, a mediation of the effects of several personality traits on dynamic learned trust in automation through state anxiety was investigated. More specifically, depressiveness and negative dispositional self-evaluations were hypothesized to be associated with an experience of higher levels of state anxiety at the time of introduction to a new automated system, which in turn was hypothesized to be associated with lower trust in this system.

While differential effects of negative and positive emotions on trust in automation have been investigated before (e.g., Merritt, 2011; Frison et al., 2019; see chapter 1.4.3), research on the influences of state anxiety on trust in automation is scarce. Anxiety was defined as an emotional state "characterized by subjective, consciously perceived feelings of tension and apprehension, and heightened autonomic nervous system activity" (Horikawa & Yagi, 2012, p. 1) and was found to influence the processing of ambiguous stimuli (e.g., Mathews & Mackintosh, 1998). Essentially, the experience of anxiety might influence trust through the assessment of risks in the interaction with such a technology as a theoretical link between the individual tendency to experience anxiety, risk aversion, and trust suggests (in the interpersonal context discussed, e.g., by Thielmann & Hilbig, 2015). On this basis and in line with both the affect-as-information and the affect infusion model, it was hypothesized:

> *Hypothesis 1 (H1): Higher state anxiety at the time when an unfamiliar automated driving system is introduced is associated with lower trust in the system.*

As individual differences in anxiety have been found to be predicted by personality, amongst others by depression (Lovibond & Lovibond, 1995) and self-evaluations (e.g., Xiao et al., 2014), it can be concluded that depressiveness is positively and dispositional positive self-evaluations (self-esteem, self-efficacy, and an internal LOC) are negatively associated with individual differences in the experience of anxiety at the time of being introduced to a new automated system. In line with this, depressive symptoms are commonly found to be associated with higher levels of anxiety (e.g., Murphy et al., 2004; Cummings et al., 2014). This relationship was also found for subclinical

depression (e.g., Lovibond & Lovibond, 1995). As depression and diminished levels of interpersonal trust have also been associated (e.g., Lester & Gatto, 1990; Kim et al., 2012), such a relationship might generalize to limitations of depressive people to build up trust in automated systems. It was hypothesized:

Hypothesis 2 (H2): Depressiveness is negatively related to trust in an automated driving system.

Hypothesis 3 (H3): The effect of depressiveness on trust in an automated driving system is mediated by the experience of anxiety at the time of the introduction of the system.

Based on several studies supporting an interrelation between self-confidence and trust in automation (see chapter 1.4.3), this study further investigated the role of the dispositional self-evaluations self-esteem (e.g., Harter, 1993), self-efficacy (Bandura, 1991) and internal LOC (Rotter 1966, 1990) for the experience of anxiety when being introduced to a new automated system and how these relationships are associated with dynamic learned trust in the system. Self-esteem, as a positive global evaluation of oneself (e.g., Harter, 1993), has been found to be associated with self-confidence (e.g., Owens, 1993). Also, self-efficacy has been discussed as closely related to self-confidence (e.g., Schunk, 1991). Thus, it can be assumed that these traits lead to a more positive evaluation of the own ability to successfully and safely interact with automated technology in line with one's needs and goals and, on this basis, a tendency to trust such systems more (as for example for the own ability to react adequately to any difficult or unforeseen situations in the course of the interaction). Furthermore, people with an internal LOC view significant outcomes to be within their own control and not mainly controlled by external, uncontrollable events, which, on the other hand, externals perceive to control their fate (e.g., Ajzen, 2002). In line with this, an internal LOC has been related to increased intrinsic task motivation and problem-focused coping strategies (e.g., Galvin et al., 2018). Consequently, internals should perceive themselves as being more in control of a positive interaction with automated technology and the self-awareness of this capability might, in turn, establish a tendency to trust such technology more. On this basis, it is hypothesized:

Hypothesis 4 (H4): Self-esteem is positively related to trust in an automated driving system.

Hypothesis 5 (H5): Self-efficacy is positively related to trust in an automated driving system.

Hypothesis 6 (H6): An int. LOC is positively related to trust in an automated driving system.

Also, the associations of these traits with anxiety have been well supported in the literature. Self-esteem has been found to be negatively related to anxiety, both in cross-sectional and longitudinal designs (Sowislo & Orth, 2013). Additionally, a negative correlation between task self-efficacy and computer anxiety has been reported (Brosnan, 1998). Also, an internal LOC has consistently been found to be related to lower levels of anxiety than an external LOC (e.g., Spector, 1982). From these theoretical and empirical associations of the investigated traits to both the experience of anxiety and trust in automation, it can be hypothesized that people who perceive themselves as being more capable of succeeding in various tasks and in control of the related outcomes, also experience less anxiety when being introduced to an unfamiliar, potentially dangerous, and risky automated driving system. On the basis of H1 and H4-6, it can be further hypothesized that through this reduced anxiety, these people also tend to trust such a system more. As in the case of depressiveness, the relationships of the included self-evaluative traits to trust in automation are hypothesized to be mediated by state anxiety:

Hypothesis 7 (H7): The effects of self-esteem (H7.1), self-efficacy (H7.2), and locus of control (H7.3) on trust in an automated driving system are mediated by the experience of anxiety at the time of introduction of the system.

3.2.2 Method

The study hypotheses were investigated in a driving simulator experiment, in which the variables of interest for this study were collected at the beginning of the study procedure before and directly after a practice trial. As not of interest for this study, the complete study procedure is described in detail in chapter 3.4 (study 6). After being welcomed, participants answered questionnaires on the investigated personality variables and several emotional states (positive and negative affect, state anxiety) before being introduced to an automated driving system (but after being introduced to the forthcoming interaction with an automated driving system). Afterwards, they drove with the system in a practice trial, both in manual and automated mode, which was followed by a measurement of their trust in the system with the LETRAS-G (see Appendix A). This measurement was used as the outcome variable in this study. For the assessment of the investigated personality variables, the following questionnaires were used: Becks Depressions Inventory-II (BDI-II, Hautzinger et al., 2009) for depressiveness, the TOR-IE by Rotter (1966; with higher values indicating a more external LOC) for LOC, the Rosenberg Self-Esteem Scale (Ferring & Filipp, 1996), the General Self-Efficacy Scale (GSE; German version; Jerusalem & Schwarzer, 1999), the PANAS (Krohne et al., 1992) for positive and negative affect, and the STAI-6 (Marteau & Bekker, 1992) for state anxiety. The final sample consisted of $N = 47$ participants (27 females, $M_{age} = 27.45$, $SD = 9.75$).

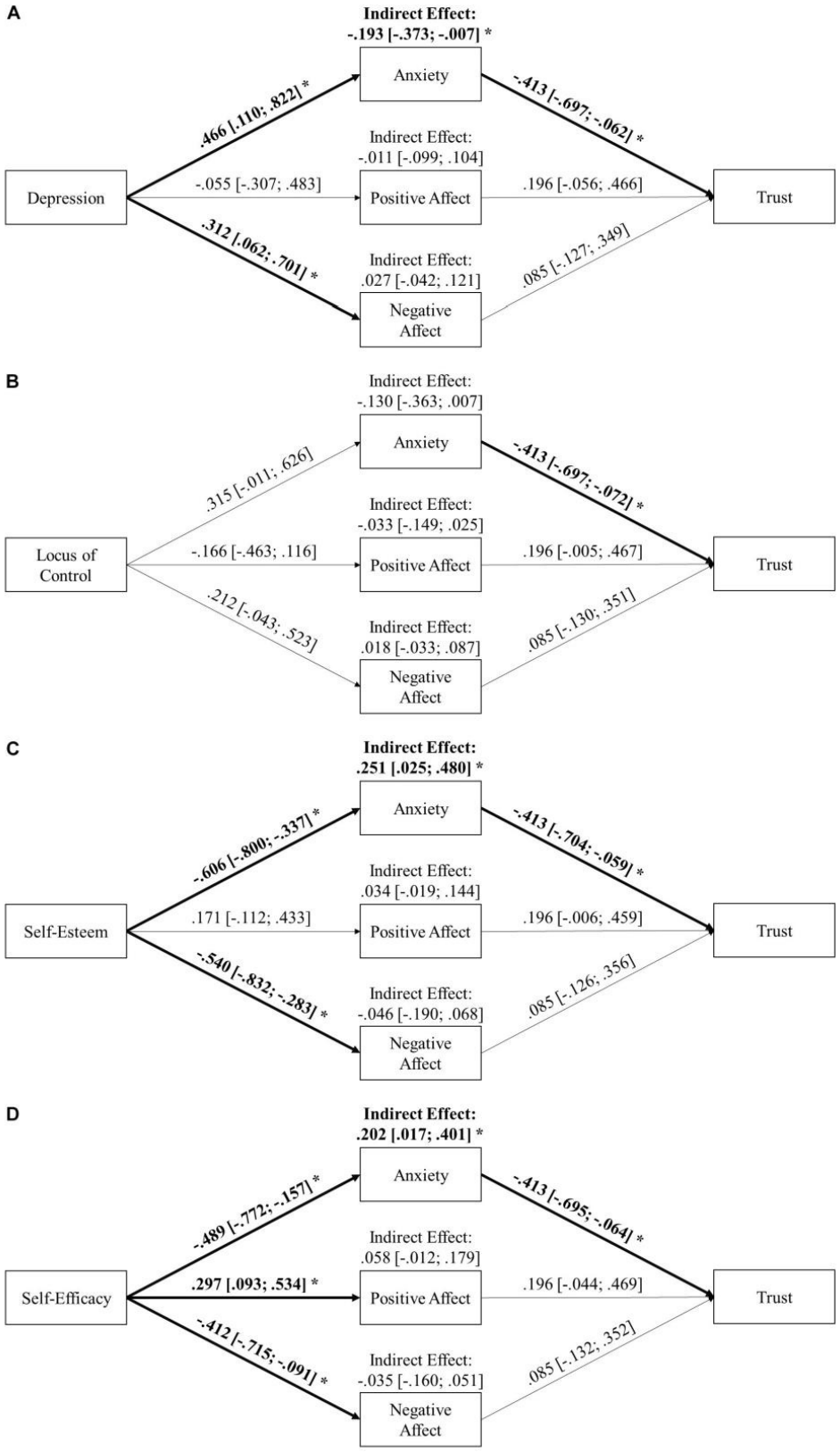
For testing the direct bivariate relationships (H1, H2, H4, H5), Pearson correlations were calculated. For the mediation hypotheses (H3 and H7), path models with the R package lavaan (Rosseel, 2012; version 0.6-2; R version 3.4.3) were calculated and bootstrapped direct and specific indirect effects were investigated (5000 samples, Hayes, 2009) and the hypotheses were evaluated by inspecting the 95% CIs. As a first step in the analysis, four path models for the individual mediation effects from each of the investigated personality traits were calculated in which, in addition to state anxiety, also positive and negative affect were included to assess the specific effect of anxiety in combination with these two emotional states (see Figure 7 A-D). As a second step, a path model with all four personality traits as predictors and state anxiety as the only mediator was calculated to investigate specific indirect effects on trust of the personality traits in a combined consideration. Standardized effects were reported. Some of the variables showed derivations from normality, as indicated by a series of Shapiro-Wilk tests. As the Pearson coefficient has been shown to be robust against such violations (e.g., Bishara & Hittner, 2012) and bootstrapping was used in the path models (making distribution assumptions obsolete), the reported findings can be validly interpreted.

3.2.3 Results

The hypothesized relationship of state anxiety and trust (H1) was supported by a significant negative correlation, $r = -.400$ ($p < .001$), and a significant negative direct effect in the path models, $\beta = -.413$, [-.697; -.062], (see Figure 7). The correlations for positive and negative affect with trust in automation were not found to be significant and neither were any of the direct or indirect effects on trust in the mediation models (see Figure 7). While depressiveness was not directly related to trust, $r = -.262$, $p > .05$, a significant negative indirect effect of depressiveness on trust mediated by state anxiety, $\beta = -.193$, [-.373; -.007], in partial support of H2 and full support of H3 was found (Figure 7A). The hypotheses on the positive relationships of self-esteem and self-efficacy with trust in automation were fully supported (H4, H5, H7.1, H7.2). Self-esteem, $r = .370$, $p < .05$, and

Figure 7

Results of the separate path models for the mediations of the effects of the four investigated personality variables on trust in automation by the three investigated emotional states. (Figure from Kraus, Scholz, Messner, et al., 2020, CC BY 4.0, https://creativecommons.org/licenses/by/4.0/).



Note. Asterisks indicate significance at α = .05; above each of the emotional states the respective specific indirect effect is depicted; bootstrapped 95% CIs are depicted.

self-efficacy, $r = .291$, $p < .05$, both showed significant positive correlations with trust in automation and these relationships were mediated by state anxiety (self-esteem, H7.1; $\beta = .251$, [.025; .480]; Figure 7C, and self-efficacy, $\beta = .202$, [.017; .401], Figure 7D). For LOC, neither bivariate correlations nor direct or indirect effects to trust in automation were found to be significant in contradiction to H6 and H7.3 (Figure 7B). In the path model which investigated the paths from the four personality traits simultaneously, only the direct effect from self-esteem on state anxiety remained significant ($\beta = -.459$, [−.800; −.019]). This indicates that, in line with their substantial theoretical overlap, the investigated personality traits showed a considerable degree of multi-collinearity and the relative effect on dynamic learned trust in automation was highest for self-esteem.

3.2.4 Conclusion

First of all, the study supported a prediction of individual differences in dynamic learned trust in an automated driving system after the first experiences with that system by the individual level of anxiety at the time of introduction of the system. In this study, the relationship between anxiety and dynamic learned trust was higher than that of positive or negative affect to trust. Furthermore, personality differences in regard to dispositional self-evaluations predicted differences in the experience of anxiety and in dynamic learned trust. More specifically, as indicated by the significant mediations of the effects of self-esteem and self-efficacy on dynamic learned trust through state anxiety, people who hold a favorable view about themselves and perceive themselves as able to perform in various task successfully, are less anxious when being confronted with a new technological system and in turn trust it more after some interaction. Again, comparable to the findings of studies 1 and 2, the effects associated with LOC were smaller than those for self-esteem and self-efficacy and the role of this trait for the formation of trust in automation remains unclear (as further discussed in chapter 4.1). As supported by an indirect effect of depressiveness on dynamic learned trust, a pessimistic view on the world and on one's own capabilities might emphasize the consideration of potential risks and dangers of automated driving systems. The nature of this effect could be addressed in future studies.

Taken together, this study is the first of its kind to provide evidence for a mediation of the effect of personality variables on dynamic learned trust by state anxiety. Notably, the influence of state anxiety before the interaction on dynamic learned trust after some experience with a system provides a new finding in the area and again indicates that interpersonal differences influence the process in which trust in automation is built up. The supported mediations provide various implications on a theoretical and practical level regarding the role of personality and emotional states in the familiarization with automated systems. In terms of the Three Stages of Trust framework, this study underlines the influence of personality differences not only in the early stages of trust formation but also in the formation of dynamic learned trust at stage 3 (proposition 8.1). Above this, the reported mediations of these effects by the experience of anxiety when being first introduced to the system (stage 2) further underline the interrelation of the three trust stages (propositions 1 and 3). Furthermore, the findings suggest a possible process through which personality might affect the perception of unfamiliar automated systems in terms of an attribution of experienced affect to the evaluation of an automated system in line with the affective route proposed by Lee and See (2004) and, for example, the affect-as-information model (e.g., Schwarz & Clore, 1988). As discussed in chapters 4.1, 4.2, and 4.3, the reported study findings support a consideration of dispositional self-evaluations, anxiety in the face of new technology as well as their association in both research and practice in regard to automated technology.

3.3 Study 4: Two Routes to Trust Calibration: Effects of Reliability and Brand Information on Trust in Automation

Kraus, J. M., Forster, Y., Hergeth, S., & Baumann, M. (2019). Two routes to trust calibration: Effects of reliability and brand information on trust in automation. *International Journal of Mobile Human Computer Interaction, 11*(3), 1-17. https://doi.org/10.4018/IJMHCI.2019070101

3.3.1 Introduction and Study Hypotheses

Especially in suspicious or anxious users, the information prior to the first system use might be essential to generate an initial trust level exceeding a minimum threshold necessary for the intention to use an unfamiliar system at all. At the same time, from a perspective of trust calibration, the initial trust level should not exceed a realistic assessment of system capabilities leading to overtrust and system misuse. As for this vital role of prior system information, it is essential to gain a deeper understanding of how this information is used in the process of trust formation and calibration to build up a realistic picture of the automated system. The presented framework proposes that the formation of learned trust in automation is based on a system expectation built up using available information (trust cues, proposition 5). It is further assumed that at stage 2 trust cues affect initial learned trust through attitude formation processes (proposition 6). Against this background and based on the notion by Lee and See (2004) that trust can be influenced by provided information through three different routes – the analytical, the analogical and the affective route – this study investigated if trust cues entailed in prior system information are used in line with the predictions of the Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1986; see chapter 2). In this study, the effect of two types of information about an automated driving system (information about system reliability and the manufacturer) on initial learned trust was investigated. In line with the assumptions of the ELM, both types of information were hypothesized to influence initial learned trust in their respective direction (either positive or negative). First, the presentation of prior reliability information was hypothesized to result in attitude changes through the central route of processing. In line with this, in two studies with driving automation at different levels it was found that prior information on the systems' functional scope affected initial trust levels (Begiatto & Krems, 2013, Körber et al., 2018). Accordingly, it was hypothesized:

Hypothesis 1 (H1): An automation that is described as reliable is trusted more than one that is described as unreliable.

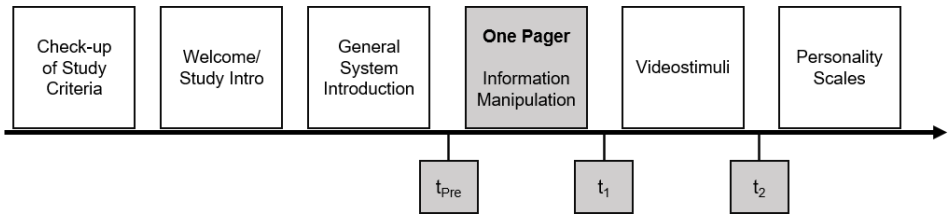
Second the information about the system manufacturer (original equipment manufacturer, OEM brand) was manipulated in the study. This more heuristic information was assumed to affect trust formation through the ELM's peripheral route. It was hypothesized:

Hypothesis 2 (H2): A better OEM brand reputation of an automated system leads to higher trust in the system.

Elaboration of the content and arguments of a presented message depends on the motivation and cognitive capacity to do so (Petty & Cacioppo, 1986). As mere support for these two hypotheses would not provide evidence for the effect of two routes in trust formation, additionally, the motivation of users to elaborate the content of the received information seemed worthwhile to be considered in the study. On the level of individual differences, need for cognition (NFC), defined as an "individual's tendency to engage in and enjoy effortful cognitive endeavors" (Cacioppo et al., 1984, p. 306), has been found to influence the probability of elaboration. For example, findings

Figure 8

Procedure of study 4. (Figure from Kraus, Forster, et al., 2019, ©2019 IGI Global, www.igi-global.com. Posted by permission of the publisher).



from a study by Cacioppo et al. (1983) support the notion that people with higher NFC engage in central processing more often than those with lower NFC. In line with this, it was hypothesized:

Hypothesis 3 (H3): NFC moderates the usage of information as follows: People with higher NFC use the reliability information more than people with lower NFC (H3.1), People with lower NFC use the brand reputation information more than people with higher NFC (H3.2).

Based on the theorizing of studies 1 and 2, three further personality traits were included in the study: regulatory focus (promotion and prevention focus, Higgins, 1997), materialism (Belk, 1985) and the PAS (Merritt, Unnerstall, et al., 2015) with its two factors (see chapter 1.4.1, please find a more detailed argumentation for their inclusion in Appendix B). Based on their theoretical and empirical relevance for trust in automation, it was hypothesized:

Hypothesis 4 (H4): Regulatory focus, materialism, as well as the perfect automation scheme, predict initial learned trust in an automated system.

3.3.2 Method

Study hypotheses were investigated in an online experiment with a two-factorial design in which the presentation of information on system reliability was manipulated in two (low vs. high) and the OEM brand was manipulated in seven levels (baseline without brand and six different OEM brands). Additionally, NFC was assessed and the sample was split into two groups on the basis of a median split as an additional factor for some analyses. The included OEM brands were selected on the basis of results from a cluster analysis in a pilot study to resemble three categories of brand reputation (below average: Kia and Dacia; average: Opel and Skoda; above-average: BMW and Tesla). The sample was recruited with an online panel company. Participants were randomly assigned to one of the 14 combinations of the two manipulated factors. The dependent variable trust in automation was measured three times during the experiment (see Figure 8 for the study procedure). At the beginning of the study, participants read a general description of automated driving systems, after which trust in the system was assessed for the first time (t_{pre}). Then a one-pager was presented in which a conditionally automated driving system was introduced along with the results of a fictive road test. As a manipulation of the brand reputation factor, the manufacturer of the system was prominently mentioned two times in the one-pager. Information on system reliability was manipulated in the provided results of the (fictive) road test. While in the high-reliability condition the system was said to have had nine malfunctions during the test and ranked second of 22 tested systems, in the low-reliability condition the system was reported to have shown 90 errors and ranked 16th. After the presentation of the one-pagers, trust was rated a second time (t_1). After this, participants were presented with four videos allegedly showing a driver using the system in a driving simulator (about 30 seconds each). The videos entailed different interactions with the system: activation by the driver, automated overtaking, system-initiated TOR, and a deactivation of

Table 1

Means and standard deviations for initial learned trust in automation at t_1 and t_2 split by reliability and brand reputation groups (left) and split by reliability groups and median-split groups for NFC (right). (Table adapted from Kraus, Forster, et al., 2019. ©2019 IGI Global, www.igi-global.com. Posted by permission of the publisher).

			Brand reputation			Need for cognition median-split	
Reliability			<average	average	>average	NFC low	NFC high
Trust t_1	low	*M*	3.29	2.99	3.41	3.30	3.16
		(SD)	(1.01)	(1.06)	(1.09)	(1.04)	(1.09)
	high	*M*	3.88	3.84	4.00	3.74	4.13
		(SD)	(0.94)	(1.15)	(1.12)	(0.99)	(1.13)
Trust t_2	low	*M*	3.85	3.66	3.87	3.82	3.76
		(SD)	(1.23)	(1.19)	(1.23)	(1.14)	(1.31)
	high	*M*	4.30	4.20	4.16	4.04	4.47
		(SD)	(1.00)	(1.19)	(1.21)	(0.99)	(1.26)

Notes. *M* = Mean, *SD* = Standard Deviation

the system by the driver. After the videos, participants rated their trust in the system a last time (t_2). Afterwards, the personality questionnaires were filled out and participants were debriefed.

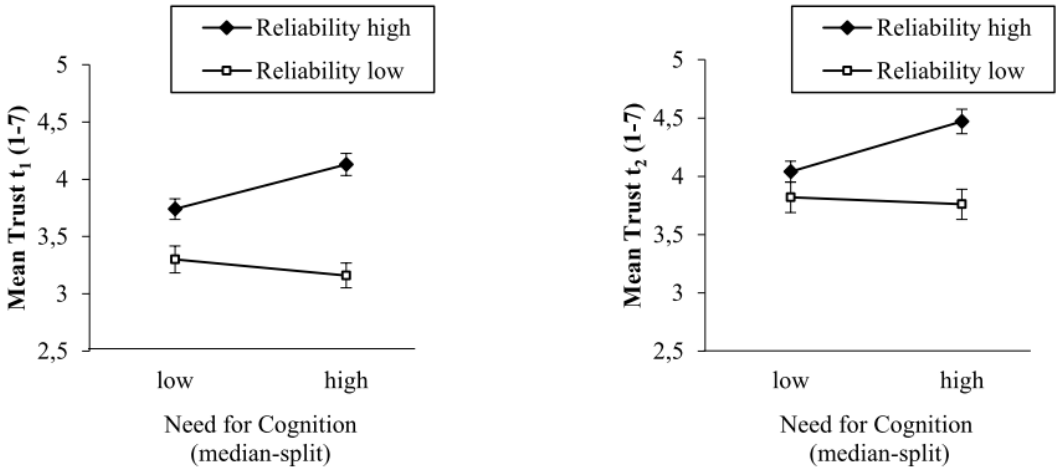
Trust in automation was measured with a 12-item German translation of the ATS Scale (Jian et al., 2000). For assessing NFC, eight items of the German scale by Bless and colleges (1994) were selected on the basis of item difficulty and factor loadings. For the assessment of the regulatory focus, the scale by Fellner et al. (2007) was used. For the two remaining personality variables, scales were shortened on the basis of factor loadings: the scale by Richins and Dawson (1992) for materialism and the scale by Merritt, Unnerstall, et al. (2015) for the PAS. From $N = 623$ participants, who completed the survey, 104 had to be excluded for one of the following two reasons: unsuccessful manipulation checks (wrongful memory of the brand or the overall result of the road test in the one-pager) or speeding (study completion in less than 24 min.) resulting in a final sample of $N = 519$ (55.3% females, $M_{age} = 45.5$, $SD = 14.35$).

3.3.3 Results

As for unequal group sizes, type-III sum of squares (Howell, 2007) were reported. At t_{pre}, the experimental study groups did not significantly differ in their trust mean scores, as indicated by a two-way ANOVA with brand reputation and reliability as between factors, $F_{max} = 1.08$, all $p > .05$. In line with the results of the pilot study, OEM brands were grouped in the three OEM reputation groups. For investigating the hypotheses on the effects of the information manipulations, the baseline condition was left out, resulting in an $n = 435$ in these analyses. For the correlations of the personality traits and trust the full, cleaned sample was used. A manipulation check with an ANOVA (reputation groups as between factors, the average image score of the pilot study as dependent variable, see Forster et al., 2018) indicated a successful manipulation, $F(2,432) = 27.49$, $p < .01$, $\eta_p^2 = .113$, $M_{below\ average} = 4.03$, $SD = 1.09$, $M_{average} = 4.41$, $SD = 1.16$, $M_{above\ average} = 4.95$, $SD = 0.94$. In Table 1, trust means for the six study groups at t_1 and t_2 are presented on the left side. For testing the moderation of information use by NFC (H3), a median split in the sample without the baseline condition was conducted ($Mdn = 4.75$) resulting in two groups: low NFC, *n*

Figure 9

Effects of reliability information and NFC median-split groups on trust in automation. Error bars reflect 1 SE. (Figure from Kraus, Forster, et al., 2019, ©2019 IGI Global, www.igi-global.com. Posted by permission of the publisher).



= 240, $M_{NFC\ low}$ = 4.10, *SD* = 0.54, and high NFC, *n* = 195, $M_{NFC\ high}$ = 5.44, *SD* = 0.48. On the right side of Table 1, the respective group means for the groups by the NFC median split and the reliability manipulation are displayed. H1-3 were tested with two three-factorial ANOVAs (factors: reliability, brand reputation, NFC median-split) and trust at t_1 and t_2 as the dependent variables. Neither at t_1 nor at t_2 a main effect for the NFC median-split groups was found. In line with the hypothesis that reliability information affects trust in automation (H1), in both ANOVAs a main effect for reliability was supported, t_1: $F(1,423) = 48.14$, $p < .001$, $\eta_p^2 = .102$; t_2: $F(1,423) = 16.34$, $p < .05$, $\eta_p^2 = .037$. In support of an influence of brand reputation on trust (H2), at t_1, a significant main effect was found for the OEM brand reputation groups, $F(2,423) = 3.17$, $p < .05$, $\eta_p^2 = .015$, while there was no effect of the brand reputation at t_2, $F(2,423) = 0.58$, $p = .559$, $\eta_p^2 = .003$. A Tukey's HSD post hoc test indicated that the effect at t_1 reflected a significant difference between the average and the above-average reputation group. The two-way interaction between the two manipulated variables was not found to be significant at both times of measurement. For the evaluation of the moderation of information use by NFC (H3), the two-way and three-way interactions including the NFC median-split groups were inspected. The two-way interactions with brand reputation and the three-way interactions were neither found to be significant at t_1 nor t_2. Thus the two NFC groups did not show to influence the use of brand reputation information (H3.2). On the other hand, NFC was found to influence the use of the reliability information (H3.1) as supported by a significant interaction between the reliability condition and the NFC median-split groups for trust at t_1, $F(1,423) = 6.84$, $p < .01$, $\eta_p^2 = .016$, and t_2, $F(1,423) = 4.54$, $p < .05$, $\eta_p^2 = .011$ (Figure 9). The interactions indicate that the trust difference between the high and low reliable systems is higher in the high NFC as compared to the low NFC group at both times of measurement. In line with H3.2, people with higher NFC seem to use the reliability information more to form their trust than those with lower NFC. A further analysis of a quartile split of NFC indicated a non-linear relationship between NFC and the extent to which the reliability information was used to form trust in the system. To evaluate the prediction of trust in automation by the investigated personality traits (H4), bivariate Pearson correlations were inspected. For NFC, no significant correlation with trust in automation was found. While the correlations between the PAS factor "all or none" at t_{pre} as well as the correlation between promotion focus and trust at t_1 were not found to be significant,

for the remaining personality variables, significant relationships with trust were found at all three times of measurement ($r_{promotion\ focus}$ = .108 – .183, $r_{prevention\ focus}$ = .127 - .139, $r_{materialism}$ = .191 - .238, $r_{PAS\ high\ expectations}$ = .599 - .613, $r_{PAS\ all\ or\ none}$ = -.090 to -.159, all $p < .05$).

3.3.4 Conclusion

The reported findings support an influence of prior system information about system reliance and brand on the formation of initial learned trust. While reliability information influenced trust at both times of measurement, an effect of brand reputation was only supported for a short time interval between the presentation of the information and the trust measurement. In this regard, in line with the assumptions of the ELM, the effect of reliability information, which is supposedly processed through the central route, showed higher stability over time as compared to the effect of the more heuristic brand information, which is likely to be processed through the peripheral route. Furthermore, differences in the personality variable NFC affected the weighting of the reliability information in trust formation: participants with a higher NFC changed their trust ratings more into the direction of the provided information as those with a lower NFC. Furthermore, the results support relationships of three personality traits to initial learned trust in automation. First, the significant relationships of both layers of the regulatory focus support the role of individual differences in the consideration and weighting of gains and losses in trust formation. Second, people with higher materialism tended to trust the driving automation more. This points into the direction that the higher importance materialists put on material goods might be associated with an overly optimistic evaluation of innovative technology and that this is a possible source of bias in the direction of overtrust. Third, a considerable positive relationship of the PAS factor "high expectation" with initial trust was found, which is in line with earlier research (e.g., Lyons & Guznov, 2019). The factor "all or none" showed smaller, yet significant negative relationships with initial learned trust.

Taken together, this research is among the first to investigate trust processes in automated driving systems in terms of dual-process theories and by this supports several of the propositions of the Three Stages of Trust framework. Essentially and central for the framework, the findings support a transfer of attitude formation processes to the process in which initial learned trust is built up on the basis of prior system information (propositions 5 and 6). Furthermore, a moderation of information processing at stage 2 by individual user differences was supported (proposition 9). Also, in line with the findings of studies 1 to 3, the role of personality and general technology-related attitudes for individual differences in initial learned trust was supported (proposition 8.1). These findings underline the essential relevance of trust formation processes before the actual interaction with an automated system for trust processes. While the study groups started with comparable trust levels in the experiment, the presentation of reliability information was associated with stable trust differences, which presumably relate to different expectations towards the system, which might influence trust calibration during later system use. Taken together, these findings point into the direction that the transfer of theories and models from the domain of attitude formation to the research on trust in automation seems fruitful (as discussed in more detail in chapter 4.2.3). Furthermore, these findings entail implications for the design of prior system information with the goal of facilitating initial trust calibration. A detailed discussion of the practical implications of a transfer of attitude theories is provided in chapter 4.3.

3.4 Studies 5 and 6: The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction and System Transparency

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

3.4.1 Introduction and Study Hypotheses

The Three Stages of Trust framework proposes that trust is built up and calibrated prior and during the interaction with an automated system along the provided information at stages 2 and 3, at which the user learns to trust a system (proposition 5). Against this background, to further enhance the understanding of the nature of the proposed feedback process at stage 3, studies 5 and 6 investigated how dynamic learned trust in automation is calibrated during early use of an automated driving system. In each of the studies a driving simulator experiment was conducted in which trust in automation was measured prior and repeatedly during the interaction with an automated driving system in different conditions. Both studies compared trust calibration in conditions with and without system malfunctions, while in study 5, participants additionally experienced TORs in both conditions. Furthermore, the effect of prior system information about system limitations on the interpretation of a system malfunction was investigated in study 6. In combination, these research questions represent a test of several of the propositions of the Three Stages of Trust framework. First, the character of the comparison between system expectation and actual system behavior in the proposed feedback loop of trust calibration was investigated (proposition 7.2). In this regard, additionally, study 6 aimed at an investigation of how prior system information affects the calibration of dynamic learned trust at stage 3 (proposition 8.2).

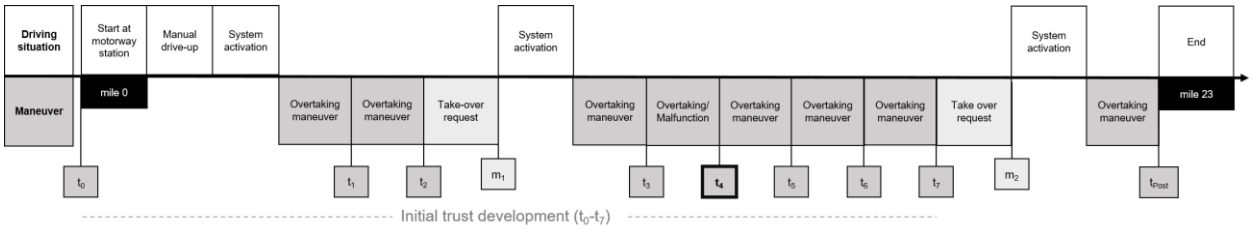
In several studies which measured trust repeatedly over time with flawless system functioning, trust in automation was found to increase in different settings and over different periods of time: in simulated automated process control scenarios over the course of a single episode of system interaction (e.g., Muir & Moray, 1996; Moray & Inagaki, 1999) and in real-life use of adaptive cruise control (ACC) over the first days (Kazi et al., 2007) and weeks (Beggiato & Krems, 2015). Several studies (Hergeth et al., 2016, 2017, 2015; Hartwich et al., 2019) provided initial evidence that such a trust increase also occurs in the interaction with higher-level driving automation. Accordingly, it was hypothesized:

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

In highly automated vehicles (SAE levels 3 and 4), automated driving is restricted by system limitations and system malfunctions (see chapter 1.2). The depicted studies investigated the effect of both, TORs and system malfunctions, on the formation of dynamic learned trust during the first drive with an automated driving system. As users were found to tend to expect automated technology to work extremely reliable or even flawlessly (e.g., Dzindolet et al., 2002; Merritt, Unnerstall, et al., 2015), derivations from perfection should be associated with trust decreases. For system

Figure 10

Procedure and experimental drive of study 5. (Figure from Kraus, Scholz, et al., 2019. ©2019 Human Factors and Ergonomics Society.)



limitations, such trust decreases were shown with single-item trust measures in the domain of highly automated driving (Hergeth et al., 2015) and for system malfunctions in automated decision aids (Dzindolet et al., 2003). In regard to the processes after such a trust decrease, exemplarily, research in the domain of automated plant simulations (Moray & Inagaki, 1999; Lee & Moray, 1992) and automated driving (Hergeth et al., 2015) showed that trust in these automated systems after a transition of control or error was at least to some degree remediated in further system use. Based on this, it was hypothesized that such a re-establishment of trust also takes place in continued interaction with an automated driving system after experiencing a TOR or a malfunction:

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

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).

These temporal trust decreases may arise from a lack of knowledge and understanding of the system malfunctions. Accordingly, prior system information on the reasons, nature, and results of system limitations/malfunctions might prevent such trust decreases. In line with this, it was shown that transparent information during the use of automated driving systems affects trust (Ekman et al., 2016) and that information prior to using an automated driving system affects trust in the system during subsequent system use (e.g., Beggiato et al., 2013). Therefore, it was hypothesized:

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.

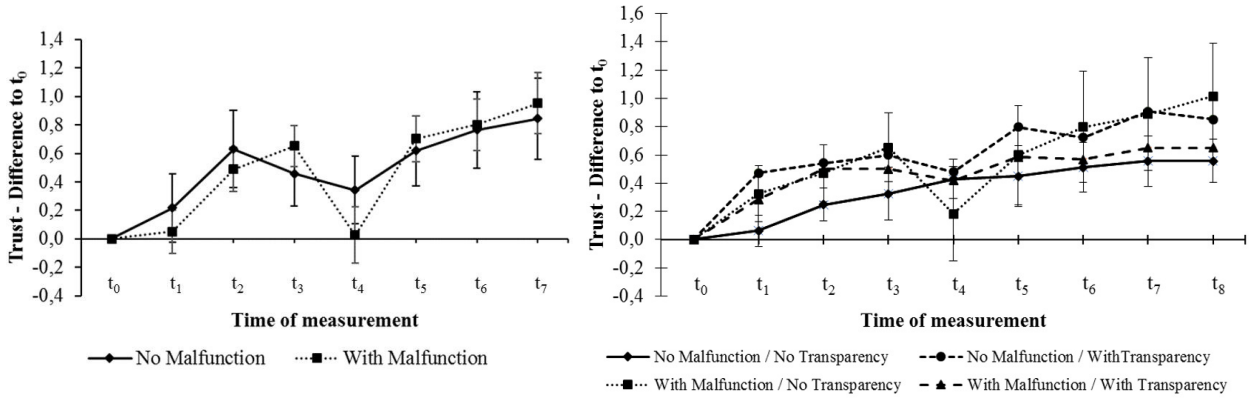
These hypotheses were tested in two driving simulator experiments in which trust was repeatedly measured during the first drive with an automated vehicle. In study 5, initial trust formation was investigated in two groups: error-free system and system with a malfunction. In study 6, additionally, the effect of prior system information about system limitations was investigated.

3.4.2 Study 5 - Method

In a one factorial design (between factor: system malfunction), trust was repeatedly measured over time during an initial drive with an automated driving system in the driving simulator of the department of Human Factors at Ulm University. Participants were randomly assigned to a malfunction (MF+) or no-malfunction group (MF-). Altogether N = 31 participants (18 females; M_{age} = 23.84, SD = 3.66) took part in the study. In Figure 10, the procedure of the study is illustrated. First, participants were introduced to the automated driving system and the experimental task. The

Figure 11

Trust development in study 5 (left side) and study 6 (right side) at the different times of measurement prior to and during the experimental drive for the two respectively four study groups. The malfunction took place before t_4. Differences to t_0 are displayed. (Figures from Kraus, Scholz, et al., 2019, ©2019 Human Factors and Ergonomics Society).



automation was introduced as an SAE level 3 system allowing for highly automated driving on the highway with automated overtaking. The system had two modes: a manual and an automated one. Participants were told that they could take over manual control at all times during the experimental drive. As a secondary task during the drive, participants were playing a tablet-based Tetris game (Electronic Arts, 2013). After the introduction and filling out questionnaires with the dependent variables (amongst others trust in automation), two practice trials for familiarization with both driving modes were conducted. In the practice trial, a system malfunction was implemented, which resulted in a safe mode established by an alarm sound and a visual signal in the user interface, activation of the indicator light, slowing down, and stopping at the emergency lane. After the practice trial, the experimental drive (see Figure 10) started during which, at t_1 – t_7, after automated overtaking and at m_1 and m_2 after two TORs, the simulation was halted and trust was measured. While in the MF- group the automation worked flawlessly during the complete drive, the MF+ group experienced a system malfunction and a safe mode of the system. After the drive, trust was assessed again, followed by a questionnaire on demographics. Trust was measured with the long version of the LETRAS-G scale at critical times (t_{pre}, t_0, t_4, and t_{post}) and at the remaining times a further shortened five-item version of the scale was used (see Appendix A). In the analysis, trust development in the interval between t_0 (after system introduction) and t_7 (before the effect of the second TOR) was inspected. The two study groups did not show significant differences in trust at the two times of measurement before the experimental drive, t_{pre}: $t(29) = 0.08$, $p = .937$, $d = -0.03$; t_0: $t(29) = -0.36$, $p = .720$, $d = 0.13$.

3.4.3 Study 5 - Results

Figure 11 (left side) shows the trust development over the course of the experimental drive relative to the trust measurement at t_0. To test the hypothesis on a trust increase over time (H1), a repeated measures ANOVA with the trust measurements from t_0 - t_7 as the dependent variables was conducted in the MF- group. To account for a violation of the assumption of sphericity, Huynh-Feldt corrected degrees of freedom were reported. In line with H1, an F-test showed a significant effect of the time of measurement on trust, $F(4.27,59.77) = 3.35$, $p < .05$, $\eta_p^2 = .19$, and a polynomial contrasts analysis supported a positive linear trend for trust over time in the MF- group, $F(1,14) = 6.01$, $p < .05$, $\eta_p^2 = .30$. Furthermore, in support of a temporary trust decrease after the experience

of a TOR (H2), a polynomial contrast analysis with the trust measures at t_2, m_1, and t_3 in the complete sample supported a quadratic but not a linear trend for TOR 1, linear: $F(1,29) < 0.01$, $p = .966$, $\eta p^2 = .00$, quadratic: $F(1,29) = 5.60$, $p < .05$, $\eta_p^2 = .16$. Unlike assumed in H2, for TOR 2 in both groups, a significant negative linear trend, but not a quadratic trend was supported, indicating a more stable trust decrease, MF-, linear: $F(1,14) = 10.05$, $p < .01$, $\eta p^2 = .42$, MF+, linear: $F(1,15) = 9.71$, $p < .01$, $\eta_p^2 = .39$. Further, to investigate the effect of system malfunctions on trust formation (H3.1 and H3.2), separate contrast analyses for the two study groups were conducted with trust measured before (t_3), directly after the malfunction (t_4), and after some further driving (t_5). While for the MF+ group a quadratic, but not a linear trend was supported, linear: $F(1,15) = 0.35$, $p = .564$, $\eta p^2 = .02$, quadratic: $F(1,15) = 15.32$, $p < .01$, $\eta p^2 = .51$, both trends were not significant in the MF- group, linear: $F(1,14) = 1.71$, $p = .212$, $\eta p^2 = .11$, quadratic: $F(1,14) = 1.91$, $p = .189$, $\eta p^2 = .12$. This pattern of findings supported a temporary trust decrease after a system malfunction (H3.1) and a subsequent trust recovery (H3.2). In an additional two-tailed t-test at t_7, no group differences for trust were found, $t(29) = -0.77$, $p = .449$, $d = 0.28$, indicating that the trust impairment in the MF+ group after the malfunction was not permanent and trust was reestablished. Taken together, all investigated study hypotheses were supported (H1-3) in study 5.

3.4.4 Study 6 - Method

The procedure of study 5 was slightly adapted in study 6 to investigate the effect of the prior transparency information. First, all TORs were removed and participants remained in automated driving mode after the initial take-over. Second, an attention assistant (a beep in a 1-minute interval after which participants had to direct their view to the road and touch the steering wheel) was implemented to enhance balanced task engagement between the primary and secondary task. Third, instead of Tetris, the participants played Angry Birds (Rovio Entertainment Corporation, 2010) as this game allows for task disruption. Fourth, additionally, to the experience of a malfunction, transparency was manipulated in two levels. In the low transparency groups (LT), the malfunction and safe mode were excluded from the practice trial. In the high transparency groups (HT), the malfunction and safe mode were part of the practical trial and participants received additional information on the technical reasons for malfunctions as well as on the character of the safe mode. Fifth, an additional overall evaluation (besides others of trust) of the system after the experimental drive ended was included in the study design ($t_{overall}$). This resulted in a 2 (malfunction) x 2 (system transparency) experimental design with four groups, MF-/LT, MF+/LT, MF-/HT, and MF-/LT. The final sample consisted of $N = 47$ participants (27 females, $M_{age} = 27.45$, $SD = 9.75$). As a control for the transparency manipulation, a three-item scale on the anticipation of a malfunction was filled out after the experimental drive in the MF+ groups. A t-test supported a higher transparency for the MF+/HT, $M = 3.82$, $SD = 1.12$, than the MF+/LT group, $M = 2.39$, $SD = 1.03$, $t(20) = 3.10$, $p < .01$, $d = 1.32$.

3.4.6 Study 6 - Results

The development of trust over time during the experimental drive is depicted in Figure 11 on the right side. Again, as for violations of sphericity, Huynh-Feldt-corrected degrees of freedom were reported. No group differences in trust at the two initial times of measurement were found, t_{pre}, $F(3,43) = 1.80$, $p = .162$; t_0, $F(3,43) = 1.15$, $p = .339$, indicating that the additional information provided in the HT groups did not affect initial trust levels. All findings of study 5 in regard to H1 and H3 were validated. In support of a trust increase over time in flawless system functioning (H1), in the MF-/LT group, in a repeated measures ANOVA (t_0 - t_8) a significant effect for time

of measurement on trust was found, F (2.46,29.51) = 3.22, $p < .05$, $\eta^2_p = .21$ and a t-test between the trust measures at t_0 and t_8 revealed a significant difference, $t(12) = -3.64$, $p < .01$, $d = 0.79$. In the MF-/LT group, this trust development corresponded with a positive linear trend, $F(1,12) = 6.31$, $p < .05$, $\eta^2_p = .35$. Furthermore, the temporal trust decrease after a system malfunction (H3.1) and subsequent recovering of trust (H3.2) was supported by an according pattern of trends in the two LT groups of the trust measurements at t_3, t_4, and t_5. In line with H3.1, while for the MF-/LT group no trend was significant, linear: $F(1,12) = 0.35$, $p = .568$, $\eta_p^2 = .03$; quadratic: $F(1,12) = 0.15$, $p = .709$, $\eta_p^2 = .01$, for the MF+/LT group, the quadratic, but not the linear trend was significant, linear: $F(1,10) = 0.09$, $p = .772$, $\eta_p^2 = .01$; quadratic: $F(1,10) = 5.65$, $p < .05$, $\eta_p^2 = .36$. In line with the hypothesized trust recovery over error-free system use subsequent to a malfunction (H3.2), at t_8 no significant difference for trust between the two LT groups was found, $t(22) = 0.78$, $p = .446$, $d = 0.32$. In order to test the hypothesis that information on reasons and consequences of malfunctions prevents a trust decrease after a malfunction (H4), polynomial contrasts were calculated for trust at t_3 to t_5 in the two HT groups. In both groups, in line with the hypotheses, no significant trend was statistically supported, MF-/HT, linear: $F(1,10) = 1.83$, $p = .206$, $\eta p^2 = .16$, quadratic: $F(1,10) = 2.95$, $p = .117$, $\eta_p^2 = .23$; MF+/HT, linear: $F(1,11) = 1.36$, $p = .269$, $\eta_p^2 = .11$; quadratic: $F(1,11) = 1.87$, $p = .199$, $\eta_p^2 = .15$. In combination with the significant quadratic trend in the MF+/LT group, this indicates that the presentation of transparency information prevented a trust decrease after the system malfunction in the MF+/HT group. Interestingly, the presentation of the transparency information was not associated with a significant trust difference of the two HT, $M_{HT} = 5.19$, $SD = 1.01$, and the two LT groups, $M_{LT} = 5.33$, $SD = 0.98$, at t_0 as indicated by a non-significant t-test, $t(45) = 0.485$, $p = .630$, $d = -0.14$.

3.4.7 Studies 5 and 6 - Conclusion

The findings of studies 5 and 6 support the notion of a dynamic nature of trust calibration during the early phase of interaction with an automated driving system. In both studies, trust increased linearly over time when the system functioned without error and decreased temporarily after a system malfunction. The combined findings support a rapid recovery of trust after the experience of a first malfunction in an automated driving system. Additionally, as supported by the findings of study 2, in the case of a system malfunction, the presentation of prior system information on the underlying reasons and the character of the safe mode prevented a temporary trust decrease. At the same time, the presentation of the prior system information on the malfunction did not lead to a trust decrease in the groups that received such information as compared to those that did not get such information. In study 1, a temporary trust decrease after the experience of a first TOR was found, while after a second TOR users' trust was more permanently reduced.

Taken together, the studies provide conclusive support for dynamic trust calibration during the early use of automated driving systems that entails several characteristics of the proposed processes of trust calibration from the presented framework. Thereby, the studies support the proposition that dynamic learned trust is calibrated in a continuous feedback process in which expected and observed system behavior are compared against each other and derivations are associated with adaptions of the level of trust (proposition 7.2). This is reflected in the successive increases of trust in error-free system functioning and the temporary reductions of trust if the expectation of flawless system function was violated either by a TOR or a system malfunction. In this regard, the study findings further support a link between trust stages 2 and 3 in the sense that prior system information is used to interpret the observed system behavior during the interaction to build up system

expectations, which in turn are used as a standard for comparison of the actual system behavior in trust calibration (proposition 8.2). Furthermore, for the domain of highly automated driving, the reported findings are among the first to show a reestablishment of trust in automation after the experience of a first system malfunction. In sum, the findings of studies 5 and 6 underline the importance of early system use for trust calibration and point into the direction that neither prior information about limitations or potential shortcomings of a system nor the actual experience of a system error is necessarily associated with trust decreases. Rather, it might be that such information and experiences support trust calibration before and during early system use by providing transparent insights into potential worst-case scenarios. Theoretical and practical implications of these findings are further discussed in chapters 4.2 and 4.3.

4 General Discussion

Trust has been found to be essentially related to behavioral decisions and reliance on automated systems (see chapter 1.1; e.g., Biros et al., 2004; Muir & Moray, 1996). There is a consensus in the field that trust in automation is built in a dynamic process over time (Lee & See, 2004) and that in this process trust is influenced by various user, system and situation variables (see chapter 1.4; e.g., Schaefer et al., 2016). While the two models by Lee and See (2004) and Hoff and Bashir (2015) contributed essentially to theorizing on trust in automation, the assumptions of these models have not yet been integrated and require further specification and empirical validation. Against this background, the first goal of this dissertation was to provide a theoretical integration of assumptions of these two models into a psychological framework for investigating trust in automation. Based on this integration and further theoretical extensions and specifications, the Three Stages of Trust framework was presented along ten basic propositions on trust processes prior and during the interaction with automated systems. Against this background, six studies were presented that build on the theoretical basis of the proposed framework. In these studies, trust processes prior and during early use of automated driving systems were investigated.

In the following, the findings of the single studies are discussed in relation to the proposed framework (chapter 4.1), and, on this basis, an integrative theoretical discussion and identification of future research questions along the three trust stages are provided (chapter 4.2). Furthermore, in chapter 4.3, practical implications for facilitating calibrated trust are discussed. In the last two sections, the strengths and limitations of the presented work are discussed (chapter 4.4) and a conclusion is drawn (chapter 4.4).

4.1 Summary of Findings and Evaluation of the Proposed Framework

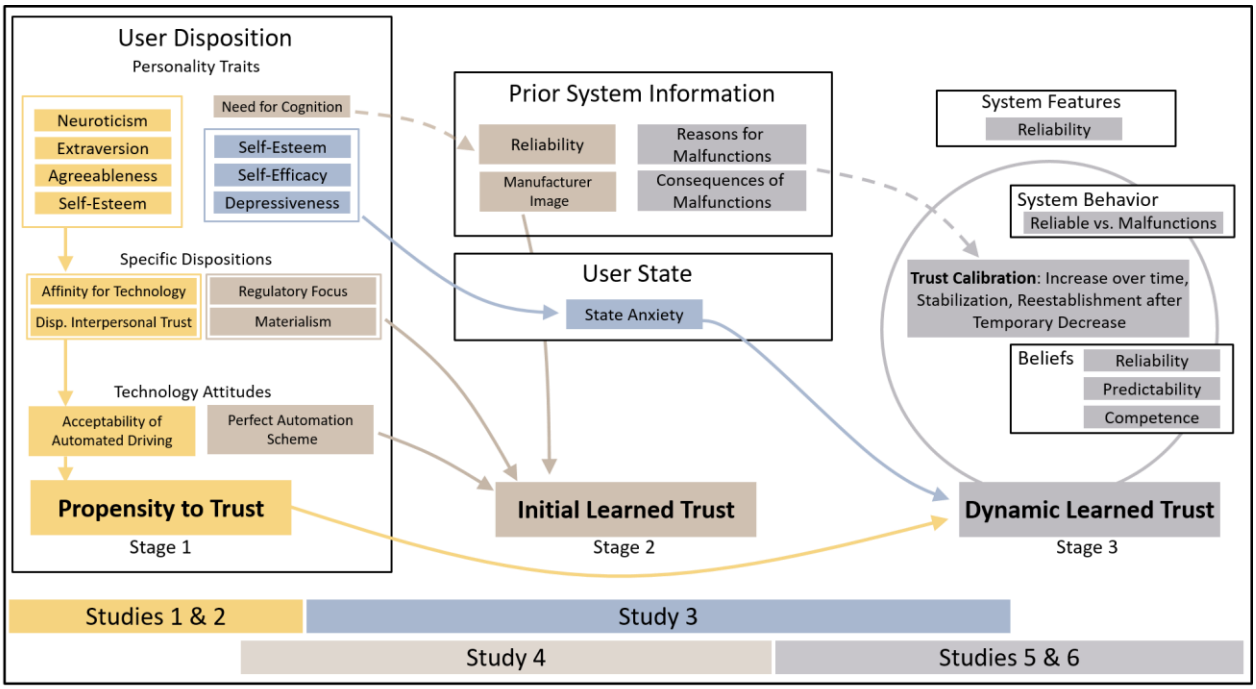
Taken together, the combined findings of all presented studies underpin the theoretical importance of an integrated consideration of trust processes at all three stages of the proposed framework for enhancing the understanding of how trust in automated systems is built up and calibrated. Figure 12 provides an integration and summary of the findings of the six studies along the Three Stages of Trust framework. Taken together, the presented studies provide initial support for nine of the ten underlying propositions of the Three Stages of Trust framework (Figure 12). The combined evidence of all six studies supports the three basic framework propositions of a dynamic and history-based character of trust development (proposition 1), the influence of person-, system- and

situation-related factors in the emergence of the different trust layers (proposition 2), and the differentiation of three subsequent and interrelated trust stages (propositions 3 and 8). Above this, the findings of the single studies provide initial support for some of the proposed trust processes at the three trust stages (propositions 4-7). In the following for each study, the findings are discussed in more detail in relation to the Three Stages of Trust framework.

Studies 1 and 2 investigated the role of user personality (besides others of dispositional interpersonal trust) and general technology-related attitudes for both the propensity to trust in automation and dynamic learned trust in automation (link between stages 1 and 3). The findings support the general proposition of the framework that different trust layers build on each other at different stages of trust formation (proposition 1 and 3). In support of such a link between stages 1 and 3, dynamic learned trust was found to be predicted by the propensity to trust. This finding is in line with previous findings from other technological domains (e.g., Merritt et al., 2013). Additionally, in support of proposition 4 of the framework, the propensity to trust in automation was explained by dispositional interpersonal trust, which in turn was predicted by extraversion, agreeableness, and self-esteem. These findings provide initial support for the popular hypothesis of a common psychological basis of trust in other people and trust in automated technology (e.g., Lee & Moray, 1992; Hoff & Bashir, 2015). Especially, the mediated relationships of extraversion and self-esteem through interpersonal trust and the propensity to trust in automation to dynamic learned trust underline the common personality basis of the different investigated trust variables. However, the magnitude of the proportion of explained variance also suggests that despite a certain commonality between the two types of dispositional trust, a substantial proportion of variance in the general tendency to trust automation cannot be related to the psychological mechanisms of interpersonal trust. This is particularly supported by the high variance proportions in dynamic learned trust, that were explained by general attitudes towards automated technology and automated driving. This prediction of dynamic learned trust by a pre-existing technology attitude supports the importance of considering the individual learning history and expertise in regard to (automated) technology in the explanation of individual differences in trust in automation. Furthermore, the negative relationship of neuroticism and the affinity for technology in both studies, points into the direction that individual differences in the openness and active approach of new technology might in part be grounded in differences in the dispositional tendency to be worrying and risk-averse in general. Furthermore, the role of self-esteem in the investigated personality model points into the direction that the self-assessment of one's own capability to use such a system in a functional manner might be an important prerequisite for trust in automation. This was further investigated in study 3. All in all, the established relationships of the more specific dispositions with theoretically well-established, broader traits help to gain a better understanding for the foundation of the psychological mechanisms involved in trust formation in more basal psychological functions like the user's self-evaluations, emotional stability, the tendency to trust other people and a general affinity for technology. In summary, the suggested link of the general personality of users to experiences they make in their individual learning history and the establishment of more specific technology-related personality traits and general technology-related attitudes in this process seems essential to be considered at the different stages of trust formation in research and practice.

Figure 12

Summary of findings of the six presented studies and their integration in the Three Stages of Trust framework. In the upper area, variables for which the studies supported a role in the formation of trust in automation are depicted along with their supported relationships. Solid lines represent direct relationships and mediations. Dashed lines represent moderations and interaction effects. In the lower part, study findings and their relation to the propositions of the framework are summarized for each study or pair of studies respectively.



Combined evidence: inital support for the basic propositions of the framework, that trust in an automated system is built up in a history-based process which begins before the interaction **(P1)**, is influenced by factors on the side of the person, the situation and the system **(P2)**, and is formed and calibrated along three sequential stages **(P3)**

Studies 1 & 2: Hierarchical Personality Model

Key Findings
- Traits/attitudes at different levels affect the formation of the propensity to trust and dynamic learned trust
- Role of: **neuroticism, extraversion, agreeableness, self-esteem, affinity for technology, acceptability of automation**
- **Interpersonal trust → propensity to trust in automation → dynamic learned trust in automation**

Framework
- Personality and global attitudes influence dyn. learned trust **(P8.1)** & the prop. to trust automation **(P4)**
- Trust stages build sequentially on each other **(P8.1)**

Study 3: Personality - Anxiety Mediation

Key Findings
- Dynamic learned trust was predicted by **the experience of anxiety** after receiving system information
  Personality differences in **depressiveness, self-esteem** and **self-efficacy** predict differences in anxiety
- **Mediation: Personality → state anxiety → dynamic learned trust**
- Unclear role of locus of control

Framework
- Personality and emotional states influence formation of dynamic learned trust **(P3)**
- Personality-related differences in emotional states influence the formation of learned trust **(P8.1)**

Study 4: Attitude Formation by Prior System Information

Key Findings
- Both **prior system information on reliability and manufacturer** influences initial learned trust
- In line with predictions of **dual-process theories**, trust changes based on reliability infomation were more stable over time and the use of information is **influenced by need for cognition**
- Regulatory focus, materialism and the **perfect automation scheme** are related to initial learned trust

Framework
- Initial learned trust is built up along prior system information **(P5)**; information use is moderated by personality **(P9)**; Formation of initial learned trust follows attitude formation processes **(P6)**

Studies 5 & 6: Dynamic Trust Calibration

Key Findings
- Dynamic learned trust increases over time with flawless system function and is **temporarily decreased after a system malfunction** and subsequently **reestablished** with further error-free system use
- **Transparent information** about reasons and consequences of system malfunctions **can prevent such a temporary decrease of trust in automation** after the experience of a system malfunction

Framework
- Dynamic learned trust is formed in a dynamic feedback process **(P7.2) along provided information (P5)**
- Information provided prior to system interaction influences the formation of dynamic learned trust **(P8.2)**

In **study 3,** a mediation of the effects of dispositional self-evaluations (depressiveness, self-esteem, self-efficacy, and locus of control) on dynamic learned trust by the experience of anxiety was investigated (link between stages 1, 2, and 3). This study was one of the first to focus on the relationship between the experience of anxiety and trust in automation. In this study, a negative relationship between state anxiety when being introduced to unfamiliar technology and dynamic learned trust in the system after some interaction with it provides support for a transfer of models proposing influences of emotional states on decision-making (e.g., Schwarz & Clore, 1988; Forgas, 1995) to the domain of trust in automation. Furthermore, the study findings supported the proposed mediations of the positive effects of self-esteem and self-efficacy on trust by state anxiety. Also, a negative indirect effect of depressiveness on trust was found. Persons, who viewed themselves as worthy and capable of performing well in general, experienced less anxiety when being introduced to an unfamiliar automated system and subsequently trusted such a system more. This underlines the essential role of one's assessment to use a new system in a functional manner and corroborates findings from early studies on trust in automation on an interrelation between trust in automation and self-confidence in their association to reliance behavior (see chapter 1.4.3; e.g., Lee & Moray, 1994; de Vries et al., 2003). At this point, it seems that a generalized positive self-evaluation of users reduces the level of anxiety when being introduced to an unfamiliar automated system. This reduced anxiety might be grounded in the self-assessment to be able to functionally react to all eventualities of the forthcoming interaction with the system. In this manner, users with higher self-esteem and self-efficacy might perceive themselves to be able to balance out potential shortcomings or errors of the automated system. This, in turn, might lead to higher trust levels as for the perceived own ability to use such a system more functionally, which makes the whole situation of using such a system seem less risky and dangerous. An integrative investigation of dispositional and situational self-evaluations, state anxiety, other personality variables, and reliance behavior in future research endeavors could provide insights into the character of the association of these variables. Like in studies 1 and 2, the effect of LOC in this study was considerably smaller than those of self-esteem and self-efficacy. In the same manner, LOC has been associated with mixed findings in other research areas before (e.g., driving behavior in general, Özkan et al., 2005). While the general direction of the relationships of LOC was comparable to those of the other two self-evaluative traits, these showed higher correlations and the mediation through anxiety to trust was supported. Accordingly, in terms of the self-evaluative personality basis of trust in automation, self-esteem and self-efficacy seem to assess these differences in users more clearly than LOC. Nevertheless, future research might further explore the relative role of different self-evaluations for the development of trust in automation. To conclude, the findings of study 3 support the role of personality (proposition 2) and state anxiety in trust formation (proposition 8.1). Specifically, in order to facilitate trust calibration, the role of anxiety during the introduction to unfamiliar automated systems and dispositional self-evaluations seemingly contributing to this negative emotional experience might be considered both in future research (chapter 4.2) and practice (chapter 4.3).

Findings of **study 4** support the role of prior system information in the establishment of initial learned trust at stage 2. In line with proposition 5 and 6 of the framework, the effects of different types of information were consistent with predictions derived from a dual-process theory of attitude change, the ELM. The pattern of effects indicates an influence of information processes through both the peripheral and central route on initial learned trust. This is in support of the notion that information with different levels of attributional abstraction (e.g., purpose and performance

information; Lee & Moray, 1992) is used in several processes (e.g., analytical and analogical; Lee & See, 2004) to establish trust in automation. Above this, in the study, an interaction effect indicating that user differences influence the use of information in trust formation was found. Participants with a high need for cognition changed their trust ratings more into the direction of the provided reliability information than those with a low need for cognition. This supports the notion that user characteristics moderate the use and processing of provided information in trust formation (proposition 9). In line with proposition 1 of the framework, in regard to the stable effects of the reliability information, it can be concluded that trust calibration begins before the actual system use. For these reasons, a consideration of initial trust formation before the first interaction with a given system in research and practice is similarly important for an understanding of trust processes as dynamic trust calibration during the interaction. The supported transfer of processes of attitude formation provides the possibility to build on existing research from social psychology in the further investigation of information processing in the formation of initial learned trust. As a preliminary conclusion of such a transfer, in the design of prior system information, it seems essential to consider the complex interaction between a) message content, b) the way of message presentation, c) the situation in which the message is presented, as well as d) general dispositions of the user, and e) the situational state in which the user processes the information. Taken together, from the findings of study 4, various research questions and implications for the design of the introduction process of novel automated systems can be derived (as discussed in more detail in chapter 4.3.2).

Studies 5 and 6 investigated the dynamic development of trust in automation during the early interaction with an automated driving system (stage 3). Both studies conclusively found trust in automation to increase over time in an error-free system supporting results of previous studies (e.g., Hergeth et al., 2016; Beggiato et al., 2013; Lee & Moray, 1992). While in study 5 an early TOR was followed by a temporal decrease of trust, confirming findings by Hergeth et al. (2015), after a second TOR, such a decrease was not observed, supporting a more permanent trust decrease by the second TOR. Additionally, in both studies, trust in automation decreased after a first system malfunction and recovered quickly in subsequent error-free system use. Furthermore, in study 6 it was found that this temporal trust decrease after a malfunction could be prevented by prior system information providing more transparency on the character of malfunctions. Taken together, these findings support the notion that dynamic learned trust is dynamically calibrated in a feedback loop, in which system expectation and actual system behavior are compared and discrepancies are reflected in trust adjustments (propositions 7.2). As long as the system performs well, it tends to surpass the initial expectations leading to a trust increase over time. An unexpected error of the system, on the other hand, results in a trust decrease. Such a decrease was not found after an error, that was anticipated on the basis of the provided information before the interaction with the system. This supports the role of prior system information in the calibration of dynamic learned trust which is in line with earlier findings suggesting an influence of prior information about system limitations on trust calibration (e.g., Beggiato et al., 2013; Körber et al., 2018). The use of prior system information for trust calibration during system interaction, thereby underlines the sequential interrelation of trust processes at stages 2 and 3 (propositions 3 and 8). Interestingly, the study groups that received the information on system malfunctions before system interaction did neither trust the system less before nor after the first interaction with it as compared to those groups that not receive such information. This emphasizes that trust is not necessarily negatively affected by the transparent communication of potential shortcomings or even actual system malfunction. Rather, it points

into the direction that not necessarily the "new" negative information (that the system is not flawless) leads to trust reduction, but that actually the divergence between expected and actual system behavior is the reason for trust reduction after a system malfunction. Taken together, this study supports the assumption that in trust calibration information prior to the system use and observed system behavior during the interaction are integrated into a system expectation that, in turn, influences trust formation in a feedback loop in which this expectation is compared and updated along the observed system behavior. As for the complex dynamics in trust calibration during the establishment of initial and dynamic learned trust, provided information both before and during early system interaction should be taken into consideration in the design and dissemination of automated systems (as discussed in chapter 4.3).

4.2 Theoretical Integration and Directions for Future Research

In the following, the combined findings of the studies and perspectives for future research are discussed in relation to the framework. This discussion is organized in six sections: a first section that covers the underlying hypothesis of the dynamic and history-based nature of trust in automation, one for each of the proposed trust stages, a section that covers a discussion of the influences of individual, system-specific and situational variations, and a final section, in which implications for the design of research on trust in automation are discussed.

4.2.1 Trust as a Dynamic and History-based Variable

In line with theories on trust formation in the interpersonal (e.g., Rempel et al., 1985) and technological domain (e.g., Lee & Moray, 1992), the proposition that trust in automation is dynamically built up in three sequential stages was supported in the presented studies. First, the positive and substantial correlations of the propensity to trust and dynamic learned trust in studies 1 and 2 indicate that previous experiences with other automated systems considerably influence the emergence of trust in a newly introduced system. Second, in studies 3 and 6, the reported effects of available information and individual reactions at stage 2 to the formation of dynamic learned trust at stage 3 supports the interrelation between the two layers of learned trust. In study 3, anxiety after an introduction to an automated driving system was found to be a predictor for the individual level of dynamic learned trust after the interaction with an automated system. In study 6, it was shown that the information received at stage 2 affected the interpretation of system behavior at stage 3. Thus, it can be concluded that the information provided prior to system interaction and the individual level of initial trust affects the interpretation of system behavior at stage 3. The supported interrelations of trust processes at the different stages of the proposed framework allow the derivation of a multitude of research questions for future studies to gain a better understanding of trust in automation.

At this point, the process of how the preceding trust stages feed into the later ones is still underspecified and provides many open questions for future research. Essentially, the relative importance of the propensity to trust for the establishment of the subsequent trust stages needs further investigation. In this regard, the development of its relative importance for the formation of learned trust over time seems an interesting research question. It might be hypothesized that pre-existing tendencies to trust an automated system more strongly affect trust formation in a new system if only little or ambiguous information is available and that its influence diminishes over time with the accumulation of specific information about a system. Further, based on the findings of study 6, a further investigation of the processes through which the provided information and trust cues

at stage 2 influence trust calibration at stage 3 seems worthwhile. Further research questions in regard to the interrelations of the proposed trust stages are discussed in the following sections.

4.2.2 Propensity to Trust

The propensity to trust, as a general tendency to trust in automated systems, is proposed to be formed from a combination of broader personality variables, an individual's learning history with technology in general and automated systems in specific as well as general technology-related attitudes. It is further assumed that a higher individual expression on this trait leads to more positive expectations towards automated systems. In this regard, the propensity to trust constitutes an individual disposition with which users enter the process of learning to trust a system. This role of the propensity to trust was essentially supported by studies 1 and 2, in which it substantially predicted dynamic learned trust in line with previous studies (e.g., Merritt, Unnerstall, et al., 2015). Furthermore, a relationship between dispositional interpersonal trust and the propensity to trust supported that this generalized trusting attitude towards automated systems is, to some extent, rooted in the individual disposition to trust other persons. Accordingly, the psychological mechanisms associated with trust in other persons and trust in automated technology seem to share some similarities (as discussed in chapter 4.1). On the other hand, as this relationship was only in a range of a small effect, the propensity to trust in automation seems to have an additional unique psychological basis. In this regard, a more history-based interpretation of the relationship of the propensity to trust in automation and the affinity for technology might point into the direction of potential underlying mechanisms providing some explanatory approaches for the reported differences between dispositional trust in others human and in automated technology. As reflected in both, the scales and definitions of the utilized scale for affinity for technology (Karrer et al., 2009), people with higher affinity for technology like to possess and use new technological products more and find themselves as more capable of understanding and using these products successfully than those with lower affinity for technology. In line with the proposed framework, this points into the direction of the role of the individual learning history with technology for the emergence of the propensity to trust. As compared to people with lower accentuation in this trait, technology-affine people have a history of actively approaching technology leading to more experience with technology and more knowledge of the general abilities and functionality of technology. This, in turn, might lead to a more positive self-assessment of their ability to be able to use technological systems in a functional and effective manner and thus a tendency to start into the interaction with a considerably higher readiness to place their fate in the hands of an automated system. In line with this interpretation, the factor of competence included in the applied measure for affinity for technology reflects the self-assessment of their own ability to make good use of new technology. While such a foundation of differences in the propensity to trust in the individual learning history with technology awaits further investigation, initial support for the role of self-evaluations for the formation of trust in automation stems from study 3 in which a positive relationship of the traits self-esteem and self-efficacy and dynamic learned trust in automation was found. Furthermore, in studies 1 and 2, a consistent significant negative correlation between neuroticism and the propensity to trust in automation was found. This relationship might be – as the studies' findings suggest – mediated at least in part by the individual affinity for technology. In this regard, the individual disposition of neurotic persons to avoid risks might lead to a generalized avoidance of situations in which trust has to be put in an unfamiliar technology, as reflected in lower levels of the propensity to trust in automation.

On the basis of these findings, future studies could further investigate which broader personality tendencies build the psychological basis for differences in the propensity to trust automation. Besides this, building on the findings of studies 1 and 2, the interrelation of a general trusting personality (as reflected in dispositional interpersonal trust) and the self-attribution of technological expertise for the emergence of the propensity to trust in automation could be investigated more closely. In this regard, especially the character of psychological processes that explain the relation between self-evaluative dispositions, the experience of anxiety in face of new technology, affinity for technology and trust in automated systems might be further investigated, also with a consideration of the role of the propensity to trust automation in this process. An additional promising direction for future studies is the investigation of how the individual learning history with technology and automated systems influences the establishment of the propensity to trust in automation.

4.2.3 Initial Learned Trust

Both layers of learned trust are built up from the aggregate of the information perceived and processed up to the time of their measurement. Relevant information for initial learned trust is proposed to be established by trust cues included in prior system information from sources like advertisement or information campaigns. Trust cues were proposed to be classified in the five categories reputation, purpose, process, performance, and appearance (e.g., Lee & Moray, 1992, Thielmann & Hilbig, 2015, see chapter 2). For the formation of initial learned trust, the presented Three Stages of Trust framework proposes that the use and processing of information at this stage follows the general principles of attitude formation (e.g., Maio et al., 2018), including analytical, analogical, and affective processes (Lee & See, 2004). As initial learned trust is built up on the basis of the available, partly ambiguous information about a system before its first use, the evaluation of available information should be considerably influenced by personality, pre-existing general attitudes, personal preferences, and motivations. In study 4, first evidence for both an analogy to attitude formation and the role of user personality in the use of initial system information in the formation of initial learned trust has been provided. On the theoretical basis of the ELM, it was shown that prior information was differently processed by users with a relatively high and those with a relatively low NFC. Participants with a high NFC weighted information on the reliability of an automated system more in their trust assessment than those with a low NFC. Furthermore, it was shown that in line with the predictions of the ELM, the reliability information, which is supposedly processed through the analytical route, had a more stable effect on trust over time as compared to information on the brand of the system. Also, the considerable correlations between the high expectations factor of the PAS and initial learned trust, in line with earlier findings by Lyons and Guznov (2019), as well as the respective correlation for materialism supported the role of general prior attitudes for the formation of initial learned trust. Above this, the findings of study 3 imply that also affective processes influence initial learned trust. People who experienced higher levels of anxiety when being introduced to a new automated system, trusted the system less in a later interaction. As the variations in state anxiety were in part explained by dispositional differences of users, this study provides additional support for the notion that also the processes at stage 2 are considerably influenced by personality. Taken together, research on the emergence of initial learned trust might combine the perspective of attitude formation with theories that view emotions as an important source for judgements like the affect-as-information model (e.g., Schwarz & Clore, 1988) or the affect infusion model (e.g., Forgas, 1995). Furthermore, top-down processes resulting from user personality and general attitudes might be considered.

While recently, several studies on the effect of prior system information on trust have been published (e.g., Körber et al., 2018; Beggiato et al., 2013), future research endeavors might take the attitude formation processes in the formation of initial trust more into account. In this regard, a first interesting research question is which different kinds of information are essential for the formation of initial learned trust and how these interact in their effect. In this regard, the suitability of the proposed classification of trust cues from chapter 2 could be further explored. Second, the role of the information source might be considered in future research projects. For example, it is a promising research question to investigate if the information provided by system manufacturers is used differently from information provided by more neutral sources. Third, it might be interesting to investigate how users deal with conflicting information from different sources or at different levels of processing. Fourth, the role of the order of provided information, their relation to previous experience with other systems, and a moderation of information use by individual differences might be further explored. Fifth, based on the reported mediation of the effects of self-evaluative traits to trust by the level of anxiety at the time of being introduced to a new automated system, also the exploration of the role of these and additional trait and state variables in the process of building up initial learned trust from provided information seems worthwhile. Sixth, characteristics of the provided information like salience or simplicity should essentially influence the probability that it is used for building up trust and, therefore, might be addressed in the future. Seventh, the role of mental models and beliefs for building up an initial system expectation and initial learned trust might be further investigated. Eighth, the relative importance of bottom-up (the use of provided information) and top-down processes (influence of global technology attitudes and the propensity to trust) in the formation of initial learned trust is worthwhile to be further investigated. Ninth, in the same manner, the role of situational factors for the transfer of pre-existing attitudes to new systems might be more closely investigated. For example, it might be that such top-down processes are of higher relevance if a user has restricted capabilities to process new information. Reasons for such restrictions involve a lack of motivation, high situational or task-related workload, or a temporary salience of the existing attitude. In this regard, also potential individual moderators of the relationship between attitudes and trust formation like the accessibility of attitudes (e.g., Fazio, 1990) or self-monitoring (e.g., Ajzen, 2005) might be worthwhile to consider.

4.2.4 Dynamic Learned Trust

The presented findings of studies 5 and 6 support the general notion of the proposed framework that dynamic learned trust is calibrated during the interaction with an automated system in a psychological process with characteristics of a feedback loop, in which observed system behavior is compared to system expectations. The framework further proposes that initial learned trust provides a starting point for trust calibration in the first interaction with a system. While this specific assumption has not been investigated in the presented studies, initial support for an interaction of the processes at stages 2 and 3 can be derived from the findings of study 6, which support a combined consideration and integration of information provided at these stages in building up dynamic learned trust. Additionally, the findings of study 6 underline that prior system information on potential shortcomings or flaws of an automated system does not necessarily negatively affect trust in the system. Rather, as such prior system information might serve as a buffer against a trust reduction in the face of a system malfunction, such allegedly negative information might, thus, be even associated with a long-term trust-enhancing effect. The hypothesized feedback process of trust calibration is proposed to continue after the first reliance decision by comparing the observed

system behavior to the initial trust expectation, which is stabilized or adjusted and, as a consequence, so is the level of dynamic learned trust. This adjusted level of trust is proposed to subsequently build the basis for the forthcoming reliance decision. In the resulting feedback process, trust is proposed to be steadily updated during the interaction with the observed system behavior and output of the user interface. In this dynamic feedback process, new information is proposed to be used to calibrate trust leading to either stabilization of this assessment in the case of affirmative new information or an adjustment in the case of conflicting new information. Empirical evidence for such a feedback process has been reported in studies 5 and 6 (see chapter 4.1), in which dynamic learned trust was found to be calibrated dynamically according to the observed system behavior as reflected in a linear trust increase over time in flawless system interaction and temporary trust declines after the experience of a first TOR or system malfunction. To conclude, it seems recommendable for future research to incorporate design measures to reflect both the dynamic nature of learned trust during system interaction and the enhanced complexity of the combined consideration of information provided prior and during system interaction in trust calibration to account for a complete picture of the comparison processes at play during the interaction. Several recent studies in the domain of automated driving or advanced driver assistance systems followed such a paradigm in controlling the provided information prior to early system use and measuring trust repeatedly in the course of the study (e.g., Beggiato, et al., 2013; Koustanaï et al., 2012; Hergeth et al., 2016; Körber et al., 2018).

On this basis, a number of research questions to further enhance the understanding of the nature of the calibration of dynamic learned trust in automation can be derived from the proposed framework. First, the development of trust over time as a function of a) system behavior, b) expectations, and c) the interaction of expectations and system behavior might be further addressed (e.g., consequences of exceeding or disappointing expectations). Furthermore, the role of interface output for the interpretation of system behavior seems worthwhile to be further explored. For example, it seems reasonable to hypothesize that system behavior deviating from user expectations has a less substantial effect on trust in the system if appropriate explanations for such a deviation are given in a user interface. Above this, it is of interest to further investigate how user differences like prior experience, attitudes, and personality traits influence the dynamics of trust calibration. In this regard, the investigation of the role of different states in which users encounter a system (e.g., anxiety) for dynamic trust development provides a potential for future research. Additionally, a further specification and empirical investigation of the proposed trust mediation cascade (mental models, beliefs, intention, reliance behavior) in orientation to Ajzen (1991) might help to gain essential insights in how trust in automation is related to reliance behavior and thereby investigate proposition 10 of the presented framework, which has not been covered by the presented studies. In this regard, an integration of the variables specified in the Technology Acceptance Model (TAM, Davis et al., 1989) and the Unified Theory of Acceptance and Use of Technology (UTAUT, Venkatesh, et al., 2003) seems promising. In line with this, Buckley et al. (2018) found trust to explain additional variance in the intention to use an automated vehicle above other variables entailed in the PLABE and the TAM. Furthermore, the role of the mental model for dynamic learned trust should be further investigated. These studies might build on the work by Beggiato et al. (2013, 2015), who already reported on an essential role of mental models as a basis for trust calibration.

4.2.5 Influences of Person Characteristics, System Features, and Characteristics of the Situation
The Three Stages of Trust framework proposes that trust processes are decisively influenced by characteristics of the user, situational conditions, and system characteristics (propositions 2, 8, 9, and 10). The presented studies investigated variables of all three groups.

First, in terms of dispositional characteristics, studies 1 to 4 found considerable relationships of user disposition with the propensity to trust, initial, and dynamic learned trust. In particular, the studies support a prediction of differences in trust in automation by general and technology-specific personality traits as well as general attitudes, besides others towards technology and automation. The combined evidence of these studies underlines in specific the role of dispositional self-evaluative processes and global technology attitudes for differences in trust in automation. Above this, in study 2, a partial overlap between trust in other people and automation trust was found. The findings of study 4, additionally, point into the direction that the individual representation of gains and losses might influence trust in automation (regulatory focus, e.g., Higgins, 1997). The relevance of considering direct and indirect effects of personality and technology-related attitudes in future research has been discussed in the respective sections covering the trust stages above.

Second, in the presented studies, the effect of several communicated and experienced system characteristics on trust in automation was investigated both at stages 2 and 3. In specific in studies 4, 5, and 6 system reliability and the manufacturer brand were manipulated and considerable effects were found. In terms of directions for future research on the effects of system features and characteristics on the emergence of trust, the different attributional levels of information (Lee & Moray, 1992) and the different processes in which available information is processed (e.g., Lee & See, 2004) need further consideration (as detailed in the section on initial learned trust). Furthermore, the role of the provided information in the user interface for the interpretation of system behavior has only been scarcely investigated. In this regard, especially the interaction of information provided at different times in the trust formation process (e.g., at stage 2 vs. stage 3), through different channels before the interaction (e.g., public campaigns vs. advertisement) and during the interaction (actual system behavior vs. information in the user interface) could be further investigated. Essentially, also the interaction of conflicting information at the trust stages 2 and 3 (e.g., advertisement vs. actual system behavior) in regard to system characteristics seems worthwhile to be addressed in future research. In this regard, an orientation on the psychological mechanisms proposed in the presented framework (attitude formation and feedback process with actual/target comparison) might support an identification of meaningful research questions.

Third, in terms of situational factors (on the side of the users or the environment), the presented framework proposes that situational factors essentially influence the formation and calibration of trust in four processes: a) in the perception and interpretation of available information prior to system interaction, proposition 9, b) in the perception and interpretation of system behavior and associated output of the system's user interface, proposition 9, c) in the translation of behavioral intentions to actual reliance behavior, proposition 10, and d) directly through the attribution of emotional states on the system or the interaction situation (e.g., affect-as-information), proposition 8.1. In line with proposition 8.1, in study 3, the emotional state of the users when being first introduced to an automated driving system was found to be directly related to dynamic learned trust. While at this time, mainly studies on the role of positive and negative affect are available (Merritt, 2011, Stokes et al. 2010, Frison et al., 2019), in the presented study, a strong relationship between state anxiety and dynamic learned trust was found. This finding supports the fourth route, through

which situational variations in users' emotional states might affect trust in automation as proposed by the affect-as-information model (e.g., Schwarz & Clore, 1988) and the affect infusion model (e.g., Forgas, 1995). On the other hand, as anxiety was found to mediate the relationships between personality variables and dynamic learned trust, the emotional experience in the investigated situation was, at the same time, strongly influenced by user dispositions. Future research might address the role of anxiety for trust formation more closely and thereby aim at the differentiation of situational and dispositional variances in the experience of emotional states during the familiarization with automated systems and their relative effects on trust in automation. The same holds true for a more detailed investigation of the role of other emotional states. Besides situational fluctuations in user emotion, future research might also address the influences of situational variations of other psychological states like motivation, cognitive capacity, or situational awareness on trust formation. In the same manner, like momentary psychological states influence trust formation, more objective situational variables and task characteristics like the constellation of tasks, time pressure, and physical conditions like lighting and surrounding noise could also affect trust formation, especially in terms of information processing, and might be considered in future research endeavors.

4.2.6 Recommendations for the Design of Research on Trust in Automation

First, based on the sequential process of trust formation over time, measuring learned trust repeatedly during the course of a study seems essential to understand the individual development of trust. Therefore, optimally, trust should be measured before, repeatedly during, and after the interaction with a system under consideration. Specifically, measuring trust before the actual interaction provides possibilities to account for pre-existing trust differences in participants. In this sense, it seems recommendable to use increments of pre and post measurements of trust to be able to assess the actual effects of manipulation in terms of system characteristics or experiences with the system on trust. Above this, a repeated measurement of trust allows for the assessment of individual relative trust variations (gains and losses) over time and in reaction to specific pieces of information (stage 2) and a specific system behavior (stage 3).

Second, for the assessment of (initial and dynamic) learned trust, it is essential to use a subjective trust measurement that clearly only measures trust as a subjective attitude (Lee and See, 2004) and is not confounded with the integration of antecedents or consequences of trust. In this regard, the available measurement approaches vary considerably in their underlying trust definitions and number of included trust dimensions (Chien et al., 2014; Madsen & Gregor, 2000; Muir & Moray, 1996; see Jian et al., 2000 for a detailed discussion). In line with Jian et al. (2000), it seems advisable to conceptualize trust as a unidimensional construct. Besides this, eye-tracking (Hergeth et al., 2016) and skin-conductance (Mühl et al., 2019; Walker et al., 2019), which have been recently found to reflect trust variations, might provide reasonable additions to subjective trust measures. In derivation from the definition of trust as an attitude, it seems further worthwhile to clearly distinguish trust in automation from its objective correlates or consequences. While trust is conceptualized as a subjective psychological variable, reliance behavior is an objective behavior, which is assumed to be influenced by trust (see chapter 1.1). Furthermore, a strong, direct association between subjective trust measures and objective behavioral outcomes cannot always be expected as this relationship might be moderated, e.g., by situational variables or task characteristics. From this, a clear differentiation and conceptual separation of subjective trust measures and measures of behavioral consequences of trust can be concluded. This is in line with Wiegmann et al. (2001), who claim "that automation trust as a psychological construct might appropriately be inferred only

through subjective measures, such as reliability estimates. On the other hand, objective measures of automation usage, such as agreements with an aid, actually reflect automation reliance […]" (p. 364). Nonetheless, objective measures provide a reasonable supplement in most trust research and a combined measurement of subjective trust measures and behavioral measures seems meaningful for investigating both the relevance of trust for reliance behavior and the nature of this relationship in the studied scenarios. Furthermore, from the recommendation to measure trust several times over the course of a single study, it can be followed that such a scale should be economical. Based on these considerations, in Appendix A, the LETRAS-G is introduced, which was used in five of the six presented studies.

Third, the different nature of the propensity to trust (stage 1) and the learned trust layers (stages 2 and 3) needs to be accounted for with the implementation of a specific measure for the propensity to trust in automation in the study setup before the interaction with a specific automated system. Such a measure needs to provide the possibility to capture stable inter-individual differences in the general tendency to trust various automated systems over different situations. In studies 1 and 2, to measure the propensity to trust in automation, four items of the Propensity To Trust Scale (Merritt, 2011; Merritt et al., 2013) were rephrased for "automation technology" instead of their original reference to "machines". This scale showed convincing reliabilities of Cronbach's α of about α = .91 in both studies and was found, as reported, to predict dynamic learned trust well.

Besides these considerations for the measurement of trust, additional recommendations for the research design of studies investigating the formation and calibration of trust in automation can be derived. As for the presumed moderations of trust processes by variables on the side of the user (e.g., traits, motivation), the situation, and task environment, on the basis of theoretical considerations, variables possibly influencing trust formation should be controlled for in research designs. Furthermore, in order to make the experience of the system under consideration comparable for different people, it seems recommendable to keep the information participants are addressed with before the beginning of an experimental interaction as constant as possible. For this purpose, also, a comparable background of experience of participants can be enhanced by introducing a respective criterion in sample selection or by the presentation of information before the experimental interaction to facilitate alignment of the level of experience of different participants. This also means that in many cases, the interaction situation with the system should be designed as simple as possible to facilitate experimental control and the interpretation of the findings. Another critical factor is the creation of a maximum degree of psychological reality in the interaction. Here, the existence of a certain degree of uncertainty in the system interaction and a certain degree of subjective risk resulting from a failed cooperation with the system is particularly important for investigating trust in automation. Both components are essential prerequisites for valid conclusions about the role of trust in automation for actual reliance behavior in practical settings. In this regard, to be able to understand and interpret study findings, it seems reasonable to conduct exit interviews in the process of debriefing participants.

4.3 Practical Implications

In the following, based on the Three Stages of Trust framework and the presented study findings, recommendations for the design, deployment, and introduction of automated systems are derived. These are general for automated technology if not explicitly framed for automated driving.

4.3.1 Designing for Calibrated Trust, Diverse Users, and Situational Variability

In all endeavors to educate and assist users in learning how to use an automated system in a functional manner, the promotion of calibrated trust provides an essential goal. Although a certain level of trust is necessary to motivate users to get involved with an automated system at all, an unconditional maximization of trust is irresponsible and dangerous. The proposed trust processes and the empirically supported relationships provide directions to prevent overtrust and distrust as well as to facilitate trust calibration. In line with the presented framework and study findings, calibrated trust can be promoted by providing information on a system's capabilities and limitations prior to (e.g., user manuals, tutorials) and during system use (e.g., real-time information on sensor perception of surrounding, sensor status, the state of the automation, and warnings). In this regard, as the system is more or less a black box for users at the beginning of the familiarization process, all information provided prior and during system interaction helps to build up a mental model, beliefs, a system expectation, and calibrated trust. A further significant contribution of this research is the identification of intra- and inter-individual differences in users that influence trust formation and calibration. Additional to these individual differences, also situational variables have been found to influence trust formation (see chapter 1.4.3). For the practice of developing and disseminating automated systems, these complexities can be addressed by designing both prior information and the actual system (its functionality and user interface) with consideration of individual and situational variances. Approaches for both are discussed in the following two sections. Taken together, to account for the complexities of designing automated systems, in line with Lee and See (2004), it can be argued that it is the task of developers to design automated systems in a way that they evoke a calibrated level of trust in the users. In this regard, in the following several recommendations for facilitating trust calibration before and during the use of automated systems are provided.

4.3.2 Trust Calibration Prior to Interaction

In regard to considerations how to design prior system information in the introduction of both new automated technologies on a societal level and specific automated systems on an individual level, it is an essential task for all stakeholders to provide future users with information aimed at creating a realistic picture of the functionality, abilities, and limitations of systems. In line with the assumptions of the presented framework and the findings of study 4, to maximize the effectiveness of such information, it should be designed to address attitude formation processes. For example, in line with the ELM, the type of information provided should be selected in a way that fits the expected cognitive resources and motivation of the users who process the information at the time of their presentation (like NFC in study 4). In the same manner, any other reason for motivated processing of information should lead to more impact of information that needs a certain effort to be fully understood. For example, highly motivated processing should be more likely in situations in which users perceive the information to be highly relevant for a current goal or in regard to a potentially risky or dangerous decision they have to make. This might, exemplarily, be the case when users are addressed with prior system information just before they about to use an automated system for the first time (if they are not experiencing excessive anxiety which might prevent central processing again). The other way around, if the information has either to be presented in a situation that leaves the user with limited time to process the information or if the information is directed at comparatively little motivated users, it should be designed in a way that heuristic processing leads to the desired outcomes. Such information should be easy to understand with surface characteristics of the information fostering the intended direction of attitude change (e.g., heuristic

information). In this regard, the reported findings of study 4 point into the direction that the provided information should not only be designed in consideration of the situation but also in a way to address the individual personality (and capabilities) of the user. Furthermore, in terms of user dispositions, in studies 1, 2, and 3, technology-specific traits (e.g., the propensity to trust and the affinity for technology) and prior attitudes towards the domain of automation (e.g., acceptability of automated driving, PAS) were found to play a role in the formation of learned trust. Based on the reported studies and in line with the framework, it can be concluded that such technology-specific traits and technology attitudes play an essential role in the perception and interpretation of prior system information. As a means to address these attitudes, information campaigns might be designed by public institutions or manufacturers in order to systematically evoke realistic global attitudes toward automated technology (e.g., acceptability of these systems). This includes, for example, realistic general information on the advantages, functional scope, capabilities, reliability, and potential dangers of such systems. Also, on a more individual level, differences in personality and attitudes might be addressed to facilitate trust calibration, for example, in personal contact or along with user profiles. In terms of findings from studies 1 and 2, people with higher neuroticism might be more sensitive to information on possible risks and dangers of using an automated system. The same holds true for users with low levels of self-esteem or self-efficacy, who might react with excessive anxiety when being presented with a new automated system (study 3). At the same time, those who are more technically inclined (higher affinity for technology) might tend to ignore such negative information and instead focus on the innovative character of the technology more. To address these individual differences in the weighting of the information, measures to personalize the provided information can be implemented, for example, in information campaigns, prior user training, and specific trust-related measures to provide "soft" learning possibilities. Based on the findings of study 6, these measures might include realistic information about possible malfunctions, as these do not necessarily reduce trust in user and might be important for trust calibration. For example, to prevent distrust in overly skeptical users (as reflected, for example, in negative general technology attitudes), specific information material or education videos could be produced that provide realistic information on risks and potential consequences of system error. The same possibilities for soft training measures prior to the first use might support users with a negative self-evaluation to build up a more realistic picture of their ability to functionally use the system. Furthermore, such users might be provided with possibilities to get to know a system under conditions that allow them to make first experiences with the system without potential negative outcomes (e.g., virtual reality simulations, supervised training). In these measures specifically irrational fears and overrepresented risks could be addressed to diminish potential distrust in these users. On the other hand, users who are overly optimistic about the system's reliability and might, therefore, misuse it and thereby endanger themselves and others in the surrounding of the task could be addressed with salient information on dangers, risks, and the associated negative consequences. In this manner, for example, training programs prior to system use could produce system malfunctions and clearly illustrate the consequences. In the face of the far-reaching consequences of prior information for the establishment of trust in an automated system and the strong relationship of trust in automation to actual system reliance, the power of advertisement should not be underestimated. There is a particular danger that by creating an overly positive impression of a system, advertising may lead users to overtrust a system. For this reason, advertising campaigns should strictly adhere to the described principle of calibrating rather than maximizing trust. In this

sense, advertising can also contribute to a realistic picture of a system's functions and capabilities at an early stage of trust formation.

For the domain of automated driving, a collection of promising directions to address users' attitudes towards automated driving is entailed in Lee and Kolodge's (2020) recent paper, in which they collected reasons for user attitudes towards self-driving vehicles in a text analysis approach. In regard to trust, the authors report mainly topics in regard to the dimension of performance (Lee & Moray, 1992) like system errors or situations beyond the functional scope of the driving automation. Above this, an implementation of a specific driver training for automated driver assistance within existing structures provides an excellent possibility to give drivers the possibility to learn to trust these systems realistically from the beginning. Driver education thereby should aim at systematically building up a realistic mental model of automated driving systems in general and supporting the establishment of realistic beliefs and attitudes. Furthermore, driver training could be implemented as part of the acquisition process of automated driving systems.

4.3.3 Trust Calibration During Interaction

In line with the propositions of the framework and findings of studies 4 and 6, trust calibration during the system interaction (stage 3) is essentially influenced by the processes at the preceding trust stages. In this regard, the measures discussed in the preceding paragraph are also essential for trust calibration during system interaction. On this basis, and in line with the findings of studies 5 and 6, trust is further calibrated during the interaction in a feedback process in which the system expectation is continuously compared to the observed system behavior and information in the user interface. Accordingly, a transparent system design enhances users' understanding of what the system does, why it does it, and what it is going to do (Seong & Bisantz, 2008; Chen et al., 2014; Christoffersen & Woods, 2002). Also, it should be transparent for the user if and how he can take back control and at what time which parts of a task can be conducted automatically. In the same manner, feedback on users' input might help them to better understand the system. While some of these information needs might be addressed in the design of the system itself, for example by enhancing observability of the system's internal processes, most of the necessary information in terms of transparency, feedback, and understanding has to be provided in a user interface bridging the gap between the user and the system. Further practical design implications for designing for system transparency are discussed, for example, in Lyons et al. (2017) or in the literature on human-machine cooperation (e.g., Christoffersen & Woods, 2002).

Also, user differences, for example, in terms of preferences and motivation, to attend to and use specific information should be taken into account in interface design to "better accommodate the trusting tendencies of diverse operators" (Hoff & Bashir, 2015, p. 415). In this regard, a personalized design of both the appearance and the contents of the user interface might enhance trust calibration. Findings of studies 1, 2, and 3 provide starting points for individualization in the design of user interfaces. For example, the user interface could provide different levels of detail on technological parameters like sensor functioning to account for individual differences in technology affinity. In the same manner, interfaces could provide neurotic or depressive users with information reducing irrational fears or provide those with low levels of self-esteem with positive feedback that strengthens their self-assessment to interact with the system functionally. In this respect, design decisions should be made on a case-by-case basis for each automated system in the specific context of the task, the characteristics of the user, and the situations in which it is used. In consideration of this complexity, it seems advisable to implement some kind of adaptability to changing

situations or changing states and preferences of users (e.g., de Visser & Parasuraman, 2011; Saffarian et al., 2012). In Lee and See (2004) as well as Hoff and Bashir (2015), further recommendations for a trustworthy design of automated systems are provided.

In the design of automated driving systems, besides others, user interfaces should provide transparency on the current status of the automation (automated or manual mode activated), on the automation's current representation of the situation (environment detection of the sensors), and maneuver planning (e.g., Debernard et al., 2016). An exemplary design approach to implement transparency is the augmentation of elements in the traffic scenarios in the driver's field of view, together with a graphical evaluation by the automation (e.g., Wintersberger et al., 2017). Intermediate levels of automation (SAE levels 3 to 4) call for the implementation of additional information in the user interface that enables dynamic task allocation and cooperative interaction between the driver and the automated vehicle, including, for example, the current allocation of subtasks and the temporal perspective of the current task allocation (e.g., Debernard et al., 2016; Dekker & Woods, 2002). The implementation of such information for cooperative real-time maneuver planning, control sharing, and feedback in the user interface is likely to enhance trust calibration during system use as it can be used to update the system expectation and trust. In this regard, interface design should be carefully fine-tuned to the dynamic capabilities of the automation, driver capabilities, expectations, changing traffic situations, and road conditions in order to optimize the process of trust calibration. Another important implication of the presented studies is that an experience of system malfunctions during the drive does not always result in negative long-term reductions in trust (studies 5 and 6). Instead, such experiences might also contribute to a realistic picture of the capabilities of an automated vehicle. Furthermore, if the consequences of such a malfunction are not extreme and do not endanger the driver, irrational fears in favor of trust calibration might be reduced. Thus, it might be advisable that automated driving systems provide transparent feedback on minor problems and malfunctions, helping the driver to better understand system limitations. On the basis of the ideas of cognitive work analysis, human-machine cooperation and transparency, Debernard et al. (2016) provided 12 principles for the decision which information to display during the drive in automated vehicles. Additionally, Ekman et al. (2018) provided an overview of design measures to facilitate trust calibration at different phases before and during the interaction with automated vehicles. Furthermore, additional implications for the interface design of cooperative automated vehicles are discussed in Walch et al. (2017) as well as Hoc et al. (2009).

All in all, while the described measures likely support trust calibration, not all might be feasible for all systems, contexts of use, or user groups. Therefore, and as the dangers of miscalibrated trust cannot be emphasized enough, stakeholders should carefully plan and evaluate all measures addressing trust in automated systems and test strategies thoroughly before implementation.

4.4 Strengths and Limitations

In the following, several strengths and limitations of the provided framework and studies are discussed on an integrative level.

4.4.1 Evaluation of the Proposed Framework

The proposed framework builds on and integrates theoretical work and empirical findings from the last decades of research in the domain of trust in automation. Essentially, it integrates, extents, and further specifies the assumptions of the models by Lee and See (2004) and Hoff and Bashir

(2015). Furthermore, theorizing from interpersonal trust was integrated into the framework along with a specification of psychological processes in the formation and calibration of trust in automation based on theorizing from other domains of psychology. In the framework, several specifications for the included variables and assumed processes at the different trust stages have been provided. For example, the theoretical conceptions for different trust layers have been extended by differentiating the propensity for trust from learned trust on a conceptual level. Furthermore, mental models, beliefs, and system expectations have been included. Also, study findings support the benefits of an application of attitude formation theories in providing a theoretical systematization for the investigating of prior system information in the formation of initial learned trust. Above this, attitude formation theory allows the integration of the effects of dispositional and situational variables to further enhance the understanding of trust formation prior to system use. Additionally, the concept of trust cues has been transferred to the domain of trust in automation and a preliminary classification of trust cues has been proposed. Furthermore, the proposed framework serves as a research basis for the integration and further investigation of trust in automation. Essentially, it can be used as a methodological guideline for the design of future research. On this basis, in chapter 4.2.6, a number of design recommendations for empirical studies on trust in automation have been provided. In this regard, one essential conclusion from the propositions of the framework is the distinction of the measurement of the propensity to trust in automation and learned trust. On this basis, in Appendix A, a scale for measuring learned trust in automation is included that was used in five of the six presented studies. While most propositions of the Three Stages of Trust framework gained initial empirical support in the depicted studies, all propositions yield further empirical investigation. Specifically, various of the proposed interrelations of variable groups were not covered in the presented studies. This includes, besides others, the investigation of propositions 7.3 and 10. In the same manner, for most of the entailed variable groups, only single variables have been investigated in the presented studies and the role of many other variables needs further investigation. Also, theorizing and research in the interpersonal domain of trust in other people has been growing over the last decades to a considerably complex and rich field. It is likely that the proposed framework can be further extended and specified on the basis of this work. Two possible directions that seem worthwhile to consider for extending the proposed framework would be a game-theoretic trust approach (Cabon-Dhersin & Ramani, 2007) or Brunswick's lens model (Brunswick, 1959; Morita & Burns, 2014). In the same manner, further theorizing from the field of human factors could be implemented into the framework, for example, the model of situation awareness (Endsley, 1995, 2015), which might enhance the understanding of using information to build up a representation of the current and future system interaction during the calibration of dynamic learned trust. Although the proposed framework is complex and, clearly, the proposed complexity is unlikely to be addressed within single research endeavors, the framework is intended to build a basis for integrating research incorporating some of this complexity. Hereby, the framework aims at providing directions for research and research design to further investigate the nature and dynamic of the formation and calibration of trust in automated systems in the future. In this regard, the depicted suggestions for future research may help to further specify the psychological processes involved in the formation and calibration of trust in automation and thereby contribute to an iterative refinement of the proposed framework in the scientific process.

4.4.2 Evaluation of the Studies

The depicted studies combine the strengths of different methods and study designs. While studies 1, 2, and 3 follow correlative designs offering the possibility to simultaneously investigate the

relationships of several variables, the experimental designs of studies 4, 5, and 6 allow for more experimental control and isolation of the effects of interest. In terms of the utilized study paradigms, studies 1, 2, and 4 were conducted in an online setting enabling an investigation of larger samples with more power and studies 3, 5, and 6 were conducted in a high-fidelity driving simulator allowing for higher external validity. Through the combination of these approaches, the individual shortcomings of the studies are, to some extent, balanced out, strengthening the combined evidence for the processes involved in trust formation. Furthermore, for all studies, besides the framework propositions, thorough theoretical considerations from the areas of general, social, and differential psychology, as well as human factors, provided the basis for hypothesizing and research design, which facilitates the understanding and integration of the study results. Also, the pairs of studies 1 and 2, as well as 5 and 6, provide combined evidence for several of the investigated hypotheses. All studies addressed the complexity of dynamic trust formation processes through a consideration of either mediated or moderated effects. Furthermore, trust was measured repeatedly over the course of most of the depicted studies, allowing for a clearer picture of trust development over time. To conclude, the technical domain of highly automated driving was kept constant in the reported studies, which further facilitates the integration of study findings.

However, as every research, the depicted studies have limitations, of which some will be discussed in the following. First, all studies have been conducted in the setting of automated driving. Although, this can be regarded as an advantage as described above and most of the established relationships and effects can be assumed to generalize to other domains of the interaction with automated technology, studies in other domains might be conducted to ensure the generalizability of the findings. Second, all studies were conducted within paradigms that only allow for restricted reality in terms of the external validity of findings. While the studies in the driving simulator provide a higher degree of realism, system interaction in the online studies was restricted to texts, pictures, and videos. While it is likely that the observed effects generalize to more realistic settings, some of the findings, in specific those of the online studies, should be validated in paradigms with a higher degree of realism. This is especially the case in terms of the perception of uncertainty and risk as these perceptions are per definition essential for the experience of trust. Third, the duration of the introduction processes and interactions with the reported automated driving systems were rather short-term. Future studies might explore the reported effects within longer time frames and repeated interaction with the system over several days or even longer time intervals. Fourth, at this time, the investigated highly automated driving systems are not available on the market in the presented form. Accordingly, the study participants had no prior experience with these systems and had to build up their mental models and expectations during the course of the studies. Fifth, a further limitation of all reported studies is a lack of the inclusion of a behavioral measure, which would allow for the investigation of generalization of the observed effects for trust in automation to reliance behavior. While the inclusion of such a behavioral measure would strengthen the implications of the reported studies, existing support for the relationship between trust in automation and actual reliance behavior is substantial (see chapters 1.1 and 1.2). Furthermore, as the relationship between trust in automation and actual behavior is assumed to be subject to moderation by individual and situational variables, even if such an objective behavioral measure would have been included, results would be difficult to be transferred into practical settings. Future research might further address the character of the relationship between trust in automation and reliance. Above this, the quality of trust calibration could be objectively assessed, for example, with the operationalization proposed by Merritt, Lee, et al. (2015). Sixth, for some studies, the investigated sample

was considerably small (studies 2 and 3) and therefore, these studies were prone to a higher beta error and neglecting relevant effects. In study 2, a procedure of effect size comparison with the findings of study 1 was applied to prevent an overly excessive beta error. In study 3, the power of the small sample was sufficient to provide statistical support for the reported findings, further strengthening the relevance of these effects. Seventh, for some of the studies, the average age of participants was relatively low, which might restrict the generalizability of the results of these studies. Eighth, for some variables, the utilized scales were rather short. While short scales were purposely used to prevent bias by unmotivated study processing, future studies might use longer scales for some of these variables. Ninth, related to the first criticism, as all study findings, the investigated relationships and effects might be validated in settings with other automated systems, task environments, and situations. Also, second-order studies like meta-analyses provide a promising perspective to further enhance the understanding of the generalizability of influencing variables and psychological processes in trust formation. A thorough discussion of additional limitations of the studies is included in the articles in Appendix B.

4.5 Conclusion

As calibrated trust in automation is essential for using automated systems appropriately, it is important to enhance the understanding of how trust in automation is built up, relates to reliance behavior, and how a calibrated level of trust can be supported in practice. The Three Stages of Trust framework integrates and extends the theoretical basis for human factors research on trust in automation by proposing a process in which trust in a specific automated system is formed through three successive trust stages. Additionally, the framework integrates variables on the side of the user, the system, and the situation. In this respect, the proposed framework provides a structure for the consideration of a certain complexity in trust formation for both research and practice. The framework, in particular, emphasizes that the level of trust in a certain system at a specific point of time is the result of a personal learning history in which different information is used in the formation of three different trust layers at the three trust stages. Taken together, on the theoretical basis of the framework propositions, numerous research questions and hypotheses for future research can be derived (chapter 4.2) and decisions regarding the design of research on trust in automation can be informed (as provided in chapter 4.2.6).

In this dissertation, six studies were presented which support some of the central propositions of the framework. In particular, they support the role of personality differences, general technology-related attitudes, and users' emotional states in the development of learned trust. Also, it was supported that the type of available information influences trust formation and is thereby interpreted differently by different users, depending on their personal preferences and abilities. Additionally, the presented results provide insights into the processes involved in trust calibration. In this regard, the findings of studies 5 and 6 point into the direction that dynamic learned trust develops during the interaction with an automated system in a dynamic feedback loop in which the perceived system behavior and the associated output of the user interface are compared. On the basis of both the framework and the presented studies, practical implications were discussed (chapter 4.3).

Taken together, calibrated trust in automated systems is essential for a safe, efficient, and satisfying interaction with such systems, and therefore, a thorough understanding of its psychological basis is important for designing socio-technical systems with respect to these goals. The presented dissertation substantially supports this endeavor by providing: a) a framework to integrate current

research, b) new empirical evidence for how trust in automation is built up and which variables influence this process, c) directions for future research, and d) practical implications for educating users and for the design of trustworthy automated systems that facilitate trust calibration and in the result a calibrated level of trust. Through these contributions, this work might help to accumulate knowledge on the formation of trust in automation. Through an integration of this knowledge into the practice of designing sophisticated automated systems and the strategies to introduce them, future users are essentially supported in building up an appropriate level of trust and using these systems for their individual and society's benefit. In this manner, human factors research might further enhance a consequent consideration of psychological processes underlying the interaction with technology at all stages of product development and market introduction and thereby essentially contribute to creating technical products that meet the needs and psychological requirements of users and therefore are likely to be safe and enjoyable to use.

References

Ajzen, I. (1988). *Attitudes, personality, and behavior.* (1st ed). Dorsey Press.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179-211.

Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology*, *52*(1), 27-58. https://doi.org/10.1146/annurev.psych.52.1.27

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology, 32*(4), 665-683. https://doi.org/10.1111/j.1559-1816.2002.tb00236.x

Ajzen, I. (2005). *Attitudes, personality, and behavior.* (2nd ed.). Open University Press.

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

Albarracín, D., Chan, M. P. S., & Jiang, D. (2018). Attitudes and Attitude change: Social and personality considerations about specific and general patterns of behavior. In K. Deux, & M. Snyder (Eds.), *The Oxford handbook of personality and social psychology* (pp. 439-463). Oxford University Press.

Allport, G. W. (1961). *Pattern and growth in personality*. Holt, Reinhart & Winston.

Atkinson, D., Hancock, P., Hoffman, R. R., Lee, J. D., E., Stokes, C., & Wagner, A. R. (2012). Trust in computers and robots: the uses and boundaries of the analogy to interpersonal trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 303-307. https://doi.org/10.1177/1071181312561071

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, *50*(2), 248-287. https://doi.org/10.1016/0749-5978(91)90022-L

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. https://doi.org/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. https://doi.org/10.1016/j.trf.2015.10.005

Beiker, S. (2016). Deployment scenarios for vehicles with higher-order automation. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous driving: Technical, legal and social aspects* (pp. 193-211). Springer. https://doi.org/10.1007/978-3-662-48847-8_10

Belk, R. W. (1985). Materialism: Trait aspects of living in the material world. *Journal of Consumer Research, 12*(3), 265–280. https://doi.org/10.1086/208515

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino: Multivariate Software.

Billings, D. R., Schaefer, K. E., Chen, J. Y. C., & Hancock, P. A. (2012). Human-robot interaction. *Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction - HRI '12*. https://doi.org/10.1145/2157689.2157709

Biros, D. P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, *13*(2), 173-189. https://doi.org/10.1023/b:grup.0000021840.85686.57

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods, 17*(3), 399. https://doi.org/10.1037/a0028087

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F., & Schwarz, N. (1994). Need for cognition: eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben [Need for cognition: a scale measuring engagement and happiness in cognitive tasks]. *Zeitschrift für Sozialpsychologie*, *25*, 147-154.

Briggs, P., Burford, B., & Dracup, C. (1998). Modelling self-confidence in users of a computer-based system showing unrepresentative design. *International Journal of Human-Computer Studies, 49*(5), 717-742. https://doi.org/10.1006/ijhc.1998.0224

Brosnan, M. J. (1998). The impact of computer anxiety and self-efficacy upon performance. *Journal of Computer Assisted Learning*, *14*(3), 223-234. https://doi.org/10.1046/j.1365-2729.1998.143059.x

Brunswik, E., 1952. *The conceptual framework of psychology.* University of Chicago Press.

Buckley, L., Kaye, S.-A., & Pradhan, A. K. (2018). Psychosocial factors associated with intended use of automated vehicles: A simulated driving study. *Accident Analysis & Prevention*, *115*, 202-208. https://doi.org/10.1016/j.aap.2018.03.021

Cabon-Dhersin, M.-L., & Ramani, S. V. (2007). Opportunism, trust and cooperation: A game theoretic approach with heterogeneous agents. *Rationality and Society, 19*(2), 203–228. https://www.doi.org/10.1177/1043463107077391

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306-307. https://doi.org/10.1207/s15327752jpa4803_13

Cacioppo, J. T., Petty, R. E., Kao, C. F., & Rodriguez, R. (1986). Central and peripheral routes to persuasion: An individual difference perspective. *Journal of Personality and Social Psychology, 51*(5), 1032-1043. https://doi.org/10.1037/0022-3514.51.5.1032

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology, 45*(4), 805-818. https://doi.org/10.1037/0022-3514.45.4.805

Carroll, J. M., & Olson, J. R. (1987). Mental models in human-computer interaction. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 45-65). North Holland.

Carver, C. S. (2004). Self-regulation of action and affect. In R. F. Baumeister & K. D. Vohs (Eds.), *Handbook of self-regulation: Research, theory, and applications* (p. 13–39). The Guilford Press.

Chaiken, S., & Trope, Y. (Ed.). (1999). *Dual-process theories in social psychology.* Guilford Press.

Chancey, E. T., Bliss, J. P., Proaps, A. B., & Madhavan, P. (2015). The role of trust as a mediator between system characteristics and response behaviors. *Human Factors*, *57*(6), 947-958. https://doi.org/10.1177/0018720815582261

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. H. (2017). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors*, *59*(3), 333–345. https://doi.org/10.1177/0018720816682648

Chen, S., Shechter, D., & Chaiken, S. (1996). Getting at the truth or getting along: Accuracy-versus impression-motivated heuristic and systematic processing. *Journal of Personality and Social Psychology, 71*(2), 262-275. https://doi.org/10.1037/0022-3514.71.2.262

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

Chien S.-Y., Semnani-Azad, Z., Lewis, M., & Sycara, K. (2014) Towards the development of an inter-cultural scale to measure trust in automation. In P. L. Rau. (Ed.), *Cross-cultural design. CCD 2014. Lecture notes in computer science, vol. 8528* (pp. 35-46). Springer, Cham

Chien, S.-Y., Sycara, K., Liu, J.-S., & Kumru, A. (2016). Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and big five personality traits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 841-845. https://doi.org/10.1177/1541931213601192

Christoffersen, K. and Woods, D.D. (2002), How to make automated systems team players, *Advances in Human Performance and Cognitive Engineering Research Vol. 2*, (pp. 1-12). Emerald Group Publishing Limited. https://doi.org/10.1016/S1479-3601(02)02003-9

Conner, M., Godin, G., Sheeran, P., & Germain, M. (2013). Some feelings are more important: Cognitive attitudes, affective attitudes, anticipated affect, and blood donation. *Health Psychology*, *32*(3), 264-272. https://doi.org/10.1037/a0028500

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R professional manual*. Psychological Assessment Resources.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* University of Illinois Press.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cummings, C. M., Caporino, N. E., & Kendall, P. C. (2014). Comorbidity of anxiety and depression in children and adolescents: 20 years after. *Psychological Bulletin, 140*(3), 816-845. https://doi.org/10.1037/a0034733

Czarnecki, K. (2018). Operational Design Domain for Automated Driving Systems - Taxonomy of Basic Terms. https://doi.org/10.13140/RG.2.2.18037.88803.

Daly, M. A. (2002). *Task load and automation use in an uncertain environment* [Master's thesis, Air Force Institute of Technology]. ProQuest Dissertations and Theses.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management Science, 35*(8), 982-1003. https://doi.org/10.1287/mnsc.35.8.982

Debernard, S., Chauvin, C., Pokam, R., & Langlois, S. (2016). Designing human-machine interface for autonomous vehicles. *IFAC-PapersOnLine*, *49*(19), 609-614. https://doi.org/10.1016/j.ifacol.2016.10.629

de Visser, E., & Parasuraman, R. (2011). Adaptive aiding of human-robot teaming: effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making*, *5*(2), 209-231. https://doi.org/10.1177/1555343411410160

de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies, 58*(6), 719-735. https://doi.org/10.1016/S1071-5819(03)00039-9

Dekker, S. W. A., & Woods, D. D. (2002). MABA-MABA or abracadabra? Progress on human-automation co-ordination. *Cognition, Technology & Work*, *4*(4), 240-244. https://doi.org/10.1007/s101110200022

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human relations*, *13*(2), 123-139. https://doi.org/10.1177/001872676001300202

Donmez, B., Boyle, L. N., Lee, J. D., & McGehee, D. V. (2006). Drivers' attitudes toward imperfect distraction mitigation strategies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *9*(6), 387-398. https://doi.org/10.1016/j.trf.2006.02.001

Durso, F. T., & Gronlund, S. D. (1999). Situation awareness. In F. T. Durso (Ed.), *Handbook of applied cognition* (p. 283-314). John Wiley & Sons Ltd.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*(6), 697-718. https://doi.org/10.1016/S1071-5819(03)00038-7

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79-94. https://doi.org/10.1518/0018720024494856

Ekman, F., Johansson, M., & Sochor, J. (2018). Creating appropriate trust for autonomous vehicle systems: A framework for HMI design. *IEEE Transactions on Human-Machine Systems*, *48*(1), 95-101. http://doi.org/10.1109/THMS.2017.2776209

Ekman, F., Johansson, M., & Sochor, J. L. (2016). 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, 9*, 1-4. https://doi.org/10.1145/2971485.2971551

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

Emzivat, Y., Ibanez-Guzman, J., Martinet, P., & Roux, O. H. (2017). Dynamic driving task fallback for an automated driving system whose ability to monitor the driving environment has been compromised. *2017 IEEE Intelligent Vehicles Symposium (IV),* 1841-1847. https://hal.archives-ouvertes.fr/hal-01724931

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32-64. https://doi.org/10.1518/001872095779049543

Endsley, M. R. (2015). Situation awareness misconceptions and misunderstandings. J*ournal of Cognitive Engineering and Decision Making, 9*(1), 4–32. https://doi.org/10.1177/1555343415572631

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality, 42*(6), 1585-1593. https://doi.org/10.1016/j.jrp.2008.07.011

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol.23, pp. 75-109). AcademicPress.

Fellner, B., Holler, M., Kirchler, E., & Schabmann, A. (2007). Regulatory focus scale (RFS): Development of a scale to record dispositional regulatory focus. *Swiss Journal of Psychology, 66*(2), 109-116. https://doi.org/10.1024/1421-0185.66.2.109

Ferring, D. & Filipp, S.-H. (1996). Messung des Selbstwertgefühls: Befunde zu Reliabilität, Validität und Stabilität der Rosenberg-Skala [Measurment of self-esteem: Findings on reliability, validity, and stability of the Rosenberg scale]. *Diagnostica, 42*, 284-292.

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

Forgas, J. P. (1995). Mood and judgment: the affect infusion model (AIM). *Psychological Bulletin, 117*(1), 39-66. https://doi.org/10.1037/0033-2909.117.1.39

Forster, Y., Kraus, J., Feinauer, S., & Baumann, M. (2018). 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*, 118-128. https://doi.org/10.1145/3239060.3239070

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

French, B., Duenser, A., Heathcote, A. (2018). *Trust in automation – A literature review.* (CSIRO Report EP184082). CSIRO, Australia. http://www.tascl.org/uploads/4/9/3/3/49339445/trust_in_automation_-_a_literature_review_report.pdf

Frison, A. K., Wintersberger, P., Riener, A., Schartmüller, C., Boyle, L. N., Miller, E., & Weigl, K. (2019). In UX we trust: Investigation of aesthetics and usability of driver-vehicle interfaces and their impact on the perception of automated driving. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Gaines, S. O., Panter, A. T., Lyde, M. D., Steers, W. N., Rusbult, C. E., Cox, C. L., & Wexler, M. O. (1997). Evaluating the circumplexity of interpersonal traits and the manifestation of interpersonal traits in interpersonal trust. *Journal of Personality and Social Psychology, 73*(3), 610-623. https://doi.org/10.1037/0022-3514.73.3.610

Galvin, B. M., Randel, A. E., Collins, B. J., & Johnson, R. E. (2018). Changing the focus of locus (of control): A targeted review of the locus of control literature and agenda for future research. *Journal of Organizational Behavior*, *39*(7), 820-833. https://doi.org/10.1002/job.2275

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*(1), 1938-1942. https://doi.org/10.1177/1541931213571433

Gold, C., Körber, M., Hohenberger, C., Lechner, D., & Bengler, K. (2015). Trust in automation – before and after the experience of take-over scenarios in a highly automated vehicle. *Procedia Manufacturing, 3*(1), 3025–3032. https://doi.org/10.1016/j.promfg.2015.07.847

Gurtman, M. B. (1992). Trust, distrust, and interpersonal problems: A circumplex analysis. *Journal of Personality and Social Psychology, 62*(6), 989-1002. https://doi.org/10.1037/0022-3514.62.6.989

Hacker,W. (2006). *Allgemeine Arbeitspsychologie. Psychische Regulation von Arbeitstätigkeiten [General work psychology. Psychological regulation of work activities].* Huber.

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. M. (2007). The integrated trait--state model. *Journal of Research in Personality, 41*(2), 295–315. https://doi.org/10.1016/j.jrp.2006.04.003

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, 53*(5), 517-527. https://doi.org/10.1177/0018720811417254

Harter, S. (1993). Causes and consequences of low self-esteem in children and adolescents. In R. Baumeister (Ed.), *Self-esteem: The puzzle of low self-regard* (pp. 87-116). Plenum.

Hartwich, F., Witzlack, C., Beggiato, M., & Krems, J. F. (2019). 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 Behaviour, 65*, 522-535. https://doi.org/10.1016/j.trf.2018.05.012

Hautzinger, M., Keller, F., & Kühner, C. (2009). *BDI-II. Beck-Depressions-Inventar Revision* (2nd ed.) [Beck Depression-Inventory: BDI II; Manual]. Pearson Assessment.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*(4), 408-420. https://doi.org/10.1080/03637750903310360

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications*, 210-217. https://doi.org/10.1145/2516540.2516554

Hergeth, S., Lorenz, L, & Krems, J. (2017). Prior familiarization with takeover requests affects drivers' takeover performance and automation trust. *Human Factors, 59*(3), 457-470. https://doi.org/10.1177/0018720816678714

Hergeth, S., Lorenz, L., Krems, J. F., & Toenert, L. (2015). Effects of take-over requests and cultural background on automation trust in highly automated driving. *Proceedings of the Eighth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design,* 331-337. https://doi.org/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*(3), 509-519. https://doi.org/10.1177/0018720815625744

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist, 52*(12), 1280-1300. https://doi.org/10.1037/0003-066X.52.12.1280

Hoc, J.-M. (2000). From human-machine interaction to human-machine cooperation. *Ergonomics*, *43*(7), 833-843. https://doi.org/10.1080/001401300409044

Hoc, J.-M. (2001). Towards a cognitive approach to human-machine cooperation in dynamic situations. *International Journal of Human-Computer Studies*, *54*(4), 509-540. https://doi.org/10.1006/ijhc.2000.0454

Hoc, J.-M., Young, M. S., & Blosseville, J.-M. (2009). Cooperation between drivers and automation: Implications for safety. *Theoretical Issues in Ergonomics Science, 10*(2), 135-160. https://doi.org/10.1080/14639220802368856

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407-434. https://doi.org/10.1177/0018720814547570

Holmes, J. G., & Rempel, J. K. (1989). Trust in close relationships. In C. Hendrick (Ed.), *Review of personality and social psychology, Vol. 10. Close relationships* (p. 187-220). Sage Publications.

Horikawa, M., & Yagi, A. (2012). The relationships among trait anxiety, state anxiety and the goal performance of penalty shoot-out by university soccer players. *PloS One, 7*(4). https://doi.org/10.1371/journal.pone.0035727

Howell, D. C. (2007). The treatment of missing data. In W. Outhwaite, & S. Turner (Eds.), *Handbook of social science methodology* (pp. 208-224). Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Jerusalem, M. & Schwarzer, R. (1999). Allgemeine Selbstwirksamkeitserwartung [General expectation of self-efficacy]. In R. Schwarzer & M. Jerusalem (Eds.), *Skalen zur Erfassung von Lehrer- und Schülermerkmalen* (pp. 57-59). Freie Universität Berlin.

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

Karrer, K., Glaser, C., Clemens, C., & Bruder, C. (2009). Technikaffinität erfassen - der Fragebogen TA-EG. In A. Lichtenstein, C. Stößel, & C. Clemens (Eds.), *Der Mensch im Mittelpunkt technischer Systeme, 8. Berliner Werkstatt Mensch-Maschine-Systeme* (pp. 196-201). VDI Verlag GmbH.

Kazi, T. A., Stanton, N. A., Walker, G. H., & Young, M. S. (2007). Designer driving: Drivers' conceptual model and level of trust in adaptive-cruise-control. *International Journal of Vehicle Design, 45*(3). https://doi.org/10.1504/IJVD.2007.014909

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2017). Calibrating trust to increase the use of automated systems in a vehicle. In N. A. Stanton, S. Landry, G. Di Bucchianico, & A. Vallicelli (Eds.), *Advances in human aspects of transportation* (pp. 535-546). Springer International Publishing. https://doi.org/ 10.1007/978-3-319-41682-3_45

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2018). Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation Research Part C: Emerging Technologies*, *96*, 290-303. https://doi.org/10.1016/j.trc.2018.07.001

Kim, S. S., Chung, Y., Perry, M. J., Kawachi, I., and Subramanian, S. V. (2012). Association between interpersonal trust, reciprocity, and depression in South Korea: a prospective analysis. PLoS One 7(1), 1-8. http://doi.org/10.1371/journal.pone.0030602

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a „team player" in joint human-agent activity. *IEEE Intelligent Systems*, *19*(6), 91-95. https://doi.org/10.1109/mis.2004.74

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

Koustanaï, A., Cavallo, V., Delhomme, P., & Mas, A. (2012). Simulator training with a forward collision warning system: Effects on driver-system interactions and driver trust. *Human Factors*, *54*(5), 709-721. https://www.doi.org/10.1177/0018720812441796

Kraus, J. M., Forster, Y., Hergeth, S., & Baumann, M. (2019). Two routes to trust calibration: Effects of reliability and brand information on trust in automation. *International Journal of Mobile Human Computer Interaction,* 11(3), 1-17. https://doi.org/10.4018/IJMHCI.2019070101

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

Kraus, J., Scholz, D., & Baumann, M. (2020). What's driving me? - Exploration and validation of a hierarchical personality model for trust in automated driving. *Human Factors*. Advance online publication. https://doi.org/10.1177/0018720820922653

Kraus, J., Scholz, D., Messner, E.-M., Messner, M., & Baumann, M. (2020). Scared to trust? – Predicting trust in highly automated driving by depressiveness, negative self-evaluations and state anxiety. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02917

Krohne, H. W., Egloff, B., Kohlmann, C.-W., & Tausch, A. (1992). Untersuchung mit einer deutschen Version der „Positive and Negative Affect Schedule“ (PANAS) [Investigation with a German version of the „Positive and Negative Affect Schedule“ (PANAS)]. *Diagnostica, 42,* 139-156.

Large, D. R., & Burnett, G. E. (2014). The effect of different navigation voices on trust and attention while using in-vehicle navigation systems. *Journal of Safety Research*, *49*, 69, 1-75. https://doi.org/10.1016/j.jsr.2014.02.009

Larsen, R. J., & Buss, D. M. (2005). *Personality psychology: Domains of knowledge about human nature*. McGraw Hill

Larsen, R. J., & Ketelaar, T. (1991). Personality and susceptibility to positive and negative emotional states. *Journal of Personality and Social Psychology, 61*(1), 132-140. https://doi.org/10.1037/0022-3514.61.1.132

Lee, J. D., & Kolodge, K. (2020). Exploring trust in self-driving vehicles through text analysis. *Human Factors*, *62*(2), 260–277. https://doi.org/10.1177/0018720819872672

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

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

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Lester, D., & Gatto, J. L. (1990). Interpersonal trust, depression, and suicidal ideation in teenagers. *Psychological Reports, 67*(3), 786-786. https://doi.org/10.2466/pr0.1990.67.3.786

Letzring, T. D., & Adamcik, L. A. (2015). Personality traits and affective states: Relationships with and without affect induction. *Personality and Individual Differences, 75*, 114-120. https://doi.org/10.1016/j.paid.2014.11.011

Lewandowsky, S., Mundy, M., & Tan, G. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied, 6*(2), 104-123. https://doi.org/10.1037/1076-898X.6.2.104

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy, 33*(3), 335-343.

Lyons, J. B., & Guznov, S. Y. (2019). Individual differences in human-machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science, 20*(4), 440-458. https://doi.org/10.1080/1463922X.2018.1491071

Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., Smith, D., Johnson, W., & Shively, R. (2017). Shaping trust through transparent design: Theoretical and experimental guidelines. In P. Savage-Knepshield & J. Chen (Eds.), *Advances in Human Factors in robots and unmanned systems* (pp. 127-136). Springer International Publishing.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Lawrence Erlbaum Associates.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277-301. https://doi.org/10.1080/14639220500337708

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *11th Australasian Conference on Information Systems*, *53*, 6-8.

Maio, G. R., Haddock, G., & Verplanken, B. (2018). *The psychology of attitudes and attitude change*. Sage Publications Limited.

Marr, B. (2019, September 30). *The 7 biggest technology trends in 2020 everyone must get ready for now*. Forbes. https://www.forbes.com/sites/bernardmarr/2019/09/30/the-7-biggest-technology-trends-in-2020-everyone-must-get-ready-for-now/#561855ee2261

Marsh, S., & Dibben, M. R. (2003). The role of trust in information science and technology. *Annual Review of Information Science and Technology*, *37*(1), 465-498. https://doi.org/10.1002/aris.1440370111

Marteau, T.M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology, 31,* 301-306. https://doi.org/10.1111/j.2044-8260.1992.tb00997.x

Mathews, A., & Mackintosh, B. (1998). A cognitive model of selective processing in anxiety. *Cognitive Therapy and Research*, *22*(6), 539-560. https://doi.org/10.1023/a:1018738019346

Maurer, M., Gerdes, J. C., Lenz, B., & Winner H. (2016). *Autonomous driving: Technical, legal and social aspects* (pp. 193-211). Springer.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709-734. https://doi.org/10.5465/amr.1995.9508080335

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175-215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

Merritt, S. M. (2011). Affective processes in human-automation interactions. *Human Factors, 53*(4), 356-370. https://doi.org/10.1177/0018720811411912

Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., & Shirase, L. (2019). Automation-Induced Complacency Potential: Development and Validation of a New Scale. *Frontiers in Psychology, 10*, 225. https://doi.org/10.3389/fpsyg.2019.00225

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: effects of implicit attitudes toward automation on trust in an automated system. *Human Factors, 55*(3), 520-534. https://doi.org/10.1177/0018720812465081

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*(2), 194-210. https://doi.org/10.1518/001872008X288574

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors*, *57*(1), 34–47. https://doi.org/10.1177/0018720814561675

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors, 57*(5), 740-753. https://doi.org/10.1177/0018720815581247

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. Henry Holt and Co. https://doi.org/10.1037/10039-000

Moray, N., & Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control, 21*(4/5), 203-211. https://doi.org/10.1177/014233129902100408

Morita, P. P., & Burns, C. M. (2014) Understanding 'interpersonal trust' from a human factors perspective: Insights from situation awareness and the lens model. *Theoretical Issues in Ergonomics Science, 15*(1), 88-110. https://doi.org/10.1080/1463922X.2012.691184

Mowen, J. C. (2000). *The 3M Model of Motivation and Personality: Theory and empirical applications to consumer behavior.* Springer Science & Business Media.

Mowen, J. C., Park, S., & Zablah, A. (2007). Toward a theory of motivation and personality with application to word-of-mouth communications. *Journal of Business Research*, *60*(6), 590-596. https://doi.org/10.1016/j.jbusres.2006.06.007

Mühl, K., Strauch, C., Grabmaier, C., Reithinger, S., Huckauf, A., & Baumann, M. (2019). Get ready for being chauffeured: Passenger's preferences and trust while being driven by human and automation. *Human Factors*. https://doi.org/10.1177/0018720819872893

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*(5-6), 527-539. https://doi.org/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*(11), 1905-1922. https://doi.org/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*(3), 429-460. https://doi.org/10.1080/00140139608964474

Murphy, J. M., Horton, N. J., Laird, N. M., Monson, R. R., Sobol, A. M., & Leighton, A. H. (2004). Anxiety and depression: a 40-year perspective on relationships regarding prevalence, distribution, and comorbidity. *Acta Psychiatrica Scandinavica, 109*(5), 355-375. https://doi.org/10.1111/j.1600-0447.2003.00286.x

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

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72-78.

Nutt, D. J. (1999). Care of depressed patients with anxiety symptoms. *The Journal of Clinical Psychiatry. 60*, 23-27.

Özkan, T., Lajunen, T., & Kaistinen, J. (2005). Traffic Locus of control, driving skills, and attitudes towards in-vehicle technologies (ISA & ACC). *Proceedings of the 18th 1247 International Cooperation on Theories and Concepts in Traffic Safety (ICTCT)*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). https://doi.org/10.1126/science.aac4716

Owens, T. J. (1993). Accentuate the positive-and the negative: Rethinking the use of self-esteem, self-deprecation, and self-confidence. *Social Psychology Quarterly, 56*(4), 288-299. https://doi.org/10.2307/2786665

Parasuraman, R., Bahri, T., Deaton, J. E., Morrison, J. G., & Barnes, M. (1992). *Theory and design of adaptive automation in aviation systems*. Catholic University of America Washington DC Science Lab.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230-253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30*(3), 286-297. https://doi.org/10.1109/3468.844354

Payre, W., Cestac, J., & Delhomme, P. (2016). Fully automated driving: Impact of trust and practice on manual control recovery. *Human Factors, 58*(2), 229-241. https://doi.org/10.1177/0018720815612319

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 123-205). Academic Press.

Phillips, R. R., & Madhavan, P. (2011). The effect of distractor modality and processing code on human-automation interaction. *Cognition, Technology & Work*, *13*(4), 233-244. https://doi.org/10.1007/s10111-010-0163-2

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879-891. https://doi.org/10.3758/BRM.40.3.879

Rajaonah, B., Anceaux, F. & Vienne, F. (2006). Trust and the use of adaptive cruise control: a study of a cut-in situation. *Cognition, Technology & Work, 8*, 146-155. https://doi.org/10.1007/s10111-006-0030-3

Rani, M. R. A., Sinclair, M. A., & Case, K. (2000). Human mismatches and preferences for automation. *International Journal of Production Research*, *38*(17), 4033–4039. https://doi.org/10.1080/00207540050204894

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

Richins, M. L., & Dawson, S. (1992). A consumer values orientation for materialism and its measurement: Scale development and validation. *Journal of Consumer Research, 19*(3), 303-315. https://doi.org/10.1086/209304

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5-12 (BETA). *Journal of Statistical Software, 48*(2), 1-36.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of 1130 reinforcement. *Psychological Monographs, 80*(1), 1-28. https://doi.org/10.1037/h0092976

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality, 35*(4), 651-665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist, 35*(1), 1–7. https://doi.org/10.1037/0003-066X.35.1.1

Rotter, J. B. (1990). Internal versus external control of reinforcement: A case history of a variable. *American Psychologist, 45*(4), 489-493. https://doi.org/10.1037/0003-066X.45.4.489

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

Rovira, E., Pak, R., & McLaughlin, A. (2017). Effects of individual differences in working memory on performance and trust with various degrees of automation. *Theoretical Issues in Ergonomics Science*, *18*(6), 573-591. https://doi.org/10.1080/1463922X.2016.1252806

SAE (2018). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (J3016_201806)*. SAE International. https://doi.org/10.4271/J3016_201806

Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Reliability and age-related effects on trust and reliance of a decision support aid. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(3), 586-589. https://doi.org/10.1177/154193120404800366

Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2011). Understanding reliance on automation: Effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science, 15*(2), 134-160. https://doi.org/10.1080/1463922X.2011.611269

Satorra, A., & Bentler, E. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye, & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). SAGE Publications.

Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T. L., Chen, J. Y., & Hancock, P. A. (2014). *A meta-analysis of factors influencing the development of trust in automation: Implications for human-robot interaction*. Army Research Laboratory.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377-400. https://doi.org/10.1177/0018720816634228

Saffarian, M., de Winter, J. C. F., & Happee, R. (2012). Automated driving: Human-factors issues and design solutions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56*(1), 2296–2300. https://doi.org/10.1177/1071181312561483

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist, 26*(3-4) 207-231. https://doi.org/10.1207/s15326985ep2603&4_2

Schwarz, N., & Clore, G. L. (1988). How do I feel about it? Informative functions of affective states. In K. Fiedler & J. Forgas (Eds.), *Affect, cognition, and social behavior* (pp. 44-62). Hogrefe.

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*(7-8), 608-625. https://doi.org/10.1016/j.ergon.2008.01.007

Sherman, S. J., & Fazio, R. H. (1983). Parallels between attitudes and traits as predictors of behavior. *Journal of Personality*, *51*(3), 308-345. https://doi.org/10.1111/j.1467-6494.1983.tb00336.x

Simpson, J. A. (2007). Foundations of interpersonal trust. In A. W. Kruglanski, & Higgins, E. T. (Eds.), *Social psychology: Handbook of basic principles* (pp. 587-607). The Guilford Press.

Sniehotta, F F., Presseau, J., & Araújo-Soares, V. (2014). Time to retire the theory of planned behaviour, *Health Psychology Review, 8*(1), 1-7. https://doi.org/1010.1080/17437199.2013.869710.

Sowislo, J. F., & Orth, U. (2013). Does low self-esteem predict depression and anxiety? A meta-analysis of longitudinal studies. *Psychological Bulletin, 139*(1), 213-240. https://doi.org/10.1037/a0028931

Spector, P. E. (1982). Behavior in organizations as a function of employee's locus of control. *Psychological Bulletin, 91*(3), 482-497. https://doi.org/10.1037/0033-2909.91.3.482

Steiger, J. H., & Lind, J. C. (1980). *Statistically-based tests for the number of common factors.* Handout for a presentation delivered at the meeting of the Psychometric Society. Iowa.

Steiger, J. H. (2016). Notes on the Steiger-Lind (1980) Handout. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 777-781. https://doi.org/10.1080/10705511.2016.1217487

Stokes, C. K., Lyons, J. B., Littlejohn, K., Natarian, J., Case, E., & Speranza, N. (2010). Accounting for the human in cyberspace: Effects of mood on trust in automation. *2010 International Symposium on Collaborative Technologies and Systems*, Chicago, IL, 2010, pp. 180-187.

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: the five factor model of personality. *Journal of Experimental Psychology: Applied, 17*(2), 71-96. https://doi.org/10.1037/a0024170

Thielmann, I., & Hilbig, B. E. (2015). Trust: an integrative review from a person-situation perspective. *Review of General Psychology*, *19*(3), 249-277. https://doi.org/10.1037/gpr0000046

TÜV | DEKRA arge tp 21 (2011). VICOM editor [computer software]. https://www.vicomeditor.de/

Venkatesh, V., Morris, M., Davis, G., & Davis, F. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425. https://doi.org/10.2307/30036540

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 G. Meixner & C. Müller (Eds.), *Automotive User Interfaces: Creating Interactive Experiences in the Car* (pp. 273-294). Springer International Publishing. https://doi.org/10.1007/978-3-319-49448-7_10

Walker, F., Wang, J., Martens, M., & Verwey, W. (2019). Gaze behaviour and Electrodermal Activity: Objective measures of drivers' trust in automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour, 64,* 401-412. https://doi.org/10.1016/j.trf.2019.05.021

Walker, G.H., Stanton, N.A., & Salmon, P. (2016) Trust in vehicle technology. *International Journal of Vehicle Design, 70* (2), 157-182. https://doi.org/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, 51*(3), 281-291. https://doi.org/10.1177/0018720809338842

Wang, L., Jamieson, G. A., & Hollands, J. G. (2011). The effects of design features on users' trust in and reliance on a combat identification system. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 375-379. https://doi.org/10.1177/1071181311551077

Wang, S., & Li, Z. (2019). Exploring the mechanism of crashes with automated vehicles using statistical modeling approaches. *PLoS ONE, 14*(3), 1-16. https://doi.org/10.1371/journal.pone.0214550

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

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209-231). Guilford Press.

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*(4), 352-367. https://doi.org/10.1080/14639220110110306

Wintersberger, P., Frison, A.-K., Riener, A., & von Sawitzky, T. (2019). Fostering user acceptance and trust in fully automated vehicles: Evaluating the potential of augmented reality. *PRESENCE: Virtual and Augmented Reality*, *27*(1), 46-62. https://doi.org/10.1162/pres_a_00320

Wintersberger, P., von Sawitzky, T., Frison, A.-K., & Riener, A. (2017). Traffic augmentation as a means to increase trust in automated driving systems. *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*. https://doi.org/10.1145/3125571.3125600

Woods, D. (2001). *Human-centered design of automated agents and human-automation team play.* Cognitive-Systems Engineering Laboratory, Ohio State University.

Xiao, W., Zhou, L., Wu, Q., Zhang, Y., Miao, D., Zhang, J., & Peng, J. (2014). Effects of person-vocation fit and core self-evaluation on career commitment of medical university students: The mediator roles of anxiety and career satisfaction. *International Journal of Mental Health Systems*, *8*(1), 8. https://doi.org/10.1186/1752-4458-8-8

Young, M. S., Stanton, N. A., & Harris, D. (2007). Driving automation: learning from aviation about design philosophies. *International Journal of Vehicle Design*, *45*(3), 323. https://doi.org/10.1504/ijvd.2007.014908

Appendix

A: The Short Learned Trust in Automation Scale, German Version (LETRAS-G)

B: Research Articles

Appendix A

The Short Learned Trust in Automation Scale, German Version (LETRAS-G)

In the past 30 years, a manifold of different methods has been proposed and used to measure trust in automation (e.g., Jian et al., 2000; Chien et al., 2014; French et al., 2018). For the subjective character of trust, subjective scales constitute a direct measurement of the construct. Thus, while subjective scales can be reasonably complemented with objective measures of reliance behavior (see chapter 4.2.6 for a more detailed discussion), repeated application of an economic subjective trust measure is essential in studies investigating the formation and development of trust in automation. As at this point, the discussion of how trust in automation can be optimally measured is ongoing (e.g., Billings et al., 2012; French et al., 2018), derivations based on the proposed framework are discussed in chapter 4.2.6.

According to French et al. (2018), the trust in automation scale (ATS) is the currently most widely used subjective trust scale in human-machine interaction research. Its predictive power for reliance behavior has been repeatedly demonstrated (e.g., Chancey et al., 2015; Wang et al., 2009, 2011). Building on the work by Jian et al. (2000), the German version of the Short Learned Trust in Automation Scale (LETRAS-G) is introduced. The LETRAS-G builds on the ATS items as a starting point for scale construction. The goal of this scale is to measure learned trust in automation as a unidimensional attitude in an economical way before and during the interaction with an automated system (initial and dynamic learned trust at stages 2 and 3). Therefore, to ensure content validity of the items and to reflect the unidimensional nature of trust, all items either reflecting antecedents or consequences of trust (e.g., familiarity with the system) were excluded (except for item T5 which reflects a general reliability assessment, Table A1). Furthermore, to avoid endangering unambiguous understanding and content validity by an unnatural word-by-word translation, the core meaning of the English items was translated as naturally as possible into German. Therefore, for some of the remaining original English items, two German ones were constructed, and the other way round for two overlapping English items, one German equivalent was used. The content validity of the items was established by backward translations. The suggested answering format for the questionnaire is a 7-point Likert scale with two opposing ends ("totally disagree" and "totally agree" with a neutral center). Altogether, eight items were constructed. In Table A1 the correspondence of the original English and the German items is depicted along with results from three online studies from the domain of automated driving, establishing initial support for construct validity of the scale. All studies were conducted in the setting of automated driving and participants rated their trust in a system after the presentation of either one or several animated pictures or videos of a driving situation with the system from the driver's ego perspective. For all studies, exemplary, the data from the first trust measurement after rating such a stimulus or series of stimuli is reported in Table A1. For all studies, results of a reliability analysis and factor analyses (principal component analysis) are reported. In the first study (study A), a scale reliability of $\alpha = .900$ was found. For one item (I0: "*Das System versucht mich in die Irre zu führen.")*, the resulting corrected item-total correlation and the factor loading on the general factor were considerably lower than for the other items (see Table A1). As it was in a range indicating a low correlation with the scale, this item was dropped from the scale. The inspected item statistics indicated comparable quality of item I5 as the remaining items and it was therefore kept in the scale. Taken together, this resulted in a scale with seven items reflecting learned trust in automation. This 7-

item measure for learned trust in automation showed a reliability of α = .921 with good item-total correlations and high factorial loadings on a general factor of all items. The general factor explained 61.01% of the total variance. In studies B and C, similar scale reliabilities (α = .923 and α = .941) and item characteristics were found, in summary, providing provisory support for the quality of measurement of the LETRAS-G. All factor analyses supported a unidimensional structure as clearly indicated by both the Kaiser criterion and an inspection of Scree plots, with the general factor explaining 70.33% in study B and 74.91% of the total variance in study C. On the basis of factorial loadings, five items were selected to form a further shortened version of the scale to allow for an even more economic assessment of learned trust in automation. The according items are marked with an asterisk in Table A1. The scale reliabilities of the short version in the three studies were in a range between α = .897 for study A and α = .941 for study C. The correlations between the short and long scale ranged between r = .981 - .985 (all p < .01), indicating that the short scale is a good approximation of the long scale.

Table A1

Reliabilities and item statistics of the German version of the Short Learned Trust in Automation Scale (LETRAS-G) from three online studies. Relations between original ATS (Jian et al., 2000) and LETRAS-G items are depicted.

Original English Version (Jian et al., 2000)	LETRAS-G items	Study A (N = 166)		Study B (N = 276)		Study C (N = 149)	
		α = .921[a]		α = .923		α = .941	
		IC	FL	IC	FL	IC	FL
The system is deceptive	T0: Das System versucht mich in die Irre zu führen.[b] (*excluded*)	.225	.227	-	-	-	-
The system behaves in an underhanded manner	*Not reflecting general trust*	-	-	-	-	-	-
I am suspicious of the system's intent action or output	I1: Ich misstraue den Vorschlägen des Systems.[b]	.708	.776	.720	.793	.729	.799
	I2: Ich bin dem System gegenüber misstrauisch.[b]*	.770	.839	.816	.865	.856	.894
The system's action will have a harmful or injourious outcome	I3: Die Nutzung des Systems wird zu gefährlichen oder schädlichen Konsequenzen führen.[b]*	.733	.807	.753	.820	.759	.822
I can trust the system	I4: Ich vertraue dem System.*	.854	.907	.841	.896	.894	.930
I am confident in the system	I6: Das System ist vertrauenswürdig.*	.793	.866	.830	.894	.890	.931
The system is dependable The system is reliable.	I5: Das System ist zuverlässig.*	.662	.755	.797	.870	.827	.884
I am wary of the system	I7: Ich bin skeptisch gegenüber dem System.[b]	.762	.838	.641	.718	.720	.787
The system provides security	*Not reflecting general trust*	-	-	-	-	-	-
The system has integrity	*Not reflecting general trust*	-	-	-	-	-	-
I am familiar with the system.	*Not reflecting general trust*	-	-	-	-	-	-

Note. Study B and C represent the data of studies 1 and 2 of this dissertation; IC = Corrected Item-Total Correlation; FL = Factor Loading (General Factor); Asterisks mark items of the short version.
[a] Cronbach's α for the 7-item version of the LETRAS-G
[b] Inverted items

Appendix B

Research Articles

Article I: Kraus, J., Scholz, D., & Baumann, M. (2020). What's driving me? – Exploration and validation of a hierarchical personality model for trust in automated driving. *Human Factors*. Advance online publication. https://doi.org/10.1177/0018720820922653

Article II: Kraus, J., Scholz, D., Messner, E.-M., Messner, M., & Baumann, M. (2020). Scared to trust? – Predicting trust in highly automated driving by depressiveness, negative self-evaluations and state anxiety. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02917

Article III: Kraus, J. M., Forster, Y., Hergeth, S., & Baumann, M. (2019). Two routes to trust calibration: Effects of reliability and brand information on trust in automation. *International Journal of Mobile Human Computer Interaction, 11*(3), 1-17. https://doi.org/10.4018/IJMHCI.2019070101

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

Article I: What's Driving Me? – Exploration and Validation of a Hierarchical Personality Model for Trust in Automated Driving

Kraus, J., Scholz, D., & Baumann, M. (2020). What's driving me? - Exploration and validation of a hierarchical personality model for trust in automated driving. *Human Factors*. Advance online publication. https://doi.org/10.1177/0018720820922653

Copyright: ©2020 Human Factors and Ergonomics Society

Check for updates

What's Driving Me? Exploration and Validation of a Hierarchical Personality Model for Trust in Automated Driving

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

Objective: This paper presents a comprehensive investigation of personality traits related to trust in automated vehicles. A hierarchical personality model based on Mowen's (2000) 3M model is explored in a first and replicated in a second study.

Background: Trust in automation is established in a complex psychological process involving user-, system- and situation-related variables. In this process, personality traits have been viewed as an important source of variance.

Method: Dispositional variables on three levels were included in an exploratory, hierarchical personality model (full model) of dynamic learned trust in automation, which was refined on the basis of structural equation modeling carried out in Study 1 (final model). Study 2 replicated the final model in an independent sample.

Results: In both studies, the personality model showed a good fit and explained a large proportion of variance in trust in automation. The combined evidence supports the role of extraversion, neuroticism, and self-esteem at the elemental level; affinity for technology and dispositional interpersonal trust at the situational level; and propensity to trust in automation and a priori acceptability of automated driving at the surface level in the prediction of trust in automation.

Conclusion: Findings confirm that personality plays a substantial role in trust formation and provide evidence of the involvement of user dispositions not previously investigated in relation to trust in automation: self-esteem, dispositional interpersonal trust, and affinity for technology.

Application: Implications for personalization of information campaigns, driver training, and user interfaces for trust calibration in automated driving are discussed.

Keywords: trust in automation, human-automation interaction, trust formation, personality, affinity for technology, structural equation modeling

Address correspondence to Johannes Kraus, Universitat Ulm Fakultät für Naturwissenschaften, Albert-Einstein-Allee 41, Ulm, Baden-Württemberg, 89081, Germany; e-mail: johannes.kraus@uni-ulm.de

HUMAN FACTORS
Vol. 00, No. 0, Month XXXX, pp. 1
DOI:10.1177/0018720820922653
Article reuse guidelines: sagepub.com/journals-permissions
Copyright © 2020, Human Factors and Ergonomics Society.

INTRODUCTION

As efforts to get automated cars on the road increase, human factors research focuses more and more on interaction with such vehicles. One of the most important psychological variables in this regard is trust in automation. Trust in automation has been defined as "the attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p. 51). Following this definition, trust is a subjective assessment of the trustworthiness of an automated technological system and determines to a substantial degree whether and in what manner an automated system is used (e.g., Hoff & Bashir, 2015). Accordingly, the individual level of trust in automation has been found to be a subjective predictor for reliance and behavior in the interaction with various automated systems (e.g., Biros et al., 2004; Lee & Moray, 1992; Lewandowsky et al., 2000; Muir & Moray, 1996; Sanchez et al., 2004) and specifically with automated driving systems (e.g., Hergeth et al., 2016; Payre et al., 2016; Rajaonah et al., 2006).

If automation is not trusted to an appropriate degree the dangers and disadvantages of disuse (using the automation less than its capabilities allow) and misuse (excessive use of the automation) are increased (Parasuraman & Riley, 1997). A scientific understanding of trust formation processes is essential in order to limit the negative effects of such maladaptive usage patterns. In this regard facilitation of a calibrated level of trust—a situation in which a person's trust reflects the capabilities and performance of the automation (e.g., Muir, 1987)—is an important goal as it minimizes misuse and disuse.

There is an ongoing discussion about how trust is dynamically built up prior and during

interactions with automated systems and which factors play a role in this process (e.g., Hoff & Bashir, 2015; Kraus, Scholz, et al., 2019; Lee & See, 2004). Although recent theoretical publications have emphasized the importance of user personality as a source of variance in trust development (e.g., Hoff & Bashir, 2015; Schaefer et al., 2016) there have been few empirical studies (e.g., Schaefer et al., 2016). An understanding of the role of individual differences in trust formation would provide essential insight into the psychological mechanisms of trust calibration and could be used to personalize automated technology. This research investigates a hierarchical personality model of dynamic learned trust in highly automated driving integrating a variety of personality traits. The model's theoretical foundation is Mowen's 3M model (2000), which posits that behavior is influenced by a hierarchy of personality factors, ranging from broad to specific traits. In a first study, the hypothetical personality model is explored and reduced to a set of dispositional variables and mediation paths providing an economic prediction of trust in highly automated driving (final model). This economic final personality model was replicated and confirmed in a second independent study.

Personality and Trust

This research builds on the definition of personality as a "set of psychological traits and mechanisms within the individual that are organized and relatively enduring and that influence his or her interactions with […] environments" (Larsen & Buss, 2005, p. 4). It follows from this definition that personality influences information processing and thus interactions with automated systems. Hence personality has been treated as a predictor of trust in automation in both theory and empirical research. Studies, reviews, and meta-analyses of factors affecting trust in technology (e.g., robots, automated driving) have emphasized the role of personality variables in interindividual trust differences (e.g., Hancock et al., 2011; Hoff & Bashir, 2015; Kraus et al., 2020; Körber & Bengler, 2014; Schaefer et al., 2016). Taken together, personality traits are hypothesized to affect trust formation. Their influence might be especially strong if the reliability of an automated system is questionable, as it is likely in early interactions with a new system when trust calibration is still unstable.

Trust is established in a dynamic psychological process (e.g., Lee & See, 2004). As the character of trust changes along the dynamic process of trust formation, to understand how personality affects trust, several different trust concepts have to be considered. A first differentiation can be made between *dispositional* and *learned trust* (e.g., Hoff & Bashir, 2015): dispositional trust exists prior to any kind of knowledge or interaction with a specific system, whereas learned trust emerges from experience with a particular system. Thus, dispositional trust has been characterized as a stable personality trait reflecting "a diffuse expectation of others' trustworthiness […] based on early trust-related experiences" (Merritt & Ilgen, 2008, p. 195) that determines the initial level of trust in another entity at a first encounter. Dispositional trust can be broken down into the two traits *dispositional interpersonal trust* (e.g., Rotter, 1967, 1980; Gurtman, 1992; also referred to as propensity to trust elsewhere) and *propensity to trust in automation* (e.g., dispositional trust in Hoff & Bashir, 2015). Dispositional interpersonal trust was defined by Rotter (1967) as an individual's expectation that he or she can rely on a promise or written communication from another human being. Hoff and Bashir (2015, p. 413) defined the propensity to trust in automation as "an individual's overall tendency to trust automation, independent of context or a specific system." Both forms of trust are conceptualized as traits, albeit with different foci and specificity. Dispositional interpersonal trust—based on the provided definition—is assumed to be more general and thus affecting more specific forms of trust (e.g., propensity to trust in automation).

Learned trust is history-based (Merritt & Ilgen, 2008), system-specific, and "represents an operator's evaluations of a system drawn from past experience or the current interaction […] directly influenced by the operator's pre-existing knowledge and the automated system's performance" (Hoff & Bashir, 2015, p. 420).

Initial learned trust is an evaluation of the system based on knowledge prior to system use, and *dynamic learned trust* is based on the sum of interactions with a specific system.

Hoff and Bashir's (2015) model posits that dispositional trust is established through the combined influence of sociocultural variables and personality traits. Dispositional trust, along with the characteristics of the situation and system variables (such as performance, reliability, and design features), influences the emergence of learned trust. To capture situational influences on trust Hoff and Bashir (2015) also propose the concept of *situational trust*, which is not in the scope of this research.

Although the relationship between dispositional trust and learned trust has been considered on a theoretical level (e.g., Hoff & Bashir, 2015), empirical evidence is scarce (e.g., Merritt & Ilgen, 2008; Merritt et al., 2013) and at present the role of dispositional trust in the development of learned trust in a specific automated system remains unclear. Furthermore, it is an open question how the propensity to trust in automation can be explained in terms of broader personality traits with a more general theoretical foundation. Two of the hypotheses of this research predict that dynamic learned trust is influenced by dispositional interpersonal trust and the propensity to trust in automation, amongst other factors.

Research on the Relation of Personality to Trust in Automation

To date, few studies have investigated the relationships between personality traits, other user dispositions, and trust in automation and related outcomes (see below; e.g., intention to use; Hoff & Bashir, 2015). In terms of general traits, mainly the Big Five and locus of control (LOC) have been investigated in relation to trust in automation. Regarding more specific user dispositions most frequently dispositional interpersonal trust, propensity to trust in automation, and a priori attitude toward automation were included in studies. Although some of the reported relationships have been replicated, in some cases effect sizes and the direction of relationships are somewhat inconsistent (e.g., Schaefer et al., 2016). These inconsistencies might in part result from different study contexts. While some personality traits might affect individual trust levels irrespective of context, the influence of other traits might only be activated by specific situational or task characteristics. This being said, at this point it is unclear which traits show a strong cross-situational relation to learned trust. Furthermore, some nonsignificant or contradicting findings for the investigated traits (e.g., the Big Five) might result from an investigation of direct correlations of general traits with trust in automation ignoring the specificity of traits for trust in automation. As "traits reside in a structure in which more abstract, cross-situational traits influence narrower situation-specific behavioral tendencies, which in turn influence behavior" (Mowen et al., 2007, p. 591), neglect of different trait specificity is associated to the so-called *bandwidth-fidelity dilemma* (Cronbach & Gleser, 1965) broadly discussed in IO (e.g., Ones & Viswesvaran, 1996) and consumer psychology (e.g., Mowen, 2000). The bandwidth-fidelity dilemma consists of two opposed problems. On one hand, while specific personality constructs provide a good explanation of variance in a specific behavior, their theoretical foundation tends to be superficial and different constructs often show very high conceptual and statistical overlap (shared variance). On the other hand, broader and theoretically well-established personality constructs like the Big Five often provide only limited explanation of variance in specific behavior.

Hierarchical personality models provide a solution to this dilemma by integrating traits at different levels of specificity for the explanation and prediction of specific phenomena. This research builds on the theory of hierarchical personality models (e.g., Allport, 1961; Mowen, 2000) to integrate and extend on the current state of research on personality factors in trust in automation and investigates a hierarchical personality model for dynamic learned trust in automation.

Hierarchical Personality Model: The 3M Model

Hierarchical modeling of personality traits enables the prediction of specific

outcomes—such as trust in automation—from a combination of broad and specific traits, thereby reducing the disadvantages of both (Mowen, 2000). At the same time hierarchical modeling of personality traits allows for the structured analysis of traits along the nomological network of an outcome of interest. Hierarchical personality models such as the metatheoretic model of motivation and personality (3M model; Mowen, 2000) were successfully applied in studies in other domains of consumer behavior (e.g., Bosnjak et al., 2007; Schneider & Vogt, 2012).

This research applies the 3M model for an integrative investigation of personality traits that have been related to trust in automation and additional, theoretically relevant traits. The 3M model posits that behaviors of interest are directly or indirectly influenced by "a combination of traits from the different levels of the hierarchy" (Mowen et al., 2007, p. 592). It proposes a hierarchy of four levels of personality traits ordered according to their specificity for a behavior. The bottom level of the trait hierarchy consists of *elemental traits*, broad behavioral predispositions embedded in genetic heritage (e.g., the Big Five; Saucier, 1994). On the second level the model orders *compound traits*, cross-situational, one-dimensional behavioral dispositions resulting from the interaction of elemental traits, individual learning history, and culture (e.g., task orientation). On the third level *situational traits* are placed, which are stable behavioral tendencies with regard to a broader general situational context (e.g., shopping enjoyment; Mowen et al., 2007). On the fourth level *surface traits*, which result from the interplay of situational traits, attitudes, and personal involvement, reflect behavioral tendencies that apply to specific products or consumer domains. By applying the hierarchical personality theory as a framework to the domain of automation trust and using structural equation modeling (SEM) for analyzing the relationships of the included variables, an integrative understanding of the role of single personality traits can be provided along unique and shared variance proportions, intercorrelations, as well as mediations.

Development of the Hierarchical Personality Model for Trust in Automation

The application of a hierarchical personality model involves the decisions of (a) which traits to include and (b) at which level to order them. This is especially challenging as a multitude of personality constructs exist, that in part are very similar to each other (both in terms of the specific psychological mechanism they refer to and statistically in terms of shared variance proportions; e.g., Watson & Clark, 1984). This decision is not trivial as a balanced number of constructs at every level of a model increases a model's predictive power for the modeled outcome (balance between predictivity and statistical power). For the decision of which traits to include in the model (model hypotheses, see Table 1), we carefully reviewed available studies on the relationships of personality traits and (a) trust in automation and (b) related phenomena (interaction with technology). Additionally, we included further personality variables that have not yet been investigated in relation to trust in automation or related outcomes before (self-esteem, technological self-efficacy, and affinity for technology). For these variables, hypotheses were derived from former psychological research (e.g., on dispositional interpersonal trust). For the decision of at which level to order the modeled personality traits, we followed the definitions of trait levels by Mowen (2000; see earlier).

The proposed model is composed of three trait levels (elemental, situational, and surface traits) predicting the outcome "dynamic learned trust in automation." To maximize model economy, the levels of elemental and compound traits were combined to one basic trait level (which is referred to as "elemental traits"). The following personality traits were included in the proposed hierarchical personality model for dynamic learned trust in automated vehicles (hypothesized full model; see Figure 1): the Big Five, LOC, self-esteem, affinity for technology, technological self-efficacy, dispositional interpersonal trust, and propensity to trust in automation. Personality and attitudes share some charcteristics, besides others in terms of their explanation of individual variances in behavior

TABLE 1: Hypothesized Relationships of the Variables in the Investigated Hierarchical Personality Model

Hypotheses	
Hypothesis 1	*Extraversion is positively related to trust in an automated driving system. This relationship is mediated by the modeled traits at the situational and surface levels.*
Hypothesis 2	*Neuroticism is negatively related to trust in an automated driving system. This relationship is mediated by the modeled traits at the situational and surface levels.*
Hypothesis 3	*Agreeableness is positively related to trust in an automated driving system. This relationship is mediated by the modeled traits at the situational and surface levels.*
Hypothesis 4	*Locus of control is related to trust in an automated driving system. This relationship is mediated by the modeled traits at the situational and surface levels.*
Hypothesis 5	*Self-esteem is positively related to trust in an automated driving system. This relationship is mediated by the modeled traits at the situational and surface levels.*
Hypothesis 6	*Dispositional interpersonal trust is positively related to trust in an automated driving system. This relationship is mediated by the modeled traits at the surface level.*
Hypothesis 7	*Affinity for technology is positively related to trust in an automated driving system. This relationship is mediated by the modeled traits at the surface level.*
Hypothesis 8	*Technological self-efficacy is positively related to trust in an automated driving system. This relationship is mediated by the modeled traits at the surface level.*
Hypothesis 9	*The propensity to trust in automation is positively related to trust in an automated driving system.*
Hypothesis 10	*A priori acceptability of automated driving is positively related to trust in an automated driving system.*

by dispositional differences (e.g., Ajzen, 2005). As for this similarity and for the essential personality influences in attitude formation (e.g., Albaraccín et al., 2018), we also included a technology attitude at the most specific model level: the a priori acceptability of automated driving. For the foundation of this research in the 3M model, in the following we refer to the model levels as "trait levels." It is hypothesized that the effects from the elemental level traits to trust in automation are mediated by (one or more) situational and surface traits. In line with MacKinnon et al. (2007) such a concurrent modeling of multiple mediations accounts more appropriately for the nature of many mediations in psychology. As for the explorative nature of our approach, we did not hypothesize specific paths for the mediations. Rather it is an essential part of the research question of this study to explore the relative importance of the many theoretically derivable mediations of the intervening model levels. The hypothesized model displayed in Figure 1 thus represents a starting point in analyses in the presented studies. Next, the inclusion of the modeled variables is theoretically and empirically supported in detail.

Literature Overview and Model Hypotheses

The Five-Factor Model of Personality (e.g., Costa & McCrae, 1992; McCrae & John, 1992; Saucier, 1994) has been investigated in relation to trust in automation and related outcomes in the past (e.g., Hoff & Bashir, 2015; Szalma & Taylor, 2011). While for some factors, several studies found relationships with trust, for others evidence is scarce or ambiguous. Extraversion loads highest on adjectives like sociable, fun loving, affectionate, and talkative (McCrae & Costa, 1987). Extraverted people tend to be more open to others and to exciting experiences and so one would expect them to be more open to new, potentially risky technology. In line with

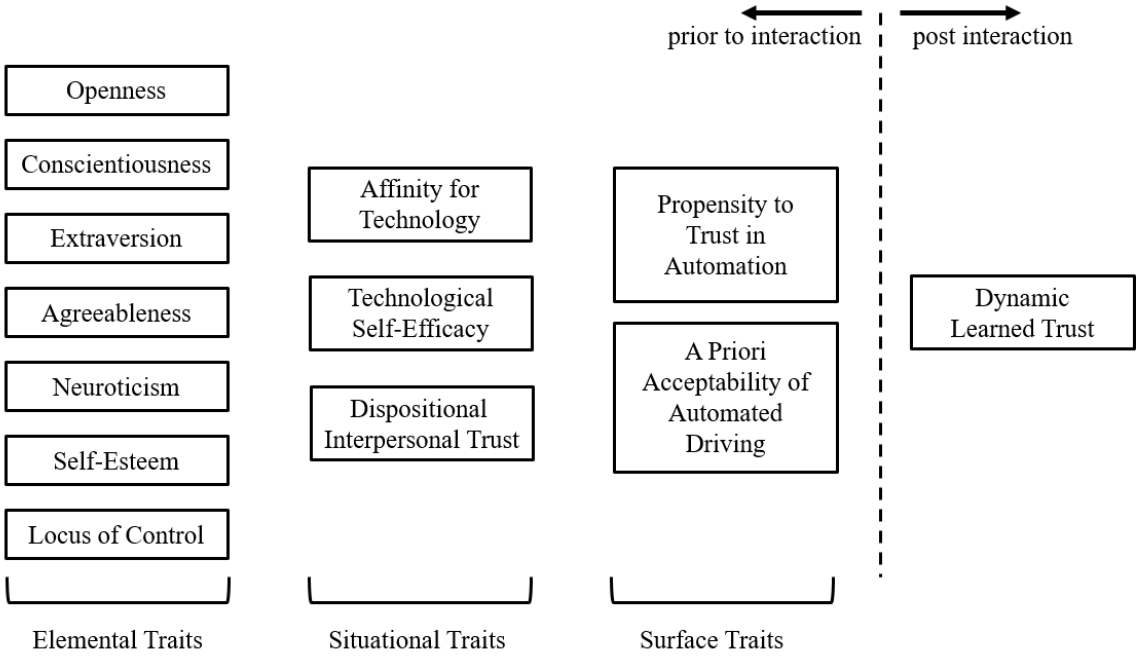


Figure 1. Included variables at the different levels of the hypothesized (full) hierarchical personality model based on the 3M model (Mowen, 2000).

their assumption that extraverts have a higher tendency to trust in interpersonal interactions, Merritt and Ilgen (2008) found extraversion to be positively related to the propensity to trust in machines. Furthermore, personality facets reflecting extraversion have been found to be positively related to dispositional interpersonal trust (e.g., Evans & Revelle, 2008; Gaines et al., 1997). Accordingly, it is hypothesized that

> *Hypothesis 1 (H1): Extraversion is positively related to trust in an automated driving system. This relationship is mediated by the modeled traits at the situational and surface levels.*

Neuroticism—best described by the adjectives worrying, nervous, high-strung, and insecure (McCrae & Costa, 1987)—was found to be related to increased threat sensitivity (Reynaud et al., 2012), a more negative reaction to uncertainty (e.g., Berenbaum et al., 2008), and higher risk aversion (Oehler & Wedlich, 2018). Accordingly, Evans and Revelle (2008) found a negative relationship between neuroticism and dispositional interpersonal trust. From this, it can be concluded that people with a higher level of neuroticism should trust automated driving systems less as their use is associated with both uncertainty and risk. In line with this reasoning, Szalma and Taylor (2011) found neuroticism to be negatively correlated with agreement to valid advice of an automated system, providing some evidence that this factor is important for trust in automation. It is therefore hypothesized that

> *Hypothesis 2 (H2): Neuroticism is negatively related to trust in an automated driving system. This relationship is mediated by the modeled traits at the situational and surface levels.*

The interpersonal trait agreeableness is described with adjectives like soft-hearted, forgiving, sympathetic, acquiescent, and trusting (McCrae & Costa, 1987). In line with its theoretical conception, agreeableness was found to be positively related to dispositional interpersonal trust (e.g., Evans & Revelle, 2008) and trusting behavior in interpersonal contexts (see Thielmann & Hilbig, 2015, for an overview). Accordingly, agreeable persons should tend to trust automated technology more, while people with lower agreeableness should have a

tendency to distrust automation. In line with this, Chien et al. (2016) found agreeableness to be positively related to the propensity to trust in automation. It is therefore hypothesized that

Hypothesis 3 (H3): Agreeableness is positively related to trust in an automated driving system. This relationship is mediated by the modeled traits at the situational and surface levels.

Due to limited existing empirical evidence for a relation of the remaining two Big Five traits, consciousness and openness to trust in automation, these traits were exploratorily investigated in this research without the formulation of directed hypotheses.

Besides the Big Five, LOC was also hypothesized to be related to trust in automated technology and its use before (e.g., Payre et al., 2014). LOC describes a person's perception of the controllability, causation, and outcome of an event or action (e.g., Ajzen, 2002; Rotter, 1966, 1990). An internal LOC is associated with a perceived ability to affect outcomes through one's actions, whereas people with an external LOC perceive that outcomes are determined by factors beyond their control. Thus, it might be that people with internal LOC trust an automated system less and are more likely object to ceding control to it as they view themselves in control of the outcome of the driving task, while persons with an external LOC on the other hand may feel liberated if automation relieves them from a potentially dangerous task they cannot fully control and thus trust such a system more. In support of this, externals were shown to have a higher increase in trust in a lane departure warning system than internals (Rudin-Brown & Noy, 2002), and external LOC was found to predict the intention to use automated vehicles in the future to some extent (Choi & Ji, 2015). On the other hand, there is also reason to assume that people with internal LOC are more likely to trust automated technology. As internals try to influence their fate to a greater extent, are more determined (Galvin et al., 2018), and are better problem solvers in ambiguous situations (e.g., DuCette & Wolk, 1973) they might also tend to perceive themselves as able to make functional decisions about the use and nonuse of the system. In contrast, an external LOC has been associated with learned helplessness (e.g., Hiroto, 1974) and thus externals might perceive themselves as less in control of functionally using the system and thus trust it less. In face of these contradicting arguments and the conflicting results for the role of LOC in driving (Özkan et al., 2005) a relationship between LOC and trust in automation is hypothesized, but the direction of this relationship is not defined in the hypothesis and is an open question of this research:

Hypothesis 4 (H4): Locus of control is related to trust in an automated driving system. This relationship is mediated by the modeled traits at the situational and surface levels.

Interpersonal trust has been linked theoretically to automation trust (e.g., Lee & Moray, 1992; Lee & See, 2004; Madhavan & Wiegmann, 2007; Muir, 1994; Parasuraman, de Visser et al., 2014), but to date the only empirical evidence that dispositional interpersonal trust might be related to trust in automation comes from other technological domains and an investigation of the explanatory power of dispositional interpersonal trust, for trust in automated technology has not been reported yet (to our knowledge). As dispositional interpersonal trust might constitute a general dispositional antecedent of more specific forms of trust, it is hypothesized that

Hypothesis 6 (H6): Dispositional interpersonal trust is positively related to trust in an automated driving system. This relationship is mediated by the modeled traits at the surface level.

For the propensity to trust in automation more empirical studies have been conducted. Merritt et al. (2013) reported higher trust toward an automated weapon detector in people with a higher propensity to trust in automation. In the same task setting, Merritt and Ilgen (2008) found higher propensity to trust in automation to be related to higher pretask trust (and

posttask trust provided the system performed well). At this point, the role of propensity to trust in automation in the emergence of learned trust in an automated driving system has yet to be explored. It is hypothesized that

Hypothesis 9 (H9): The propensity to trust in automation is positively related to trust in an automated driving system.

Also, the a priori attitude toward automated driving systems has been investigated in relation to trust in automation (e.g., Merritt et al., 2013). The acceptability of technology and the a priori attitude to technology have been found to be strongly correlated (Payre et al., 2014) and thus there seems to be a restricted theoretical foundation for distinguishing between these two. In derivation from the definition of the a priori acceptability of technology by Payre et al. (2014) the a priori acceptability of automated driving can be defined as an individual's evaluation of the technology of automated driving and its potential uses before any interaction with a specific system has taken place. The Technology Acceptance Model (Davis et al., 1989) proposes that attitudes predict the intention to use a given technology. Payre et al. (2014) found that the acceptability of automated vehicles and the attitude to them predicted the intention to use an automated vehicle. Merritt et al. (2013) found that implicit attitudes to an automated system predicted trust in that system after interaction with it. At present, it is unclear how a priori attitudes to automated driving influence trust in a specific automated vehicle. Based on the extant evidence it is hypothesized that

Hypothesis 10 (H10): A priori acceptability of automated driving is positively related to trust in an automated driving system.

Additional to these variables that have already been investigated in relation to trust in automation, we included self-esteem (Rosenberg, 1965), technological self-efficacy (Bandura, 1991; McDonald & Siegall, 1992), and affinity for technology (Franke et al., 2019) in the model.

Given that there have been many studies concerning the relationship between self-confidence, trust, and reliance in technological systems (e.g., de Vries et al., 2003; Lee & Moray, 1994, 1992; Moray et al., 2000) it is promising to investigate a broader array of personality traits related to self-evaluation, particularly as there is some debate about how trust and self-confidence interact with respect to the decision to use an automated system (Schaefer et al., 2014). Self-confidence in relation to human–machine interaction describes a situational self-assessment of one's capability to perform a task correctly (e.g., Briggs et al., 1998; Moray et al., 2000). To understand the emergence of this assessment and its relationship to trust in automation, a closer look at the role of *core self-evaluations* seems fruitful. Core self-evaluations include self-esteem and self-efficacy as well as neuroticism and LOC (Judge & Bono, 2001; Judge et al., 1998). Self-esteem has been defined as "the level of global regard one has for the self as a person" (Harter, 1993, p. 88) and reflects the extent to which people view themselves as significant, successful, worthy, and capable (Coopersmith, 1981, cited in Herz & Gullone, 1999). Self-esteem has been closely related to self-confidence both on a theoretical and an empirical level (e.g., Owens, 1993). Similarly, self-efficacy has been discussed as closely connected to self-confidence (e.g., Schunk, 1991). Bandura (1991, p. 257) defined general self-efficacy as "people's beliefs about their capabilities to exercise control over their own level of functioning and over events that affect their lives." Also, more specific forms of self-efficacy for the technological domain have been proposed. For example, *computer self-efficacy* as an individual's judgment of his or her ability to use computers has been repeatedly investigated (Dishaw et al., 2002; Madhavan & Phillips, 2010). Extending the work-related definition of *technological self-efficacy* (McDonald & Siegall, 1992) to general technology interaction, we define it as an individual's belief in his or her ability to use technological systems successfully and make decisions during interactions with such systems in a functional manner. One would expect people with relatively high self-esteem and technological self-efficacy to

consider themselves as capable of understanding and using an automated system in a safe and efficient manner. Hence it is hypothesized that

Hypotheses 5 and 8 (H5 and H8): Self-esteem (H5) and technological self-efficacy (H8) are positively related to trust in an automated driving system. These relationships are mediated by the modeled traits at the situational and surface levels.

As a third situational trait, the affinity for technology (Karrer et al., 2009) has been included in the model. Karrer et al. (2009) defined affinity for technology as a person's positive attitude to, enthusiasm for, and trust in technology and related it to interest in and acceptance thereof. Franke et al. (2019) rooted the closely related affinity for technology interaction in *need for cognition* (Cacioppo & Petty, 1982) and defined it as "the tendency to actively engage in intensive technology interaction" and propose it to be "a key personal resource for coping with technology" (p. 456). They reported correlations of affinity for technology with technology enthusiasm, control beliefs, technical system solving, and learning success. One would expect people with a relatively high affinity for technology to approach new technology more optimistically and accordingly put greater weight to potential benefits than potential risks when assessing it. Hence it is hypothesized that

Hypothesis 7 (H7): Affinity for technology is positively related to trust in an automated driving system. This relationship is mediated by the modeled traits at the surface level.

Altogether, the hypothesized model encompasses 10 hypotheses about dispositional variables for the explanation of trust in automation ordered at three levels (Table 1; Figure 1). Regarding the decision at which level to order the traits, the modeled elemental traits can be differentiated from the other trait levels easily for their broad and general character. The differentiation between situational and surface traits is based on the definitions of Mowen (2000) along with their specificity. While situational traits are predispositions to behave consistently in general situational contexts, the surface traits refer to specific consumer domains (products or consumer behavior). Affinity for technology and technology self-efficacy are related to general assessments of technology as a broader category and are therefore ordered at the situational trait level. On the other hand, the surface traits propensity to trust in automation and a priori acceptability of automated driving are more specific for the consumer domain "automated vehicles."

Overall Method

This research consists of two successive studies exploring the proposed hierarchical personality model for trust in automated driving. In Study 1, this hypothesized full model is exploratorily tested and reduced to allow for a more economical prediction of the outcome (economic final model). In Study 2 the final model is replicated in an independent sample. This research complied with the tenets of the Declaration of Helsinki. Informed consent was obtained from each participant.

STUDY 1: EXPLORING AN ECONOMIC HIERARCHICAL PERSONALITY MODEL

Method

This online study followed a correlative design investigating the relationships between personality traits, the prior acceptability of automated driving, and trust in automation. Participants first answered personality questionnaires and were then introduced to a highly automated vehicle, before watching a series of animated videos with this vehicle. After each video and at the end of the experimental trials trust was assessed. While the study also included experimental manipulations with six study groups (three levels of system transparency and two levels of system reliability), for this research data analysis was conducted in the complete sample and only the trust measurement directly after the last video was used (see precondition tests later).

Sample

Participants were recruited by mailings at Ulm University and through social media (e.g.,

Facebook). Only adults who had held a driving license for at least 1 year were eligible. Participants received 0.75-hr course credit or were entered into a prize draw for shopping vouchers with a value of €10 each. A total of 286 participants completed the study. To ensure valid answering of questionnaires we inspected the data set carefully for answer patterns and extreme response biases. Ten participants were excluded for one of the following reasons: (a) technical issues; (b) no variation in responses; (c) over-rapid completion (below 40% of the median duration; for example, discussed in Greszki et al., 2014); (d) slow completion indicating technical problems (over 2 hr). Two further participants were identified as multivariate outliers (as discussed later). The final sample comprised 274 participants, who needed 46 min and 37 s on average to complete the study. The sample consisted of 159 women (58%), 113 men, one other, and one "not specified." The average age was 33.88 years (SD = 14.75). Participants had a driving license on average since M = 17.20 years (SD = 28.83); 71.17% answered to never having used automated driving assistance before and the average experience with advanced driver-assistance systems (ADASs) was considerably small (M = 2.09; SD = 1.63, 7-point Likert scale with 1 = "very inexperienced" and 7 = "very experienced").

Procedure

The online survey was administered by EFS Survey (Questback GmbH) in July 2016. After reading the study procedure and an informed consent form, participants completed the personality questionnaires. After this, they were shown general information on ADASs and the system that was the focus of this study, the "motorway pilot." They then watched seven animated videos of the ADAS, answering a questionnaire after each. After watching the videos, participants filled out questionnaires on demographic variables, driving experience, and experience of ADASs.

Study Material

The first part of the introductory presentation consisted of general information about ADAS, informing participants about the increased availability of these systems. The lane-keeping assistant was described, as an example, and illustrated with a picture. It was also explained that manufacturers are currently working on systems that will be able to take over the driving task completely. Autonomous driving was defined as a situation in which an automated system drives sections of a route entirely independently. The second part introduced the motorway pilot, describing it as a prototype that was currently undergoing road testing by professional test drivers. Participants were informed that the study was investigating how drivers perceive the motorway pilot. They were told that the system could take over lateral and longitudinal control on the motorway, in both the deceleration and acceleration lanes, whilst complying with traffic rules. It was explained that the system uses radar and optical sensors to build up an image of the environment and to carry out a dynamic situational analysis. It was stated that this allows for (temporal) autonomous driving with maneuver planning (e.g., overtaking and lane changing). A page with manufacturer information about possible malfunctions was also presented, noting that malfunctions lead to the activation of a "safe mode" in which the vehicle drives automatically to the service lane, where it stops completely and activates the warning lights. After the safe mode is activated the driver has to manually maneuver the vehicle on the motorway and can only then activate the automation again. As the goal was to frame a malfunction as to be resulting in inconvenience and a loss of time rather than an accident and physical harm (such a system might be perceived as unacceptable for road use) and to prevent too much insecurity in the participants they were instructed that the safe mode does not constitute a safety threat as it is controlled by additional close-range sensors. For the transparency manipulation, information about system malfunctions and their causes was manipulated in three levels with increasing level of detail: (1) malfunctions might occur, (2) additional: reasons for malfunction, (3) additional: technological explanation for malfunction.

Video stimuli were created with the VICOM editor (TÜV|DEKRA arge tp 21,



Figure 2. Screenshot of the video stimuli used in Studies 1 and 2. The videos were created with the support of www.vicomeditor.de, a product of TÜV|DEKRA arge tp 21.

2011; Figure 2). In a pilot study (N = 36) 8 videos were selected from a total of 26, based on similar ratings in regard to driving style (risk, comfort, usefulness) and probability to use the automation. All video clips lasted 30 s, showed similar numbers of other road users, and included at least two lane changes and one overtaking maneuver, but the surroundings and traffic varied. All videos were recorded from the perspective of the driver using the motorway pilot. While all other videos showed noncritical situations, in the low-reliability condition, the fourth video showed a system malfunction and the subsequent "safe mode." In the high-reliability condition no malfunction occurred.

Questionnaires

To guarantee motivated study participation and participants' compliance throughout the study, short scales were utilized (Mowen & Voss, 2008). If available, validated short scales with substantial proof of psychometric quality were used. If such validated short scales were not available, existing long scales were shortened on the basis of item content, factor representation, and item statistics. Some scales were translated from English to German using a back-translation procedure. Also, for some scales adapted from other domains of technology, items had to be rephrased to suit the domain of automated driving. All scales were assessed with a 7-point Likert scale ranging from 1 = "not agree at all" to 7 = "totally agree."

The Big Five personality traits were assessed with the BFI-10, a 10-item scale by Rammstedt et al. (2013). The BFI-10 has been constructed on the basis of the BFI-44 and provides an economic measurement of the Big Five complying with psychometric quality standards (Rammstedt & John, 2007). Self-esteem was measured with five items from the Rosenberg Self-Esteem Scale (Rosenberg, 1965), which were selected on the basis of factor loadings reported in Gray-Little et al. (1997) and translated into German. Internal and external LOC were assessed with the IE-4 (Kovaleva, 2012), a validated German short scale with substantial proof of its psychometric quality (Kovaleva et al., 2012). Affinity for technology was assessed using two representative items for each of the four factors of the TA-EG scale (Karrer et al., 2009). Technological self-efficacy was measured with five items of the scale for technology self-efficacy by Holcomb et al. (2004) for which the author reported high factor loadings.

The items were translated into German and they were adapted to refer to "new technology" rather than "computers." Dispositional interpersonal trust was assessed with the German short scale by Beierlein et al. (2012). To measure propensity to trust in automation four items of the "propensity to trust" scale (Merritt et al., 2013; Merritt, 2011) were translated into German and rephrased to refer to "automation technology" instead of "machines." A priori acceptability of automated driving was measured with German translations of items from the scale developed by Payre et al. (2014). The factors "contextualized acceptance of an automated car system" and "interest in using driving automation whilst impaired" were measured using, respectively, two and three items (for the first factor the item referring to passengers and for the second factor the one referring to medication were excluded). Trust in automation was assessed with a shortened, seven-item German version of the scale developed by Jian et al. (2000). The items reflect trust as a unidimensional concept. The Jian et al. (2000) scale has been shown to predict reliance to automated systems (e.g., Wang et al., 2009, 2011; Chancey et al., 2015). Table 2 shows the means, standard deviations, reliabilities, and correlations of the scales that were used. Note that the scales for the assessment of the Big Five and LOC only consisted of two items per factor and thus Cronbach's α does not provide a representational evaluation of reliability as it is biased by test length (number of items; e.g., Cortina, 1993; Starkweather, 2012). Therefore, for these scales the Spearman–Brown coefficient is reported in Table 2. Please find additional proof of scale reliability for the BFI-10 in Rammstedt and John (2007) and for the IE-4 in Kovaleva et al. (2012). The overall trust rating in our sample was on average with $M = 5.14$ ($SD = 1.37$, 7-point Likert scale) in a range indicating a rather positive evaluation of the automated driving system, which is comparable to findings of other research (e.g., Gold et al., 2015).

TABLE 2: Means, Standard Deviation, Internal Consistency (Cronbach's α/Spearman–Brown Coefficient), and Correlations of All Traits and Trust in Automation in Study 1

	M	*SD*	α/ρ	1	2	3	4	5	6	7	8	9	10	11	12	13
1 O*	5.20	1.36	.60	—												
2 C*	4.93	1.13	.33	**.12**	—											
3 E*	4.47	1.36	.69	.11	**.17**	—										
4 A*	4.32	1.09	.14	.05	.04	−.01	—									
5 N*	3.80	1.38	.62	−.06	−.06	**−.24**	.01	—								
6 SE	5.69	0.98	.84	.03	**.22**	**.34**	.08	**−.47**	—							
7 EL*	2.93	1.16	.43	.08	−.07	−.11	−.09	**.24**	**−.36**	—						
8 IL*	5.87	0.85	.62	.06	**.16**	**.23**	.05	**−.26**	**.55**	**−.33**	—					
9 INT	4.96	1.10	.80	.01	.11	**.28**	**.47**	**−.14**	**.36**	**−.28**	**.25**	—				
10 AFT	5.05	0.99	.83	.09	.06	.06	.01	**−.29**	**.32**	−.12	**.26**	**.17**	—			
11 TSE	5.47	1.16	.90	−.00	−.01	−.03	−.12	**−.25**	**.23**	**−.12**	**.14**	−.00	**.63**	—		
12 PRO	4.98	1.23	.91	.02	.03	.10	.09	**−.13**	**.25**	−.08	**.12**	**.27**	**.47**	**.24**	—	
13 ACC	4.47	1.47	.79	−.02	−.03	.05	**.13**	−.08	.11	**−.13**	.02	**.18**	**.36**	**.23**	**.50**	—
14 Trust	5.14	1.37	.95	.10	.05	**.15**	**.12**	**−.14**	**.25**	**−.12**	**.14**	**.27**	**.43**	**.18**	**.60**	**.52**

Notes. Bold indicates a significant correlation ($p < .05$); O = openness; C = conscientiousness; E = extraversion; A = agreeableness; N = neuroticism; SE = self-esteem; EL = external locus of control; IL = internal locus of control; INT = dispositional interpersonal trust; AFT = affinity for technology; TSE = technological self-efficacy; PRO = propensity to trust in automation; ACC = a priori acceptability; for variables with asterisks Spearman–Brown coefficient is reported instead of Cronbach's α.

Statistical Procedure

To test the study hypotheses, *SEM* with scale means (path analysis) was used (note that we use the term "*SEM*" in the following as path analysis represents a special case of *SEM*). Data analysis was conducted with the R package *lavaan* (Rosseel, 2012; version 0.6-2; R version 3.4.3). The model included structural but not measurement equations. In accordance with the recommendations of Preacher and Hayes (2008), the residuals of the mediating variables at each model level were allowed to co-vary to prevent biased standard errors and model misspecification.

The goal of the first study was to explore the relationships between the modeled personality variables. Therefore, model fit was optimized iteratively in a stepwise approach by eliminating paths and variables that did not account for variance in subsequent model levels. Model fit was evaluated with the χ^2 test (Schermelleh-Engel et al., 2003; Ullman, 2013; West et al., 2012), the comparative fit index (CFI; Bentler, 1990), the root mean square error of approximation (RMSEA; Steiger & Lind, 1980; Steiger, 2016), and the standardized root mean square residual (SRMR; Bentler, 1995). The following criteria for good model fit were adopted (Hu & Bentler, 1999; West et al., 2012): CFI > .95, RMSEA < .06, and SRMR < .08.

Direct and specific indirect effects are reported (MacKinnon, 2008). Direct effects capture the relationship between adjacent variables (e.g., elemental → situational) and specific indirect effects provide a statistical test for the hypothesized mediations. Indirect effects were examined using the percentile bootstrap procedure to obtain a good balance between power and type I error rate (e.g., Hayes, 2018). The null hypothesis was rejected if the 95% confidence interval (CI) did not include zero.

STUDY 1: RESULTS

Preconditions

Preconditions for *SEM* were tested. First, there was no missing data in this online study. Second, we checked for univariate outliers using *z*-scores of the scale means (Tabachnick & Fidell, 2013). Eight *z*-scores (0.2%) exceeded 3.29, but as a small number of univariate outliers are acceptable in large samples these cases were retained in the sample. Calculation of the Mahalanobis distance identified two multivariate outliers ($> \chi^2$ (13) = 37.59, $p < .001$), which were excluded from analyses. Third, multivariate normality was assessed by inspecting histograms of scale mean distributions and QQ plots. This assumption was violated, so robust Satorra-Bentler corrected estimates of standard errors and test statistics were used (Satorra & Bentler, 1994). Fourth, to rule out bias due to group effects, we tested for interactions between each personality trait and the manipulated independent variables using a series of general linear models. The only traits involved in interactions were neuroticism, self-esteem, and internal LOC. Thus for Study 1, results on these three traits should be interpreted with the possibility that effects may be moderated by additional factors as reflected in the independent variables. We additionally compared the group values for trust in automation using a 2 × 3 ANOVA to rule out any effects of the manipulations. The ANOVA revealed that the two manipulations (malfunction and transparency) did not affect the criterion variable of Study 1, malfunction: $F(1, 268) = 0.16$, $p = .691$, $\eta_p^2 = .001$; transparency: $F(2, 268) = 0.18$, $p = .836$, $\eta_p^2 = .001$; interaction: $F(2, 268) = 0.46$, $p = .629$, $\eta_p^2 = .003$).

Study 1: Model Analysis

In line with Hayes (2018) as a first step, we tested a full parallel-serial-mediation model, in which all paths were estimated simultaneously, to get an understanding of which traits were essential in explaining variance in the personality model (Table 3). Both surface traits—predisposition to trust in automation and a priori acceptability of automated driving—significantly predicted trust (column 6). Second, the two situational traits—affinity for technology and dispositional interpersonal trust—predicted both surface traits (column 5). Technological self-efficacy, however, did not provide a significant prediction for any situational trait. At the elemental level, neuroticism

TABLE 3: Standardized β Coefficients for Direct Effects (β X→ M; β M1 → M2; β M2 → Y, Columns 4–6) and the Associated Specific Indirect Effects (β X→ M1→ M2 → Y, Column 7) of the Initial Full Parallel-Serial-Mediation Model From Study 1

			Direct Effects			Spf. Indirect Effects
1	2	3	4	5	6	7
X	M1	M2	β X → M1	β M1 → M2	β M2 → Y	β X→ M1→M2→Y
N	AFT	PRO	**−.188 [−.324, −.062]***	**.470 [.337, .607]***	**.456 [.362, .544]***	**−.040 [−.075, −.013]***
N	AFT	ACC	"	**.322 [.187, .454]***	**.293 [.184, .399]***	**−.018 [−.037, −.005]***
N	TSE	PRO	**−.188 [−.308, −.071]***	−.056 [−.188, .073]	"	.005 [−.006, .020]
N	TSE	ACC	"	.032 [−.094, .170]	"	−.002 [−.010, .006]
N	INT	PRO	.047 [−.063, .157]	**.194 [.082, .302]***	"	.004 [−.006, .016]
N	INT	ACC	"	**.121 [.010, .234]***	"	.002 [−.003, .007]
E	AFT	PRO	−.089 [−.214, .042]	"	"	−.019 [−.047, .010]
E	AFT	ACC	"	"	"	−.008 [−.023, .004]
E	TSE	PRO	**−.145 [−.272, −.014]***	"	"	.004 [−.005, .015]
E	TSE	ACC	"	"	"	−.001 [−.009, .005]
E	INT	PRO	**.198 [.086, .305]***	"	"	**.017 [.005, .034]***
E	INT	ACC	"	"	"	**.007 [.000, .018]***
O	AFT	PRO	.072 [−.043, .189]	"	"	.015 [−.009, .042]
O	AFT	ACC	"	"	"	.007 [−.004, .022]
O	TSE	PRO	.010 [−.104, .122]	"	"	.000 [−.006, .004]
O	TSE	ACC	"	"	"	.000 [−.003, .003]
O	INT	PRO	−.034 [−.126, .056]	"	"	−.003 [−.012, .005]
O	INT	ACC	"	"	"	−.001 [−.006, .002]
A	AFT	PRO	−.013 [−.122, .097]	"	"	−.003 [−.027, .022]
A	AFT	ACC	"	"	"	−.001 [−.012, .010]
A	TSE	PRO	**−.135 [−.248, −.017]***	"	"	.003 [−.005, .014]
A	TSE	ACC	"	"	"	−.001 [−.008, .004]
A	INT	PRO	**.444 [.330, .565]***	"	"	**.039 [.015, .067]***
A	INT	ACC	"	"	"	**.016 [.001, .036]***
C	AFT	PRO	−.010 [−.122, .105]	"	"	−.002 [−.027, .023]
C	AFT	ACC	"	"	"	−.001 [−.013, .010]
C	TSE	PRO	−.041 [−.166, .089]	"	"	.001 [−.004, .008]
C	TSE	ACC	"	"	"	.000 [−.004, .003]
C	INT	PRO	.020 [−.097, .101]	"	"	.000 [−.009, .010]
C	INT	ACC	"	"	"	.000 [−.004, .005]
SE	AFT	PRO	**.195 [.026, .366]***	"	"	**.042 [.005, .088]***
SE	AFT	ACC	"	"	"	**.018 [.002, .043]***
SE	TSE	PRO	**.190 [.048, .334]***	"	"	−.005 [−.020, .006]
SE	TSE	ACC	"	"	"	−.003 [−.013, .004]
SE	INT	PRO	**.214 [.088, .343]***	"	"	**.019 [.005, .040]***
SE	INT	ACC	"	"	"	**.008 [.000, .019]***
EL	AFT	PRO	.026 [−.103, .162]	"	"	.006 [−.023, .033]
EL	AFT	ACC	"	"	"	.002 [−.010, .016]

(Continued)

TABLE 3 (Continued)

			Direct Effects			Spf. Indirect Effects
1	2	3	4	5	6	7
EL	TSE	PRO	−.031 [−.177, .113]	"	"	.001 [−.005, .008]
EL	TSE	ACC	"	"	"	−.000 [−.004, .003]
EL	INT	PRO	**−.135 [−.252, −.020]***	"	"	**−.012 [−.028, −.001]***
EL	INT	ACC	"	"	"	**−.005 [−.014, −.000]***
IL	AFT	PRO	.134 [−.022, .296]	"	"	.029 [−.003, .067]
IL	AFT	ACC	"	"	"	.013 [−.002, .031]
IL	TSE	PRO	.022 [−.118, .173]	"	"	−.001 [−.008, .005]
IL	TSE	ACC	"	"	"	.000 [−.003, .004]
IL	INT	PRO	.041 [−.074, .162]	"	"	.004 [−.007, .015]
IL	INT	ACC	"	"	"	.001 [−.003, .007]

Notes. X = elemental trait; M1 = situational trait; M2 = surface trait; Y = trust in automation; N = neuroticism; E = extraversion; O = openness; C = conscientiousness; E = extraversion; A = agreeableness; SE = self-esteem; EL = external locus of control; IL = internal locus of control; INT = dispositional interpersonal trust; AFT = affinity for technology; TSE = technological self-efficacy; PRO = propensity to trust in automation; ACC = a priori acceptability; * and bold indicate a significant effect (α = .05); " indicates that this effect has already been listed in a line further up in the respective column.

was negatively (column 4) and self-esteem positively (column 4) related to affinity for technology and technological self-efficacy. Furthermore, extraversion, agreeableness, self-esteem, and external LOC were positively related to dispositional interpersonal trust. For these elemental traits specific indirect effects on trust were significant. Neither consciousness nor openness predicted trust in automated driving in this study.

Some traits on the same level were correlated, which can lead to underestimation of relationships (Hayes, 2018). To rule this out we calculated a series of mediation models in which each trait was tested individually, without other traits (Table A.1 in the Appendix). On its own, technological self-efficacy explained variance in trust, but in combination with the other situational traits this was not the case, due to multicollinearity. As for this restricted additional variance explanation (e.g., to the affinity for technology), technological self-efficacy was omitted from the final model. While internal LOC was omitted from the final model for the same reason, external LOC was found to be negatively related to trust.

On the basis of these findings, an economic exploratory personality model (final model) for trust in automated driving was constructed and those traits that did not contribute incrementally to the explanation of variance were excluded. The nonsignificant correlation between the two situational traits was also removed.

Model test for the final model (Figure 3; Table 4) indicated a very good fit to the data, χ^2 (22, N = 274) = 27.73, p = .185, CFI = .988, RMSEA = .031, SRMR = .037. The modeled relationships explained 42.4% of the variance of trust in the automated driving system. This is equivalent to an effect size of f^2 = 0.74 which, according to Cohen (1992), represents a large effect. Furthermore, between 12.6% and 37.9% of the variance in the modeled surface and situational traits could be explained accounting for a f^2 between 0.14 (small to medium) and 0.61 (large).

Altogether 9 of the 10 hypotheses were supported. Trust in automated driving was predicted by the propensity to trust in automation (H9) and a priori acceptability of automated driving (H10), which in turn were predicted by affinity for technology (H7) and dispositional interpersonal

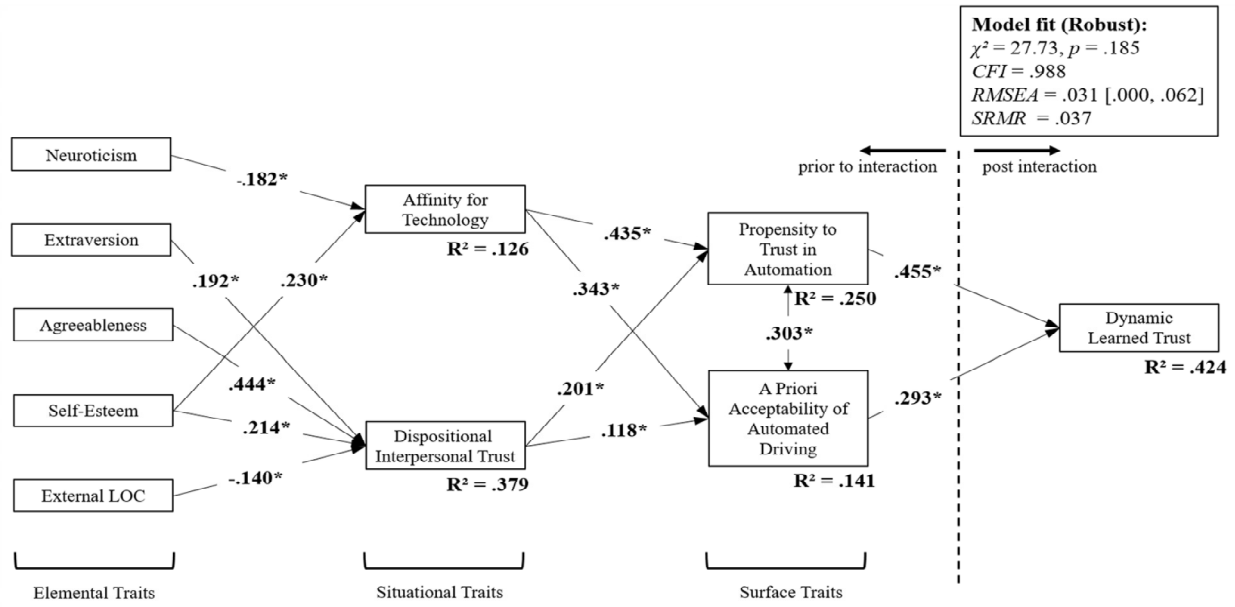


Figure 3. Fit statistics and path weights for the economic exploratory (final) personality model for trust in automated vehicles of Study 1; * indicates a significant effect ($\alpha = .05$).

TABLE 4: Standardized β Coefficients for Direct Effects (β X→ M; β M1 → M2; β M2 → Y, Columns 4–6) and the Associated Specific Indirect Effects (β X→ M1→ M2 → Y, Column 7) of the Economic Exploratory (Final) Model From Study 1

			Direct Effects			Spf. Indirect Effects
1	2	3	4	5	6	7
X	M1	M2	β X → M1	β M1 → M2	β M2 → Y	β X→ M1→M2→Y
N	AFT	PRO	**−.182 [−.307, −.060]***	**.435 [.322, .544]***	**.455 [.363, .548]***	**−.036 [−.064, −.012]***
N	AFT	ACC	"	**.343 [.229, .462]***	**.293 [.182, .401]***	**−.018 [−.036, −.005]***
E	INT	PRO	**.192 [.086, .293]***	**.201 [.084, .311]***	"	**.018 [.005, .034]***
E	INT	ACC	"	**.118 [.005, .235]***	"	**.007 [.000, .017]***
A	INT	PRO	**.444 [.328, .564]***	**.201 [.084, .311]***	"	**.041 [.016, .069]***
A	INT	ACC	"	**.118 [.005, .235]***	"	**.015 [.001, .035]***
SE	AFT	PRO	**.230 [.098, .382]***	**.435 [.322, .544]***	"	**.046 [.017, .087]***
SE	AFT	ACC	"	**.343 [.229, .462]***	"	**.023 [.008, .047]***
SE	INT	PRO	**.214 [.115, .315]***	**.201 [.084, .311]***	"	**.020 [.006, .037]***
SE	INT	ACC	"	**.118 [.005, .235]***	"	**.007 [.000, .017]***
EL	INT	PRO	**−.140 [−.252, −.028]***	**.201 [.084, .311]***	"	**−.013 [−.029, −.002]***
EL	INT	ACC	"	**.118 [.005, .235]***	"	**−.005 [−.014, −.000]***

Notes. X = elemental trait; M1 = situational trait; M2 = surface trait; Y = trust in automation; N = neuroticism; E = extraversion; A = agreeableness; SE = self-esteem; EL = external locus of control; AFT = affinity for technology; INT = dispositional interpersonal trust; PRO = propensity to trust in automation; ACC = a priori acceptability; * and bold indicates a significant effect ($\alpha = .05$); " indicates that this effect has already been listed in a line further up in the respective column.

trust (H6) at the situational level. At the elemental level, extraversion (H1), neuroticism (H2), agreeableness (H3), LOC (H4), and self-esteem (H5) showed significant relationships with the

situational traits. All specific indirect effects were significant.

Because Study 1 was exploratory the resulting model is possibly overfitted. Therefore, the model was validated in an independent sample in Study 2.

STUDY 2: MODEL VALIDATION

The aim of this study was to validate the final model explored in Study 1 using the same statistical procedure and criteria.

Method

Study 2 had a similar design as Study 1. After answering personality questionnaires and reading about automated driving and the "motorway pilot" the participants viewed seven video clips. The questionnaires were the same as in Study 1. Trust in automated driving was assessed at the end of each video. As in Study 1, manipulations of independent variables (system reliability and transparency) were carried out between the measurements of personality traits and trust in automation resulting in four study groups. While the reliability manipulation was designed in the same way as in Study 1, the transparency manipulation consisted only of two conditions: (1) high transparency: presence of a head-up display with information about system status and the current maneuver, (2) low transparency: no display. For this research, the full sample was used.

Sample

Participants were recruited as in Study 1. A total of 157 participants completed the survey; after applying the same exclusion criteria as in Study 1, the final sample comprised 149 participants (93 women, 62.4%), with a mean age of 29.96 years (SD = 11.33). The average completion time was 40 min and 17 s. For this sample, no data on driving experience and experience with ADAS was assessed. Descriptive statistics are presented in Table 5. Again, the overall mean for trust in the automated system was in a

TABLE 5: Means, Standard Deviation, Internal Consistency (Cronbach's α/Spearman–Brown Coefficient), and Correlations of All Traits and Trust in Automation in Study 2

	M	SD	α/ρ	1	2	3	4	5	6	7	8	9	10	11	12	13
1 O*	5.09	1.35	.53	—												
2 C*	4.92	1.10	.37	**.17**	—											
3 E*	4.45	1.50	.79	**.32**	**.23**	—										
4 A*	4.27	1.18	.25	.10	.08	.09	—									
5 N*	4.06	1.39	.66	.01	−.02	−.08	−.14	—								
6 SE	5.42	1.08	.86	**.22**	**.32**	**.35**	**.22**	**−.34**	—							
7 EL*	3.01	1.26	.59	−.10	−.11	**−.25**	−.15	.15	**−.30**	—						
8 IL*	5.74	1.08	.81	**.25**	**.25**	**.39**	**.23**	**−.29**	**.63**	**−.43**	—					
9 INT	4.81	1.18	.78	.13	.15	**.27**	**.50**	**−.18**	**.42**	−.16	**.40**	—				
10 AFT	4.77	1.06	.83	−.05	−.10	−.06	−.16	−.15	.08	−.09	**.20**	−.01	—			
11 TSE	5.14	1.38	.92	.06	−.12	−.02	−.12	−.16	.08	−.11	.14	−.01	**.73**	—		
12 PRO	4.62	1.32	.91	.05	.10	−.03	**.17**	**−.20**	.07	−.08	.09	.14	**.41**	**.26**	—	
13 ACC	4.43	1.48	.85	−.06	.01	−.08	−.10	−.06	−.09	.01	−.06	−.08	**.36**	**.22**	**.51**	—
14 Trust	4.81	1.44	.95	−.08	.11	−.08	.08	−.11	−.03	−.06	−.00	−.02	**.24**	.06	**.49**	**.60**

Notes. Bold indicates a significant correlation ($p < .05$); O = openness; C = conscientiousness; E = extraversion; A = agreeableness; N = neuroticism; SE = self-esteem; EL = external locus of control; IL = internal locus of control; INT = dispositional interpersonal trust; AFT = affinity for technology; TSE = technological self-efficacy; PRO = propensity to trust in automation; ACC = a priori acceptability; for variables with asterisks Spearman–Brown coefficient is reported instead of Cronbach's α.

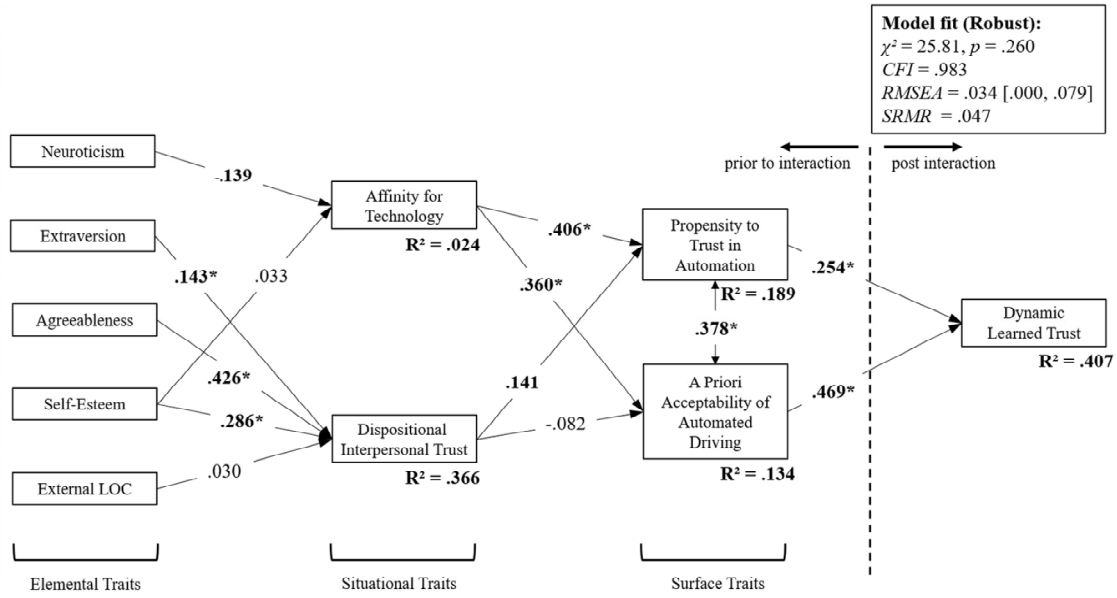


Figure 4. Validation of the reduced (final) model in Study 2. * indicates a significant effect (α = .05); bold indicates that the effect was significant in Study 1 and that the point estimate of Study 2 is within the 95% CI of the effect in Study 1.

range indicating a rather positive impression of the system (M = 4.81, SD = 1.44).

Study 2: Results

Preconditions. Once again there was no missing data, and no univariate or multivariate outliers were identified. Histograms and QQ plots indicated violations of normality, so the correction procedure used in Study 1 was applied again. Again, interactions between the traits and the independent variables were assessed. Significant interactions were found for neuroticism and propensity to trust in this study, so the relationships between these traits and trust in automation might be moderated by situational variables. Also in Study 2, group comparisons did not show any effects of the manipulations for the criterion variable, malfunction: $F(1, 144) = 1.38$, $p = .242$, $\eta_p^2 = .009$; transparency: $F(1, 144) = 0.87$, $p = .420$, $\eta_p^2 = .012$; interaction: $F(1, 144) = 0.84$, $p = .361$, $\eta_p^2 = .006$.

Study 2: Model Analysis

The overall model test (Figure 4) indicated that the final model offered a very good fit to the observed data, χ^2 (22, N = 149) = 25.81, p = .260, CFI = .983, RMSEA = .034, SRMR = .047. Similar to Study 1 the final model explained a substantial proportion of the variance in trust (40.7%; $f^2 = 0.69$; large) and between 13.4% and 36.6% ($f^2 = 0.15$ and $f^2 = 0.58$; medium to large effects) of the variance in the surface and situational traits. In this sample the model only explained 2.4% of the variance ($f^2 = 0.02$; small effect) in affinity for technology. Overall, the economic final personality model for trust in automated driving provides a good estimation of the relationships observed in a second, independent data set. However, some significant relationships from the first study were not found to be significant in Study 2 as indicated by CIs including zero (Table 6). This is, on one hand, an expected result given the exploratory nature of Study 1 and on the other hand might be a result of the lower N (and hence lower power) of Study 2 (Cumming, 2013; Maxwell et al., 2015). To provide an indication of which point estimates of the path weights of Study 2 might indicate the same effect as observed in Study 1 but missed significance due to the lower power of Study 2, we compared the point estimates of the path weights of Study 2 to the 95% CI of Study 1 in adoption of a similar procedure as discussed, for example, by the Open

TABLE 6: Standardized β Coefficients for Direct Effects (β X→ M; β M1 → M2; β M2 → Y, Columns 4–6) and the Associated Specific Indirect Effects (β X→ M1→ M2 → Y, Column 7) From the Validation of the Reduced (Final) Model in Study 2

			Direct Effects			Spf. Indirect Effects
1	2	3	4	5	6	7
X	M1	M2	β X → M1	β M1 → M2	β M2 → Y	β X→M1→M2→Y
N	AFT	PRO	**−.139 [−.313, .041]**	**.406 [.260, .552]***	**.254 [.096, .423]***	**−.014 [−.040, .004]**
N	AFT	ACC	"	**.360 [.212, .501]***	**.469 [.308, .622]***	**−.023 [−.059, .007]**
E	INT	PRO	**.143 [.001, .288]***	**.141 [−.027, .317]**	"	**.005 [−.001, .017]**
E	INT	ACC	"	−.082 [−.240, .080]	"	−.005 [−.022, .005]
A	INT	PRO	**.426 [.273, .577]***	**.141 [−.027, .317]**	"	.015 [−.003, .042]
A	INT	ACC	"	−.082 [−.240, .080]	"	−.016 [−.053, .014]
SE	AFT	PRO	.033 [−.140, .211]	**.406 [.260, .552]***	"	.003 [−.016, .025]
SE	AFT	ACC	"	**.360 [.212, .501]***	"	.006 [−.026, .036]
SE	INT	PRO	**.286 [.137, .448]***	**.141 [−.027, .317]**	"	**.010 [−.002, .029]**
SE	INT	ACC	"	−.082 [−.240, .080]	"	−.011 [−.040, .009]
EL	INT	PRO	.030 [−.107, .179]	**.141 [−.027, .317]**	"	.001 [−.006, .007]
EL	INT	ACC	"	−.082 [−.240, .080]	"	−.001 [−.014, .005]

Notes. X = elemental trait; M1 = situational trait; M2 = surface trait; Y = trust in automation; N = neuroticism; E = extraversion; A = agreeableness; SE = self-esteem; EL = external locus of control; AFT = affinity for technology; INT = dispositional interpersonal trust; PRO = propensity to trust in automation; ACC = a priori acceptability; * indicates a significant effect (α = .05); bold without asterisk indicates that the effect was significant in Study 1 and the point estimate of Study 2 is within the CI of Study 1; " indicates that this effect has already been listed in a line further up in the respective column.

Science Collaboration (2015). Path weights for three effects were in the same range as in Study 1 (N→AFT, E→INT, INT→PRO) and therefore it can be concluded that these effects are likely different from zero. On the other hand, point estimates for the effects of self-esteem on affinity for technology, of external LOC on dispositional interpersonal trust, and of dispositional interpersonal trust on the a priori acceptability do not fall within the corresponding 95% CI, so it seems plausible that in Study 1 these three effects were found due to α-error inflation. Further studies might clarify these discrepancies.

Overall, as in Study 1, the final personality model for trust in automated driving was supported by a good model fit to the observed data. For seven of the nine hypotheses supported in Study 1, at least some of the respective β coefficients in this study were in the same range as indicated by overlapping CIs. A relationship of LOC with trust in an automated driving system was not supported by the data in Study 2 (H4) and thus the role of this trait awaits further investigation. Furthermore, as no specific indirect effect for agreeableness was supported, the role of this trait for trust in automation needs further investigation. Overall, Study 2 provides strong evidence that the vast majority of the modeled personality traits contribute to the explanation of trust in automated driving.

DISCUSSION

This research provides a first integrative investigation of personality traits in relation to trust in automation. The goal of this research was to present a model that integrates different personality constructs according to their specificity for trust in automation and to investigate how the effects of more general traits are mediated through more specific traits. The indirect

effects provide insights into some mechanisms of how general traits might influence trust formation via the establishment of more specific traits and technology-related attitudes.

Evaluation of the Final Personality Model for Trust in Automation

In both studies, the economic final personality model showed a very good fit to the data and explained a substantial proportion of the variance in dynamic learned trust in an automated driving system: 42.4% of the variance in trust in automation in Study 1 and 40.7% of the variance in Study 2 could be explained by the antecedent modeled personality traits. The combined evidence from the two studies provides strong support for the notion that the modeled personality traits play an important role in the formation of trust in automated driving and might help to explain individual differences. In line with the logic of hierarchical personality models, correlations between the more specific surface traits and trust in automation are higher than those with the traits at the antecedent trait levels. In the case of most of the variables in the presented model—neuroticism, extraversion, and self-esteem at the elemental level; affinity for technology and dispositional interpersonal trust at the situational level; propensity to trust in automation at the surface level—this research provides empirical support for their prediction of trust in automated driving. Furthermore, the modeled variables explain essential variance proportions of the included situational and surface traits.

Role of Personality and Attitudes in the Emergence of Trust in Automation

The indirect effects provide insight into the psychological processes involved in trust formation and are a promising starting point for system design aiming to facilitate trust calibration in automated driving (e.g., Parasuraman & Riley, 1997). First of all, the findings of this study support a prediction of dynamic learned trust in automation by the surface traits propensity to trust in automation and the attitude toward automated driving. These two specific dispositions, in turn, are predicted well by a general affinity for technology. People who like to try out new technology and are not scared of it tend to be more open to trust and accept automated driving systems in general. The negative relationship of neuroticism to affinity for technology indicates that this openness for technology is, in part, based on the tendency to be more worrying and risk-averse in general. In this regard, new technology like automated driving seems to cause increased uncertainty in neurotic persons resulting in a less permissive attitude and lower trust in automated systems. Furthermore, dispositional interpersonal trust predicts the propensity to trust in automation, which underlines that the psychological mechanisms in the emergence of trust in other people generalize to the technological domain to some extent. The prediction of interpersonal trust by extraversion, agreeableness, and self-esteem point to the direction that a more open and positive approach to new stimuli as well as a self-evaluation as being capable of doing well in face of challenges might be a psychological basis for increased trust levels in both domains.

Role of Dispositional Trust for Dynamic Learned Trust

This research straightens out some open questions about the interrelations between different trust concepts in regard to the use of automated technology. Our findings emphasize that it is meaningful to distinguish between at least three trust variables on the basis of their generality and reference point: history-based dynamic learned trust in a specific automated system and its dispositional antecedents, the more general propensity to trust in automation and the yet more general dispositional interpersonal trust. It is therefore recommended that several trust constructs are considered simultaneously in any research on trust in automation (e.g., Hoff & Bashir, 2015), as otherwise empirical findings may be confused and difficult to interpret (Merritt & Ilgen, 2008). As this research did not include initial learned trust and situational influences on trust (situational trust), the role of these concepts in relation to the three trust concepts discussed here could be examined in future studies.

Role of Propensity to Trust in Automation and Prior Attitudes

In regard to trust formation, especially the propensity to trust in automation needs to be considered along with the a priori acceptability of automated driving. The strong relationships between these dispositions underline the importance of the propensity to trust in automation and prior attitudes towards automated driving for the formation of trust in regard to a new system. This is in line with the findings of earlier studies (e.g., Merritt et al., 2013), which mostly investigated other technological domains. Unsurprisingly previous experience with similar systems and existing knowledge of the technology play a role in the formation of the propensity to trust in automation and specific technology-related attitudes; however, the relationships between the specific attitude and more general personality variables help to understand individual differences in trust in automated driving systems. These dispositions build an essential starting point in driver education to prevent distrust and overtrust and to facilitate a well-balanced usage behavior of driving automation.

Indirect Effects of General Broader Traits on Trust in Automation

This study provides initial support for a mediation of the relationship between neuroticism and trust in automated driving by the affinity for technology. The latter seems to play an important role in both the emergence of a general tendency to trust automated technology and acceptance of automated driving. Neuroticism was negatively related to the affinity for technology in both studies, corroborating earlier studies indicating a negative relationship between neuroticism and the adoption of automated technology (e.g., Szalma & Taylor, 2011). Individuals who are less open to others and tend to worry more also seem to be somewhat less enthusiastic about technology and hence less inclined to trust automated systems. In addition, the reported studies provide initial evidence that dispositional interpersonal trust is a predictor of propensity to trust in automation. Dispositional interpersonal trust is, in turn, predicted by extraversion, agreeableness, and self-esteem. These relationships support the popular notion that there is some common psychological basis for human–human trust and the more specific trust in automation (Hoff & Bashir, 2015; Lee & Moray, 1992) and that learned trust is rooted in personality traits reflecting a positive self-evaluation and a positive and cooperative attitude toward others. Nevertheless, although interpersonal trust and trust in automated technology share some variance components in the face of the additional paths predicting trust in automated driving via the affinity for technology, they are clearly not the same and, as indicated by the data in both studies, dispositional interpersonal trust seems to be a more general trait predicting trust in automation.

The reported studies did not provide evidence of a relationship between the remaining Big Five factors (openness and consciousness) and trust in automation. This conflicts with a recent study by Chien et al. (2016), who found a significant correlation of consciousness and trust. It should be noted that technological self-efficacy showed substantial predictive value for trust in automation but was excluded from the model due to multicollinearity with other traits from the same personality cluster. Furthermore, at this point, the role of LOC for trust formation in automated driving systems remains unclear and bears further investigation. In the same manner, as none of the specific indirect effects of agreeableness on dynamic learned trust in automation was supported in Study 2, the role of this trait for trust in automation (above its role for interpersonal trust) needs further consideration in future research endeavors.

Strengths, Limitations, and Future Research

The reported studies combine a theory-based exploration and a subsequent empirical validation of a hierarchical personality model. Important strengths of the research are the simultaneous estimation of multiple paths with *SEM* and the good model fit in two independent data sets. Hereby, the differences between the two study settings and driving

systems support the generalizability and external validity of the findings. The interpretation of the combination of the included direct and indirect effects in the investigated *SEM* model provides profound insights into the psychological mechanism of the interplay of general and specific traits in the formation of trust in automation.

Like all research, the reported studies have some weaknesses. First, Study 2 had a smaller sample than Study 1 and although, overall, it provided support for the validity of the model, some of the significant relationships found in Study 1 were not significant in Study 2. As supported by overlapping CIs, most point estimates for β weights fell in the same range in both studies and thus might be viewed as supported by the data. Whether the remaining three nonsignificant paths should be omitted from the model is a question to be answered in future studies. Second, the low proportion of variance in affinity for technology explained by the model in Study 2 indicates that the foundation of this trait in broader constructs could not be conclusively clarified in this research. Third, the short scales used to assess the Big Five and LOC lead to small Cronbach's αs in our studies and might be replicated in studies using scales with more items. Whatsoever, an underestimation of reliability is in the nature of Cronbach's α in short scales (see Cortina, 1993; Lord & Novick, 2008; Starkweather, 2012). For the two scales, substantial proof of psychometric quality exists (see Rammstedt & John, 2007; Rammstedt et al., 2013, for the BFI-10 and Kovaleva et al., 2012, for the IE-4) and the reported relationships seem valid irrespective of the reported low Cronbach's αs. Fourth, while neuroticism in Study 1 was directly and indirectly and in Study 2 indirectly related to trust in automated driving (while the direct effect was not significant, but showed a similar effect size as in Study 1), the reported interactions between neuroticism and the experimental manipulations imply that the relationship between neuroticism and trust in automation might be moderated by situational factors, which could be in focus of future studies. In this sense, future research could furthermore explore interaction effects between personality variables, situational characteristics (choice between different systems, perception of risk, uncertainty, etc.), and system features (e.g., transparency, reliability) in the formation of learned trust in automated driving systems. For example, the perfect automation scheme (especially the factor "all-or-none thinking") might be a possible moderator of drivers' reaction to system malfunctions (Merritt et al., 2015). Above this, the factor "high expectations" of the perfect automation scheme might be a potential variable to include on the more specific levels of the model in the future. Thereby, its relationship to the affinity for technology and (as for the essential conceptual overlap) its unique variance explanation to the included trait propensity to trust in automation would be of particular interest. As a second example the individual tendency for complacency as "one's propensity to monitor automation sub-optimally" (Merritt et al., 2019, p. 225) might moderate the effects of automation features like system transparency. Additionally, it might be investigated how the modeled personality traits directly affect reliance on and the interaction with automated driving systems. Finally, the various specific personality variables found to be relevant for individual differences in trust levels could be used to select participants for future studies and as control variables or moderators.

Practical Implications

This research provides insights into the basic psychological processes that guide individual differences in the emergence of trust in automation. While the theoretical implications of this research have been discussed above, the reported findings on personality variables provide the foundation to further investigate how differences in driver personality direct the perception and reliance on particular automated systems in practice. As for the strong effect of personality variables at different levels in the trait hierarchy, it can be concluded that different people will show different psychological and behavioral reactions to the same automated vehicle (even in the same situation).

Above these different evaluations, the investigated personality differences might moderate

the effects of situational variables and properties of automated systems on the evaluation and adoption of automated vehicles (e.g., Szalma & Taylor, 2011). More specifically, the presented findings suggest that individual differences play a role in how available information is searched and processed in building up an impression of automated driving systems and calibrating trust at different stages of system interaction (e.g., prior to system acquisition or during system use). Different people pay attention to different aspects of information. For example, people with low self-esteem or higher neuroticism might be more sensitive to negative information about possible dangers and risks of the automation, while on the other hand, people who are more open and enthusiastic about technology might tend to ignore such information. First evidence for such interaction effects between personality and available information in the evaluation of automated systems stems from Kraus, Forster et al. (2019a), who show that need for cognition moderates the use of different kinds of information (brand information vs. reliability information) prior to the presentation of an automated driving system. In terms of the information processing in trust formation, this research provides directions for strategies for the personalized design of information campaigns (of both public stakeholders and manufacturers), driver training, and vehicle instruction, as well as the design of user interfaces and the driver–vehicle interaction in automated driving. The investigated mediation paths could be used to derive strategies for trust calibration. In the best manner, such strategies would aim at calibrated trust levels and an aggregated reliance behavior in line with the capacities and limits of the system. To clarify how this research helps to advance marketing and information campaigns, driver training design, the design of user interfaces, and the use of automated cars, in the following, we provide detailed practical implications for these areas. Thereby we provide both more specific implications from the (direct and indirect) relationships of the modeled variables and more general recommendations for a consideration of personality differences in regard to trust in automation in practical settings. As all measures to increase traffic safety, such programs should be objectively evaluated for their effectivity and economic cost–benefit ratio.

Information Campaigns on Automated Driving Systems

The information available on automated driving systems (e.g., advertisements, media reports, government information, word of mouth, and instruction manuals) influences people's expectations about its characteristics, functionality, and reliability. Based on the reported strong relationships of prior attitudes and the propensity to trust, expectations might be systematically guided by information campaigns from both public or governmental institutions and car manufacturers in order to evoke realistic attitudes to automated driving and trust levels and to facilitate trust calibration. This information can be directed at (potential) users at different points in time prior to and during their acquisition of or the interaction with automated driving systems. For example, advertisement could be designed to explicitly address people's fears about automated driving like potential system limitations and thereby triggering a more positive evaluation of the system without bias by potentially irrational fears. Thereby all communication aiming at increased initial trust should be realistic and avoid to elicit overtrust as disappointed expectations might lead to even lower levels of dynamic learned trust. Above this, the reported findings on personality differences help to identify different user groups, who might be presented with specific information along with their preferences or information processing style in their assessment of automated systems. These user groups can then specifically be addressed with information campaigns in different stages of the customer journey (e.g., prior to acquisition or during system use). Depending on the phase of the information campaign, different strategies can be thought of. For example, an information campaign by car manufacturers prior to the acquisition, based on personality features, might be based on an assessment of "typical users" or personas, with certain personality patterns (this

can be assessed by customer surveys in market research and statistically analyzed with cluster analyses).

Driver Training and Vehicle Instructions for Automated Vehicles

Driver training is essential for trust calibration in automated driving systems as it provides a possibility to systematically educate drivers on automated driving systems in general as well as the functionality of specific systems. Formal driver training on driving automation can be implemented in different ways. In countries in which formal driver training is mandatory, it can easily be integrated into the schedule and where driver training is mainly informally provided by family members or friends, corresponding test questionnaires or tasks in the practical exams can make sure that proper education on vehicle automation has been received. Furthermore, such training could also be part of the acquisition of a new vehicle that is equipped with an automated driving system (e.g., as part of the vehicle introduction at the car dealership) or part of a re-training program for elderly drivers for enabling them to use sophisticated driving assistance. Thereby driver training might have different functions. First, it may help to increase initial trust and acceptability of automated systems by providing insight into system capabilities and functionality and reducing anxiety. Second, it might be aiming at the calibration of individual trust levels and balancing of reliance behavior. In line with Szalma (2008, 2009) and as a result of this research, it can be recommended that training should be tailored on the basis of personality characteristics. For example, in order to calibrate the users' trust levels more adequately prior to the first system use, users might be assessed prior to the training with information that shapes their initial attitudes and trust levels more in the direction of a realistic assessment of risks and benefits of the system. For example, people showing overly high levels of initial trust could be addressed with information stressing the potential dangers of over-reliance or the out-of-the-loop problem. On the other hand, distrusting persons could learn about the advantages of automated driving systems, for example, in terms of traffic safety. The relationships of the surface traits with traits at the more general model levels might help to design provided information to address some of the underlying psychological mechanisms adequately.

Design of User Interfaces and Driver–Vehicle Interaction in Automated Driving

Trust calibration occurs in different phases of the interaction with automated systems. While prior to interaction with automation, available information is used to calibrate trust, trust is further calibrated during the drive along with the presented information in the user interface and the behavior of the vehicle (e.g., Hoff & Bashir, 2015; Kraus, Scholz, et al., 2019). The personality traits and mediation paths of the presented hierarchical personality model can be used in adaptive interface design in automated vehicles taking into account individual dispositions and preferences in order to "better accommodate the trusting tendencies of diverse operators" (Hoff & Bashir, 2015, p. 415). Personalization of designs (Hancock et al., 2009) and the provided information on the basis of the reported relationships might help to calibrate trust and acceptability of automated systems. For example, people with low self-esteem might gain increased trust in their own ability to use a system in a functional manner by an interface providing positive feedback for how they drive or use the system—which in turn should be associated with a higher trust in the system. In the same manner, persons with overly high initial trust levels might be presented with salient information on system limitations or real-time information of possible dangers for the functionality of the automation during the drive. In practice, in order to allow for such personalization of design, personality variables could be assessed either during or after vehicle acquisition in the car dealership, in the vehicle, or at any other digital device connected to the vehicle (e.g., vehicle app) and then be implemented in driver profiles. In this sense, personalization via driver profiles would also allow individualizing one vehicle for different drivers (e.g., personalization of the user interface in rental cars

along with user profiles). In the same manner, personalization on the basis of personality and trust levels would allow providing individualized information during the drive to persuade drivers to use the automation to the extent that maximizes comfort and safety gains.

Conclusion

Our findings underline the benefits of hierarchical personality modeling and *SEM* in human factors research. The simultaneous consideration of general and specific dispositional variables provides valuable insight into the psychological mechanisms leading to individual differences in levels of trust in automation. The reported relationships involving extraversion, neuroticism, and self-esteem help to relate the impact of more specific traits to theoretically well-understood aspects of personality. This provides a deeper understanding of the affective and cognitive processes associated with the generation of trust in automation. Future research might build on the reported findings to enhance the knowledge of how personality contributes to differences in the perception and use of automated technology.

KEY POINTS

- A hierarchical personality model for dynamic learned trust in automated driving was presented.
- The proposed model was tested in two online studies.
- Study 1 was an exploratory analysis of the hypothesized model. Relationships that were not empirically supported were removed. Study 2 replicated the final model in an independent data set.
- Both studies showed a good fit of the economic final model to the data.
- The final model supports a role of neuroticism, extraversion, and self-esteem at the most general level; affinity for technology and dispositional interpersonal trust at the second level; and propensity to trust in automation and a priori acceptability of automated driving at the most specific level for the prediction of trust in automation.
- The model explains an essential proportion of variance in dynamic learned trust in automation.
- The role of personality and attitudes in the development of trust in automation as well as implications for theory and practice are discussed.

APPENDIX

Table A1: Direct and Indirect Effects of Each Trait, When Considered Without the Other Traits on the Same Level to Avoid the Effects of Multicollinearity (Data of Study 1)

X	M1	M2	β X → M1	β M1 → M2	β M2 → Y	β X→ M1→M2→Y
N	AFT	PRO	**−.290 [−.408, −.173]***	**.467 [.355, .584]***	**.456 [.362, .544]***	**−.062 [−.101, −.032]***
N	AFT	APC	"	**.362 [.242, .483]***	**.293 [.184, .399]***	**−.027 [−.049, −.011]***
N	TSE	PRO	**−.254 [−.368, −.137]***	**.237 [.118, .362]***	"	.007 [−.008, .024]
N	TSE	APC	"	**.233 [.121, .355]***	"	−.002 [−.013, .007]
N	INT	PRO	**−.137 [−.254, −.018]***	**.273 [.146, .396]***	"	**−.012 [−.027, −.001]***
N	INT	APC	"	**.175 [.053, .297]***	"	−.005 [−.013, .000]
E	AFT	PRO	.058 [−.070, .176]	"	"	.012 [−.014, .041]
E	AFT	APC	"	"	"	.005 [−.007, .019]
E	TSE	PRO	−.030 [−.157, .096]	"	"	.001 [−.004, .008]
E	TSE	APC	"	"	"	.000 [−.004, .003]
E	INT	PRO	**.275 [.156, .387]***	"	"	**.024 [.008, .044]***
E	INT	APC	"	"	"	**.010 [.001, .022]***
O	AFT	PRO	.087 [−.030, .210]	"	"	.019 [−.007, .047]
O	AFT	APC	"	"	"	.008 [−.003, .023]
O	TSE	PRO	−.003 [−.124, .119]	"	"	.000 [−.005, .005]
O	TSE	APC	"	"	"	.000 [−.003, .003]

(continued)

X	M1	M2	β X → M1	β M1 → M2	β M2 → Y	β X→ M1→M2→Y
O	INT	PRO	.005 [−.107, .125]	"	"	.000 [−.010, .012]
O	INT	APC	"	"	"	.000 [−.005, .005]
A	AFT	PRO	.009 [−.105, .120]	"	"	.002 [−.023, .028]
A	AFT	APC	"	"	"	.001 [−.010, .013]
A	TSE	PRO	−.118 [−.239, .002]	"	"	.003 [−.004, .013]
A	TSE	APC	"	"	"	−.001 [−.007, .004]
A	INT	PRO	**.471 [.340, .604]***	"	"	**.042 [.016, .071]***
A	INT	APC	"	"	"	**.017 [.001, .038]***
C	AFT	PRO	.058 [−.060, .173]	"	"	.012 [−.013, .039]
C	AFT	APC	"	"	"	.005 [−.006, .018]
C	TSE	PRO	−.010 [−.137, .114]	"	"	.000 [−.004, .007]
C	TSE	APC	"	"	"	.000 [−.003, .003]
C	INT	PRO	.110 [−.008, .229]	"	"	.010 [−.001, .025]
C	INT	APC	"	"	"	.004 [−.001, .012]
SE	AFT	PRO	**.316 [.199, .454]***	"	"	**.068 [.035, .114]***
SE	AFT	APC	"	"	"	**.030 [.012, .056]***
SE	TSE	PRO	**.232 [.116, .369]***	"	"	−.006 [−.023, .008]
SE	TSE	APC	"	"	"	.002 [−.008, .013]
SE	INT	PRO	**.364 [.244, .468]***	"	"	**.032 [.012, .058]***
SE	INT	APC	"	"	"	**.013 [.001, .028]***
EL	AFT	PRO	**−.116 [−.236, −.001]***	"	"	**−.025 [−.056, −.000]***
EL	AFT	APC	"	"	"	**−.011 [−.026, −.000]***
EL	TSE	PRO	−.120 [−.249, .009]	"	"	.003 [−.004, .014]
EL	TSE	APC	"	"	"	−.001 [−.008, .004]
EL	INT	PRO	**−.278 [.404, −.144]***	"	"	**−.025 [−.047, −.008]***
EL	INT	APC	"	"	"	**.010 [.000, .025]***
IL	AFT	PRO	**.263 [.139, .403]***	"	"	**.056 [.026, .098]***
IL	AFT	APC	"	"	"	**.025 [.010, .047]***
IL	TSE	PRO	**.140 [.017, .275]***	"	"	−.004 [−.016, .005]
IL	TSE	APC	"	"	"	.001 [−.005, .008]
IL	INT	PRO	**.254 [.135, .374]***	"	"	**.022 [.007, .043]***
IL	INT	APC	"	"	"	**.009 [.001, .021]***

Notes. X = elemental trait; M1 = situational trait; M2 = surface trait; Y = trust in automation; N = neuroticism; E = extraversion; O = openness; C = conscientiousness; A = agreeableness; SE = self-esteem; EL = external locus of control; IL = internal locus of control; INT = interpersonal trust; AFT = affinity for technology; TSE = technological self-efficacy; PRO = propensity to trust in automation; ACC = a priori acceptability; * and bold indicate a significant effect (α = .05); " indicates that this effect has already been listed in a line further up in the respective column.

ACKNOWLEDGMENTS

Thanks to all students and colleagues for their help in conducting the studies and for helpful comments and discussions.

ORCID iD

Johannes Kraus https://orcid.org/0000-0001-7015-8477

REFERENCES

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *Journal of Applied Social Psychology*, *32*, 665–683. https://doi.org/10.1111/j.1559-1816.2002.tb00236.x

Ajzen, I. (2005). *Attitudes, personality, and behavior* (2nd ed.). Open University Press.

Albarracín, D., Chan, M. P. S., & Jiang, D. (2018). Attitudes and attitude change: Social and personality considerations about specific and general patterns of behavior. In K. Deux & M. Snyder (Eds.), *The Oxford handbook of personality and social psychology* (pp. 439–463). Oxford University Press.

Allport, G. W. (1961). *Pattern and growth in personality*. Holt, Rinehart & Winston.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, *50*, 248–287. https://doi.org/10.1016/0749-5978(91)90022-L

Beierlein, C., Kemper, C. J., Kovaleva, A., & Rammstedt, B. (2012). *Kurzskala zur Messung des zwischenmenschlichen Vertrauens: die Kurzskala Interpersonales Vertrauen (KUSIV3)*. GESIS-Working Papers 2012|22. GESIS.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS Structural Equations Program Manual*. Multivariate Software.

Berenbaum, H., Bredemeier, K., & Thompson, R. J. (2008). Intolerance of uncertainty: Exploring its dimensionality and associations with need for cognitive closure, psychopathology, and personality. *Journal of Anxiety Disorders*, *22*, 117–125. https://doi.org/10.1016/j.janxdis.2007.01.004

Biros, D. P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, *13*, 173–189. https://doi.org/10.1023/B:GRUP.0000021840.85686.57

Bosnjak, M., Galesic, M., & Tuten, T. (2007). Personality determinants of online shopping: Explaining online purchase intentions using a hierarchical approach. *Journal of Business Research*, *60*, 597–605. https://doi.org/10.1016/j.jbusres.2006.06.008

Briggs, P., Burford, B., & Dracup, C. (1998). Modelling self-confidence in users of a computer-based system showing unrepresentative design. *International Journal of Human-Computer Studies*, *49*, 717–742. https://doi.org/10.1006/ijhc.1998.0224

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Chancey, E. T., Bliss, J. P., Proaps, A. B., & Madhavan, P. (2015). The role of trust as a mediator between system characteristics and response behaviors. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*, 947–958. https://doi.org/10.1177/0018720815582261

Chien, S. Y., Sycara, K., Liu, J. S., & Kumru, A. (2016). Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and Big Five personality traits. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 841–845). SAGE Publications.

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, *31*, 692–702. https://doi.org/10.1080/10447318.2015.1070549

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Coopersmith, S. (1981). *Coopersmith self-esteem inventory (school form)*. Consulting Psychology Press.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R professional manual*. Psychological Assessment Resources.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. U. Illinois Press.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, *35*, 982–1003. https://doi.org/10.1287/mnsc.35.8.982

de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, *58*, 719–735. https://doi.org/10.1016/S1071-5819(03)00039-9

Dishaw, M., Strong, D., & Bandy, D. B. (2002). Extending the task-technology fit model with self-efficacy constructs. *AMCIS 2002 Proceedings*, 1021–1027.

DuCette, J., & Wolk, S. (1973). Cognitive and motivational correlates of generalized expectancies for control. *Journal of Personality and Social Psychology*, *26*, 420–426. https://doi.org/10.1037/h0034486

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality*, *42*, 1585–1593. https://doi.org/10.1016/j.jrp.2008.07.011

Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the Affinity for Technology Interaction (ATI) scale. *International Journal of Human–Computer Interaction*, *35*, 456–467. https://doi.org/10.1080/10447318.2018.1456150

Gaines, S. O., Panter, A. T., Lyde, M. D., Steers, W. N., Rusbult, C. E., Cox, C. L., & Wexler, M. O. (1997). Evaluating the circumplexity of interpersonal traits and the manifestation of interpersonal traits in interpersonal trust. *Journal of Personality and Social Psychology*, *73*(3), 610–623. https://doi.org/10.1037/0022-3514.73.3.610

Galvin, B. M., Randel, A. E., Collins, B. J., & Johnson, R. E. (2018). Changing the focus of locus (of control): A targeted review of the locus of control literature and agenda for future research. *Journal of Organizational Behavior*, *39*, 820–833. https://doi.org/10.1002/job.2275

Gold, C., Körber, M., Hohenberger, C., Lechner, D., & Bengler, K. (2015). Trust in automation – before and after the experience of take-over scenarios in a highly automated vehicle. *Procedia Manufacturing*, *3*, 3025–3032. https://doi.org/10.1016/j.promfg.2015.07.847

Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *23*, 443–451. https://doi.org/10.1177/0146167297235001

Greszki, R., Meyer, M., & Schoen, H. (2014). The impact of speeding on data quality in non-probability and freshly recruited probability-based online panels. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online Panel Research: A Data Quality Perspective* (pp. 238–262). Wiley.

Gurtman, M. B. (1992). Trust, distrust, and interpersonal problems: A circumplex analysis. *Journal of Personality and Social Psychology*, *62*, 989–1002. https://doi.org/10.1037/0022-3514.62.6.989

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-

analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*, 517–527. https://doi.org/10.1177/0018720811417254

Hancock, P. A., Hancock, G. M., & Warm, J. S. (2009). Individuation: The *N* = 1 revolution. *Theoretical Issues in Ergonomics Science*, *10*, 481–488. https://doi.org/10.1080/14639220903106387

Harter, S. (1993). Causes and consequences of low self-esteem in children and adolescents. In R. Baumeister (Ed.), *Self-esteem: The puzzle of low self-regard* (pp. 87–116). Plenum.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.

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.

Herz, L., & Gullone, E. (1999). The relationship between self-esteem and parenting style: A cross-cultural comparison of Australian and Vietnamese Australian adolescents. *Journal of Cross-Cultural Psychology*, *30*, 742–761.

Hiroto, D. S. (1974). Locus of control and learned helplessness. *Journal of Experimental Psychology*, *102*, 187–193. https://doi.org/10.1037/h0035910

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

Holcomb, L. B., King, F. B., & Brown, S. W. (2004). Student traits and attributes contributing to the success in online courses: Evaluation of university online courses. *The Journal of Interactive Online Learning*, *2*, 1–17.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

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. https://doi.org/10.1207/S15327566IJCE0401_04

Judge, T. A., & Bono, J. E. (2001). Relationship of core self-evaluations traits—self-esteem, generalized self-efficacy, locus of control, and emotional stability—with job satisfaction and job performance: A meta-analysis. *Journal of Applied Psychology*, *86*, 80–92. https://doi.org/10.1037/0021-9010.86.1.80

Judge, T. A., Locke, E. A., Durham, C. C., & Kluger, A. N. (1998). Dispositional effects on job and life satisfaction: The role of core evaluations. *Journal of Applied Psychology*, *83*, 17–34. https://doi.org/10.1037/0021-9010.83.1.17

Karrer, K., Glaser, C., Clemens, C., & Bruder, C. (2009). Technikaffinität erfassen - der Fragebogen TA-EG. In A. Lichtenstein, C. Stößel, & C. Clemens (Eds.), *Der Mensch im Mittelpunkt technischer Systeme, 8. Berliner Werkstatt Mensch-Maschine-Systeme* (pp. 196–201). VDI Verlag GmbH.

Körber, M., & Bengler, K. (2014). Potential individual differences regarding automation effects in automated driving. In *Proceedings of the XV International Conference on Human Computer Interaction* (pp. 1–7). ACM.

Kovaleva, A. (2012). The IE-4: Construction and validation of a short scale for the assessment of locus of control. 9. *GESIS-Schriftenreihe*. GESIS.

Kovaleva, A., Beierlein, C., Kemper, C. J., & Rammstedt, B. (2012). Eine Kurzskala zur Messung von Kontrollüberzeugung: Die Skala Internale-Externale-Kontrollüberzeugung-4 (IE-4). (GESIS-Working Papers, 2012/19). Mannheim: GESIS - Leibniz-Institut für Sozialwissenschaften. (07 July, 2019). https://nbn-resolving.org/urn:nbn:de:0168-ssoar-312096

Kraus, J. M., Forster, Y., Hergeth, S., & Baumann, M. (2019). Two routes to trust calibration: Effects of reliability and brand information on trust in automation. *International Journal of Mobile Human Computer Interaction*, *11*, 1–17.

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

Kraus, J., Scholz, D., Messner, E. -M., Messner, M., & Baumann, M. (2020). Scared to trust? Predicting trust in highly automated driving by depressiveness, negative self-evaluations and state anxiety. *Frontiers in Psychology*, *10*, 1–15. https://doi.org/10.3389/fpsyg.2019.02917

Larsen, R. J., & Buss, D. M. (2005). *Personality psychology: Domains of knowledge about human nature*. McGraw-Hill.

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

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

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*, 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Lewandowsky, S., Mundy, M., & Tan, G. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied*, *6*, 104–123.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Lawrence Erlbaum Associates.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593–614. https://doi.org/10.1146/annurev.psych.58.110405.085542

Madhavan, P., & Phillips, R. R. (2010). Effects of computer self-efficacy and system reliability on user interaction with decision support systems. *Computers in Human Behavior*, *26*, 199–204. https://doi.org/10.1016/j.chb.2009.10.005

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

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487–498. https://doi.org/10.1037/a0039400

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*, 81–90. https://doi.org/10.1037/0022-3514.52.1.81

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*, 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McDonald, T., & Siegall, M. (1992). The effects of technological self-efficacy and job focus on job performance, attitudes, and withdrawal behaviors. *The Journal of Psychology*, *126*, 465–475. https://doi.org/10.1080/00223980.1992.10543380

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors*, *53*, 356–370. https://doi.org/10.1177/0018720811411912

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

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, *55*, 520–534.

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. https://doi.org/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. https://doi.org/10.1177/0018720815581247

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, *6*, 44–58.

Mowen, J. C. (2000). *The 3M model of motivation and personality: Theory and empirical applications to consumer behavior*. Kluwer Academic Press.

Mowen, J. C., Park, S., & Zablah, A. (2007). Toward a theory of motivation and personality with application to word-of-mouth communications. *Journal of Business Research*, *60*, 590–596. https://doi.org/10.1016/j.jbusres.2006.06.007

Mowen, J. C., & Voss, K. E. (2008). On building better construct measures: Implications of a general hierarchical model. *Psychology and Marketing*, *25*, 485–505. https://doi.org/10.1002/mar.20221

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*, 527–539. https://doi.org/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. https://doi.org/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. https://doi.org/10.1080/00140139608964474

Oehler, A., & Wedlich, F. (2018). The relationship of extraversion and neuroticism with risk attitude, risk perception, and return expectations. *Journal of Neuroscience, Psychology, and Economics*, *11*, 63–92. https://doi.org/10.1037/npe0000088

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth–fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior*, *17*, 609–626. https://doi.org/10.1002/(SICI)1099-1379(199611)17:6<609::AID-JOB1828>3.0.CO;2-K

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 943-aac4716.

Owens, T. J. (1993). Accentuate the positive and the negative: Rethinking the use of self-esteem, self-deprecation, and self-confidence. *Social Psychology Quarterly*, *56*, 288–299. https://doi.org/10.2307/2786665

Özkan, T., Lajunen, T., & Kaistinen, J. (2005). Traffic locus of control, driving skills, and attitudes towards in-vehicle technologies (ISA & ACC). In *Proceedings of the 18th International Cooperation on Theories and Concepts in Traffic Safety (ICTCT)*.

Parasuraman, R., de Visser, E., Wiese, E., & Madhavan, P. (2014). Human trust in other humans, automation, robots, and cognitive agents: Neural correlates and design implications. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 58, pp. 340–344). SAGE Publications. https://doi.org/10.1177/1541931214581070

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

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*, 252–263. https://doi.org/10.1016/j.trf.2014.04.009

Payre, W., Cestac, J., & Delhomme, P. (2016). Fully automated driving: Impact of trust and practice on manual control recovery. *Human Factors*, *58*, 229–241.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891. https://doi.org/10.3758/BRM.40.3.879

Rajaonah, B., Anceaux, F., & Vienne, F. (2006). Trust and the use of adaptive cruise control: A study of a cut-in situation. *Cognition, Technology & Work*, *8*, 146–155. https://doi.org/10.1007/s10111-006-0030-3

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*, 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C., & Kovaleva, A. (2013). A short scale for assessing the Big Five dimensions of personality. *Methoden, Daten, Analysen*, *7*, 233–249.

Reynaud, E., El Khoury-Malhame, M., Rossier, J., Blin, O., & Khalfa, S. (2012). Neuroticism modifies psychophysiological responses to fearful films. *PLoS ONE*, *7*, Article e32413. https://doi.org/10.1371/journal.pone.0032413

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*, 1–36.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*, 1–28. https://doi.org/10.1037/h0092976

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*, 651–665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, *35*, 1–7. https://doi.org/10.1037/0003-066X.35.1.1

Rotter, J. B. (1990). Internal versus external control of reinforcement: A case history of a variable. *American Psychologist*, *45*, 489–493. https://doi.org/10.1037/0003-066X.45.4.489

Rudin-Brown, C. M., & Noy, Y. I. (2002). Investigation of behavioral adaptation to lane departure warnings. *Transportation Research Record*, *1803*, 30–37. https://doi.org/10.3141/1803-05

Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Reliability and age-related effects on trust and reliance of a decision support aid. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 48, No. 3, pp. 586–589). Sage Publications. https://doi.org/10.1177/154193120404800366

Satorra, A., & Bentler, E. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). SAGE Publications.

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment*, *63*, 506–516. https://doi.org/10.1207/s15327752jpa6303_8

Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T. L., Chen, J. Y., & Hancock, P. A. (2014). *A meta-analysis of factors influencing the development of trust in automation: Implications for human-robot interaction*. Aberdeen.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*, 377–400.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*, 23–74.

Schneider, P. P., & Vogt, C. A. (2012). Applying the 3M model of personality and motivation to adventure travelers. *Journal of Travel Research*, *51*, 704–716. https://doi.org/10.1177/0047287512451134

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, *26*, 207–231. https://doi.org/10.1207/s15326985ep2603&4_2

Starkweather, J. (2012). *Step out of the past: Stop using coefficient alpha; there are better ways to calculate reliability*. Benchmarks RSS Matters.

Steiger, J. H. (2016). Notes on the Steiger–Lind (1980) Handout. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 777–781. https://doi.org/10.1080/10705511.2016.1217487

Steiger, J. H., & Lind, J. C. (1980). Statistically-based tests for the number of common factors. *Handout for a presentation delivered at the meeting of the Psychometric Society*. IA.

Szalma, J. L. (2008). Individual differences in stress reaction. In P. A. Hancock & J. L. Szalma (Eds.), *Performance under stress* (pp. 323–357). Ashgate.

Szalma, J. L. (2009). Individual differences in performance, workload, and stress in sustained attention: Optimism and pessimism. *Personality and Individual Differences*, *47*, 444–451. https://doi.org/10.1016/j.paid.2009.04.019

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The five factor model of personality. *Journal of Experimental Psychology: Applied*, *17*, 71–96.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.
Thielmann, I., & Hilbig, B. E. (2015). Trust: An integrative review from a person–situation perspective. *Review of General Psychology*, 19(3), 249.
Ullman, J. B. (2013). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (pp. 681–785). Pearson.
Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors*, *51*, 281–291. https://doi.org/10.1177/0018720809338842
Wang, L., Jamieson, G. A., & Hollands, J. G. (2011). The effects of design features on users' trust in and reliance on a combat identification system. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 1, pp. 375–379). SAGE Publications.
Watson, D., & Clark, L. A. (1984). Negative affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin*, *96*, 465–490. https://doi.org/10.1037/0033-2909.96.3.465
West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). Guilford Press.

Johannes Kraus is a research assistant at the Human Factors Department at Ulm University in Germany. He received a Master of Science in psychology from the University of Mannheim in 2013.

David Scholz works at the Department of Human Factors at Ulm University. He received a Master of Science in psychology from the University of Ulm in 2019.

Martin Baumann is head of the Department of Human Factors at Ulm University. His main research interests are the psychological basis of human–machine interaction in different domains, mainly traffic, human–robot interaction, interaction with intelligent systems, and the development and validation of concepts of cooperative human–machine systems.

Date submitted: April 17, 2019
Date accepted: February 8, 2020

Article II: Scared to Trust? – Predicting Trust in Highly Automated Driving by Depressiveness, Negative Self-Evaluations and State Anxiety

Kraus, J., Scholz, D., Messner, E.-M., Messner, M., & Baumann, M. (2020). Scared to trust? – Predicting trust in highly automated driving by depressiveness, negative self-evaluations and state anxiety. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02917

Copyright: © 2020 Kraus, Scholz, Messner, Messner, and Baumann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 January 2020
doi: 10.3389/fpsyg.2019.02917

Scared to Trust? – Predicting Trust in Highly Automated Driving by Depressiveness, Negative Self-Evaluations and State Anxiety

Johannes Kraus[1], David Scholz[1], Eva-Maria Messner[2], Matthias Messner[3] and Martin Baumann[1]*

[1] Department of Human Factors, Institute of Psychology and Education, Ulm University, Ulm, Germany, [2] Department of Clinical Psychology and Psychotherapy, Institute of Psychology and Education, Ulm University, Ulm, Germany, [3] Department of Clinical and Health Psychology, Institute of Psychology and Education, Ulm University, Ulm, Germany

The advantages of automated driving can only come fully into play if these systems are used in an appropriate way, which means that they are neither used in situations they are not designed for (misuse) nor used in a too restricted manner (disuse). Trust in automation has been found to be an essential psychological basis for appropriate interaction with automated systems. Well-balanced system use requires a calibrated level of trust in correspondence with the actual ability of an automated system. As for these far-reaching implications of trust for safe and efficient system use, the psychological processes, in which trust is dynamically calibrated prior and during the use of automated technology, need to be understood. At this point, only a restricted body of research investigated the role of personality and emotional states for the formation of trust in automated systems. In this research, the role of the personality variables depressiveness, self-efficacy, self-esteem, and locus of control for the experience of anxiety before the first experience with a highly automated driving system were investigated. Additionally, the relationship of the investigated personality variables and anxiety to subsequent formation of trust in automation was investigated. In a driving simulator study, personality variables and anxiety were measured before the interaction with an automated system. Trust in the system was measured after participants drove with the system for a while. Trust in the system was significantly predicted by state anxiety and the personality characteristics self-esteem and self-efficacy. The relationships of self-esteem and self-efficacy were mediated by state anxiety as supported by significant specific indirect effects. While for depression the direct relationship with trust in automation was not found to be significant, an indirect effect through the experience of anxiety was supported. Locus of control did not show a significant association to trust in automation. The reported findings support the importance of considering individual differences in negative self-evaluations and anxiety when being introduced to a new automated system for individual differences in trust in automation. Implications for future research as well as implications for the design of automated technology in general and automated driving systems are discussed.

Keywords: trust in automation, anxiety, depressiveness, personality, self-evaluation, emotional states, trust formation, automated driving

OPEN ACCESS

Edited by:
Jan B. F. Van Erp,
University of Twente, Netherlands

Reviewed by:
Andrej Košir,
University of Ljubljana, Slovenia
Roope Raisamo,
University of Tampere, Finland

****Correspondence:***
Johannes Kraus
johannes.kraus@uni-ulm.de

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Psychology

Received: *02 August 2019*
Accepted: *10 December 2019*
Published: *23 January 2020*

Citation:
Kraus J, Scholz D, Messner E-M, Messner M and Baumann M (2020) Scared to Trust? – Predicting Trust in Highly Automated Driving by Depressiveness, Negative Self-Evaluations and State Anxiety. Front. Psychol. 10:2917. doi: 10.3389/fpsyg.2019.02917

INTRODUCTION

It is announced that in the next years, highly automated vehicles will become an affordable, everyday technology with a broad user group (Hörl et al., 2016). Commonly, the advent of automated vehicles is linked to positive outcomes like more efficient, safe, and comfortable driving (Kühn and Hannawald, 2015; Winkle, 2015). As the decision to use or not use driving automation will be made by human drivers in the next years, the associated advantages only come fully into play, if users decide to use this technology – and if so in an appropriate manner.

Trust in automation has been found to be a major subjective prerequisite for behavioral decisions in the interaction with various automated technological systems (e.g., Lee and Moray, 1994; Muir and Moray, 1996; Lewandowsky et al., 2000; Biros et al., 2004; Sanchez et al., 2004) and automated driving systems in specific (e.g., Dzindolet et al., 2003; Rajaonah et al., 2006; Hergeth et al., 2016; Payre et al., 2016). It is essential for safe and efficient use of automated driving that a calibrated trust level in accordance with a system's capabilities and performance (e.g., Muir, 1987) is facilitated as it diminishes inappropriate system use. Conversely, if users' trust is not calibrated, the dangers of disuse (underachieving the full potential of a system) and misuse (overstretching the capabilities of a system) of the system are increased (Parasuraman and Riley, 1997). To prevent these negative outcomes and in order to facilitate trust calibration, a thorough understanding of the psychological processes associated with trust formation and calibration is essential (e.g., Lee and See, 2004; Hoff and Bashir, 2015; Kraus et al., 2019b). Especially the relation of personality traits and emotional states to individual differences in trust in automation are a promising direction to gain an understanding for the underlying psychological mechanisms of trust formation and calibration (e.g., Hoff and Bashir, 2015). These processes could be addressed to personalize the design of automated technology in order to enhance trust calibration (e.g., prior information about the system, training, system functionality, and user interfaces).

Personality as a collection of "mechanisms within the individual that are organized and relatively enduring and that influence his or her interactions with [. . .] environments" (Larsen and Buss, 2005, p. 4), might shape how different persons perceive and experience a technological system. For example, on a theoretical level, personality traits like risk aversion (Glöckner and Hilbig, 2012) should influence how potential costs and benefits of automated technology are weighted and interpreted. In a similar manner, personality traits in regard to the self-assessment of one's abilities (e.g., self-efficacy; Bandura, 1991) might influence as how capable people see themselves to use new technology in a functional manner. In line with this general idea of a role of personality for trust formation, personality traits like neuroticism (Szalma and Taylor, 2011) and need for cognition (Kraus et al., 2019a) have been found to be associated with differences in trust in automation.

Furthermore, personality traits play an important role in the individual tendency to experience emotional states (e.g., Larsen and Ketelaar, 1991; Letzring and Adamcik, 2015). For example, the link between neuroticism and the tendency to experience negative affect is well supported (e.g., Larsen and Ketelaar, 1991). Emotional states have also been found to predict differences in trust in automation. For example, Stokes et al. (2010) found higher trust levels in persons experiencing positive affect than those experiencing negative affect.

Another theoretically meaningful emotional state for the explanation of individual differences in trust in automation is the experience of anxiety in the face of new technology. Anxiety might prevent rational choice in interactions with new technology and result in inefficient or even dangerous decisions (Peters et al., 2006). Such irrational decision-making would result in a reduction of an overall positive effect of the advantages of automated driving (e.g., increased road safety and more efficient and economic traffic system; see for example Kühn and Hannawald, 2015; Winkle, 2015). In line with this, in their review on interpersonal trust, Thielmann and Hilbig (2015) theoretically link the individual experience of anxiety (trait anxiety) to trust behavior. At this point, the relation between state anxiety and trust in automation has not yet been investigated.

In this research, the relationship between anxiety, when being introduced to a new automated driving system, and the initial level of learned trust after a first interaction with the system is investigated in a driving simulator study. Furthermore, the role of the personality variables depressiveness (Beck et al., 1961), self-esteem, self-efficacy and locus of control (e.g., Chamorro-Premuzic et al., 2008; Xiao et al., 2014), for the experience of anxiety and individual differences in trust in automation is investigated. The underlying idea of this research is that depressiveness and negative self-evaluations lead to a more anxious perception of and approach to an unfamiliar technological system. The resulting increased level of anxiety leads to decreased trust levels in the system. More specifically, the underlying research hypothesis is that the effect of depressiveness, self-esteem, self-efficacy, and locus of control (LOC) on trust in automation is mediated by an experience of anxiety in the face of unfamiliar automated technology (see **Figure 1**). In the following, the theoretical background of the study is provided along with an integrated derivation of study hypotheses.

THEORETICAL BACKGROUND

Factors Influencing the Calibration of Trust in Automation

Lee and See (2004) define trust in automation as "the attitude that an agent will help to achieve an individual's goal in a situation characterized by uncertainty and vulnerability" (p. 51). The psychological process, in which trust in automation is built up prior and during the interaction with an automated system is subject to several theoretical reviews and frameworks (e.g., Lee and See, 2004; Schaefer et al., 2014; Hoff and Bashir, 2015)

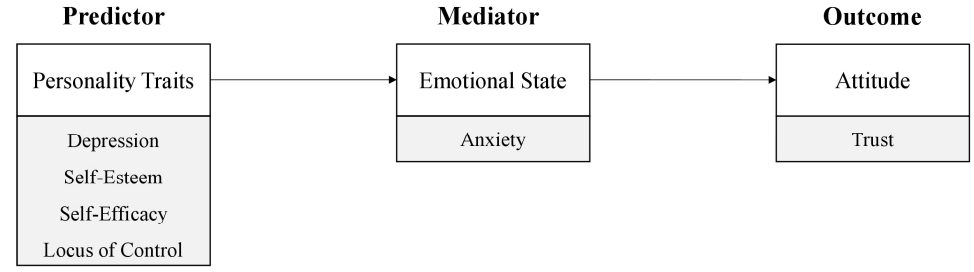


FIGURE 1 | Investigated relationship of personality, state anxiety and trust in automation. It is hypothesized that the relationship of depression, self-esteem, self-efficacy, and locus of control to initial learned trust in automation is mediated by the extent of an individual experience of anxiety at the time when a new and unfamiliar automated driving system is introduced.

as well as empirical research (e.g., Beggiato et al., 2015; Hergeth et al., 2016; Kraus et al., 2019a,b). A multitude of factors within the system, the person and the situation have been linked to trust calibration processes (see Hancock et al., 2011 for an overview). On the side of the user, personality (e.g., Merritt and Ilgen, 2008; Merritt et al., 2015; Kraus et al., 2019a) as well as emotional states (e.g., Stokes et al., 2010; Merritt, 2011) have been found to influence trust formation and explain inter-individual differences in trust toward specific automated systems. In support of this notion, reviews and meta-analyses in human-machine interaction propose personality to be an important predictor for inter-individual differences in trust in automation (e.g., Lee and See, 2004; Hancock et al., 2011; Hoff and Bashir, 2015; Schaefer et al., 2016; Hock et al., 2018). To extend the understanding of the role of emotional states and personality for the dynamic formation of trust in automation, this research investigates the relationships of depressiveness, negative self-evaluations, and anxiety to learned trust in an automated driving system. The hypothesized mediation process of this research (Personality traits → emotional state → trust in automation) gained support in previous studies in the domain of interaction with technology (e.g., Jacques et al., 2009).

Anxiety and Trust in Automation

Emotional states were defined as "a state of physiological arousal and of a cognition appropriate to this state of arousal" (Schachter and Singer, 1962, p. 380). The impact of emotional states on judgments and behavior has been subject to a great deal of research (e.g., Forgas and George, 2001; Dunn and Schweitzer, 2005; Forgas and East, 2008). Prominently, the affect-as-information model (Schwarz and Clore, 1988) proposes that people tend to misattribute a momentary experienced mood to their judgments by asking themselves how they feel about an object of evaluation. A second model proposing effects of emotional states to judgments is the affect infusion model (AIM; Forgas, 1995). Based on these theories, Dunn and Schweitzer (2005) propose that emotional states might be used as an important source of information for building up interpersonal trust. In line with this reasoning, emotional states were empirically found to be related to trust in automated systems (e.g., Merritt, 2011). Stokes et al. (2010) showed that the impact of emotional states on trust in an automated system was strongest at the beginning of the interaction when the system is unknown and diminished with repetitive system use.

While prior research mostly focused on positive and negative affect (Stokes et al., 2010), this research investigates the relation of state anxiety to trust in automation. State anxiety is conceptualized as a momentary emotional state, which is accompanied by a consciousness feeling of tension and uneasiness and a higher activity of the autonomic nervous system (Spielberger, 1966; Horikawa and Yagi, 2012). Anxiety has been found to affect information processing of ambigiuous (e.g., Mathews and Mackintosh, 1998) and threatening stimuli (Mathews, 1990). Similarly, trusting behavior refers to a risky choice of depending on another party (e.g., Thielmann and Hilbig, 2015) in a situation characterized by uncertainty, vulnerability and the inability to control the actions of another party (e.g., Mayer et al., 1995; Lee and See, 2004). Therefore, the individual level of trust in automation should essentially be influenced by perceived risk and the associated experience of anxiety. In line with this, a theoretical link between the individual disposition to experience anxiety, risk/loss aversion and trust has been proposed for interpersonal contexts (e.g., Thielmann and Hilbig, 2015). Also, in the context of automated driving, anxiety has been investigated as a dependent variable additional to trust in automation. Du et al. (2019) report a significant negative relationship between trust and anxiety although this relationship was not in the focus of their research and anxiety was measured as an attitude instead of an emotional state.

In this research, state anxiety is hypothesized to be associated with decreased initial trust in an unfamiliar, newly introduced automated driving system. Anxiety might lead to an overrepresentation of potential risks in the assessment of automated driving systems and thereby reducing trust in the systems. Additionally, in line with the affect-as-information model and the AIM the experience of anxiety in the face of a new system might be misattributed to the judgment of restricted trustworthiness of the system and therefore lead to lower trust. In the interaction with new (and potentially risky) technology, the experience of increasing levels of anxiety might progressively prevent rational choice and result in inefficient or even dangerous decisions (Peters et al., 2006). The theoretical importance of anxiety for an understanding of the psychological mechanisms in the formation and calibration of trust in automation is

further supported by its incorporation in theoretical models of technology use (e.g., Osswald et al., 2012) and empirical support for its role in the interaction with automated systems (e.g., Nomura et al., 2008). Based on this theoretical association between the experience of anxiety and trust formation and in line with predictions of the affect-as-information model and the AIM, it is hypothesized:

Hypothesis 1 (H1): Higher state anxiety at the time when an unfamiliar automated driving system is introduced is associated with lower trust in the system.

Personality traits have been linked to the experience of emotional states in a long tradition of psychological research (e.g., Larsen and Ketelaar, 1991; Letzring and Adamcik, 2015). In line with this, anxiety has been found to be predicted by inter-individual differences in personality variables – besides others by depression (Nutt, 1999) and self-evaluations (e.g., Chamorro-Premuzic et al., 2008; Xiao et al., 2014). To gain a better understanding of the emergence of anxiety differences when being introduced to or initially interacting with unfamiliar automated technology, we investigated the role of these personality variables for individual differences in the extent of anxiety when being introduced to a new automated system.

Depression and Trust in Automation

Depression is one of the most widespread psychological disorders in many societies. For example, in the United States, Major Depressive Disorder (MDD) is with a life-time prevalence of about 17% and a 12-months-prevalence of 7% (Kessler et al., 2005a) one of the most prevalent mental disorders (Kessler et al., 2005b). Not only clinical depression is affecting individuals, but subclinical depression is also widely common (Horwath et al., 1992; Cuijpers et al., 2004). An individual is referred to as subclinical depressed as he or she experiences clinically relevant depressive symptoms but does not meet the diagnostic criteria (e.g., Cuijpers et al., 2014).

As depressed individuals were shown to perceive and accept the use of technology differently from other user groups (e.g., Baumeister et al., 2014; Ebert et al., 2015), it is likely that also (subclinical) depressed persons show different attitudes toward the use of technology as compared to non-depressed individuals. Depression is characterized by a lack of interpersonal trust (Lester and Gatto, 1990; Kim et al., 2012) and this inability to build trust to other people might impair the ability to build trust into an automated driving system as well. In Hoff and Bashir's model of trust in automation (2015) it is stated that the internal variability in trust is rooted in (1) self-confidence, (2) subject-matter expertise, (3) mood, and (4) attention capacity. In (subclinical) depression, besides others self-confidence, mood, psychomotoric abilities, decision-making and concentration are distorted (e.g., Fried et al., 2016; Fried, 2017) and therefore, the ability to build trust in automated systems might be impaired. We, therefore, hypothesize that depressiveness (the amount of experiencing depression symptoms) is associated with lower trust in an automated driving system:

Hypothesis 2 (H2): Depressiveness is negatively related to trust in an automated driving system.

Mediation of the Relation Between Depressiveness and Trust in Automation by Anxiety

The experience of depressive symptoms is commonly associated with increased anxiety (e.g., Murphy et al., 2004; Cummings et al., 2014). The same holds true for subclinical depression (e.g., Lovibond and Lovibond, 1995). In fact, self-report questionnaires for depression and anxiety show strong correlations in both clinical and sub-clinical samples (Sowislo and Orth, 2013). Following from this association, the more depressive symptoms someone experiences the more anxious he or she should be when being introduced to a new technological system and accordingly trust an unfamiliar automated system less. The relation of depression and anxiety to dysfunctional interaction with technology has been already investigated in regard to smartphone addiction (Elhai et al., 2016). In accordance, it is hypothesized:

Hypothesis 3 (H3): The effect of depressiveness on trust in an automated driving system is mediated by the experience of anxiety at the time of introduction of the system.

Traits Related to Negative Self-Evaluation and Trust in Automation

Besides depressiveness, personality traits related to negative self-evaluations were found to influence trust in other persons (e.g., Chamorro-Premuzic et al., 2008; Xiao et al., 2014). Also, for trust in automation, a relevance of negative self-evaluations is supported by the strong empirical evidence for the effect of situational self-confidence on differences in trust in automation (e.g., Lee and Moray, 1992, 1994; Moray and Inagaki, 2000). Self-confidence was conceptualized as a self-assessment of the capability to sucessfully perform a task with a technical system (e.g., Briggs et al., 1998). In this research, the role of self-esteem, self-efficacy, and LOC, which play an essential role in the emergence of self-confidence judgments, for individual differences in initial trust in an automated driving system is investigated.

Harter (1993) defined self-esteem as "the level of global regard one has for the self as a person" (p. 88) affecting how worthy, capable and successful people perceive themselves to be (Coopersmith, 1981, cited in Herz and Gullone, 1999). Self-esteem has been shown to be related to self-confidence (e.g., Owens, 1993).

Similarly, self-efficacy builds a personality basis for situational self-confidence and can be understood as a trait describing a broad, trans-situational form of self-confidence in one's ability to perform well in various tasks (e.g., Locke et al., 1984; Schunk, 1991; Locke and Latham, 1994). General self-efficacy was defined by Bandura (1991) as "people's beliefs about their capabilities to exercise control over their own level of functioning and over events that affect their lives" (p. 257).

The mentioned evidence for the relationship of trust in automation and situational self-confidence points into the direction that both self-esteem and self-efficacy (as personality facets affecting self-confidence) influence the self-evaluation of one's own ability to successfully interact with automated systems. Therefore, people with comparably high levels of self-esteem and self-efficacy should perceive themselves as capable of using this technology safely and in line with their needs and goals and thus trust it more.

Hypothesis 4 (H4): Self-esteem is positively related to trust in an automated driving system.

Hypothesis 5 (H5): Self-efficacy is positively related to trust in an automated driving system.

Locus of control describes the disposition to perceive oneself as capable of controlling the outcome of an event or action (e.g., Rotter, 1966, 1990; Ajzen, 2002). The trait is formed by two antagonistic factors – internal (perception of an own ability to control events) and external LOC (outcomes are beyond own control). In line with Phares (1976), internals show more effort to control their fate, seek more information and use it better. In turn, several studies showed that internals are more efficient problem solvers (e.g., DuCette and Wolk, 1973). As a consequence, internals, through their decision-making, increase the likelihood that their experience is in line with their internal view of things (Spector, 1982). In contrast, an external locus of control has been related to learned helplessness (Seligman, 1972; Hiroto, 1974; Cohen et al., 1976). In line with these findings, it can be concluded that people with an internal LOC should be more willing to trust automation as they perceive themselves to be more in control of making functional decisions leading to positive results as compared to those with an external locus of control who do not perceive that they can control their fate. The role of LOC in the domain of driving has been investigated before with conflicting results (e.g., Özkan et al., 2005). Based on our theoretical reasoning, the following hypothesis is formulated:

Hypothesis 6 (H6): An internal locus of control (H4.3) is positively related to trust in an automated driving system.

Mediation of the Relation Between Negative Self-Evaluations and Trust in Automation by Anxiety

Self-esteem is associated with an increased feeling of safety and security and serves an anxiety-buffering function (e.g., Greenberg et al., 1992), which is a central assumption of the Terror Management Theory (e.g., Greenberg et al., 1986). In a meta-analysis by Sowislo and Orth (2013), a negative correlation between self-esteem and anxiety was supported both for measuring the constructs simultaneously and in a longitudinal manner. Also, in regard to an unfamiliar automated system, people with a higher level of self-esteem should feel more secure about themselves and thus experience less anxiety of interacting with the system and trust it more in the subsequent interaction. Therefore, it is hypothesized that self-esteem is negatively related to an experience of anxiety when being introduced to a new automated system and that state anxiety mediates the positive relationship of self-esteem to learned trust in automation (H7.1).

For self-efficacy anxiety plays a central role (e.g., Bandura, 1988) as a threat is not merely determined by external dangers of a situation but rather a combination from the situation and both an assessment of dangers and perceived coping capability. Consequently, lower self-efficacy is hypothesized to be associated with a higher degree of anxiety in the face of a forthcoming interaction with an unfamiliar technological system, whereas anxiety mediates the positive effect of self-efficacy to trust in such a system (H7.2). In support of this hypothesis, bivariate correlations (with small to medium effect sizes) between computer self-efficacy, trust, and internet anxiety have been reported (Thatcher et al., 2007).

Third, LOC has consistently been found to be related to anxiety in the way that internals experience less anxiety than externals (e.g., Spector, 1982). It can be followed that a person with an internal LOC should experience less anxiety when being introduced to a new technology he or she did never interact with before as compared to a person with external LOC. The latter will tend to perceive this situation beyond his or her control and consequently be more anxious about its consequences. Accordingly, it is hypothesized that an internal LOC is associated with less and an external LOC with higher anxiety when being introduced to a new driving automation system. Subsequently, internals should trust such a system more than externals and this relationship is mediated by anxiety (H7.3).

Following from this, in this study it is hypothesized, that the relationships of the investigated personality variables to trust in automation are mediated by the experience of anxiety when the automated system is first introduced. Support for this mediation mechanism of the influence of traits through the experience of affective states has been found in other studies in the domain of technology interaction as well (e.g., Jacques et al., 2009).

Hypothesis 7 (H7): The effects of self-esteem (H7.1), self-efficacy (H7.2) and locus of control (H7.3) on trust in an automated driving system are mediated by the experience of anxiety at the time of introduction of the system.

The hypotheses were investigated in a driving simulator study with a correlative research design. Depression and personality traits were measured before an automated driving system was introduced. Also, state anxiety of participants was assessed, together with positive and negative affect as control variables to investigate the incremental effect of anxiety on trust in automation. Subsequently, participants were introduced to the system and drove with the automation in the driving simulator and evaluated their trust in the system afterwards.

MATERIALS AND METHODS

This study investigates the relationship of depression, self-esteem, self-efficacy, LOC, and the level of anxiety when being introduced to a new automated driving system to trust in the system in a correlative design. Participants answered questionnaires with the personality and state scales and were then introduced to an automated driving system. After a practice trial in manual mode to get used to the driving simulator, participants interacted with the automated driving system they had been introduced to earlier for about 5 min. Afterwards, participants rated their trust in the system.

Sample

Participants were recruited at Ulm University with mailings and posters and were compensated with 10€ or study credit for their participation. After exclusion of two participants (automation error and non-compliance with study instructions), the final sample consisted of $N = 47$ participants, of which 27 were female (57%). Participants were on average $M = 27.45$ years old ($SD = 9.75$) and held their driving license for $M = 9.38$ years ($SD = 9.16$).

Procedure

The study was conducted in the driving simulator lab at the Human Factors department at Ulm University. After being welcomed, a brief study introduction was given and informed consent was obtained from all participants. Participants received a series of personality questionnaires and questionnaires regarding their emotional states (see below). Afterward, they were introduced to the driving simulator, settings were adjusted, and participants were introduced to the automated driving system and the driving task. The system was introduced with a manual and an automated mode, while the latter enables automated driving on the motorway under certain conditions. In a practice trial participants could familiarize with the settings of the simulator and driving under these conditions in manual mode. After 4–5 min, they were asked by the investigator if they felt safe with driving in the simulator. If they indicated to not feel safe, they could continue the practice for a while, otherwise the study procedure continued. In the instruction text and in the practice trial, transparency of the automation was manipulated in two level (low vs. high transparency), which was not in the scope of this research (Kraus et al., 2019b). Afterwards, participants were instructed that they would now drive with the automated driving system; they had been introduced to earlier. After this instruction, the participants used the automated driving system for the first time in a test drive of about 5 min. Participants were instructed to adhere to traffic rules and not exceed a speed limit of 130 kilometers per hour. The test drive started at the entrance of a two-lane motorway and the vehicle was set in manual mode for about 1 min. Afterward, participants were instructed to activate the automated mode. In the course of the remaining time, participants went through several automated overtaking maneuvers. In between the system conducted a take-over request and asked the participant to drive manually again for a while due to road works. After the latter automated driving was indicated to be available again and participants were asked to reactivate automated driving. During the simulated drive, an attention assist was implemented in the simulated drive, which impelled participants with a repeated beep sound (every 60 s) to put the hands on the steering wheel and to direct their gaze on the road. Manual control could be taken over at any time. Half of the participants experienced a malfunction of the system during the test drive as part of a second manipulation also not in the scope of this research. Data analysis was conducted with the complete sample and the trust measurement after the test drive was used as the outcome variable (test of preconditions below).

The original study procedure continued afterward with a subsequent experimental drive interrupted by several questionnaires, which was part of another study reported elsewhere. After this drive, participants answered further questionnaires, amongst others on their demographics. Subsequently, participants were compensated for their participation, debriefed, and dismissed. In total, the study took 75 min. The reported research complied with the Declaration of Helsinki.

Materials

Apparatus

The study was conducted in the driving simulator of the Human Factors department at Ulm University (see **Figure 2**). The simulator is equipped with a car interior mock-up with basic design elements and a windshield frame. For front and side view three 1920 × 1200px video projectors fill three screens of 3.3 × 2.1 m each and realize a 200° viewing angle. Rearview is implemented by two displays replacing the side mirrors and one display replacing the rearview mirror (each 7" with 800 × 400px; 16:9). For this study, an information display was located in the center console indicating the current state of the system and equipped with a touch interface to engage and disengage the automation.

Questionnaires

Depressiveness was assessed with the German questionnaire Becks Depressions Inventory II (BDI-II, Hautzinger et al., 2009),



FIGURE 2 | Driving simulator of the Human Factors Department at Ulm University. Photo by: H. Grandel/Ulm University.

which covers thoughts, feelings, behaviors, and symptoms of a major depression with 21 items to be answered on a four-point Likert scale. LOC was investigated with the locus of control Scale (TOR-IE; Rotter, 1966) with 23 items reflecting both internal and external LOC. Higher values on this scale indicate a more external LOC. Self-esteem was assessed with the Rosenberg Self-Esteem Scale consisting of 10 items (Ferring and Filipp, 1996). Self-efficacy was measured with the German version of the General Self-Efficacy Scale (GSE, Jerusalem and Schwarzer, 1999). For the assessment of positive and negative affect, the Positive and Negative Affect Schedule (PANAS, Krohne et al., 1992) was used measuring the current affect with ratings on ten positive and ten negative adjectives with a five-point Likert scale. State anxiety was measured with a translated short version of the State-Trait Anxiety Inventory (STAI-6, Marteau and Bekker, 1992), in which participants had to assess their current feelings with six statements. Trust in automation was assessed with a German shortened version of the Automation Trust Scale (ATS; Jian et al., 2000) with seven items (see also Kraus et al., 2019b). The scale was designed to provide an economic measurement of trust in automation as a unidimensional construct. Scale means, standard deviations, Cronbach's α along with the correlations of the scale values can be investigated in **Table 1**.

Statistical Procedure

Statistical analysis was conducted with R (version 3.4.3) and the R package lavaan (Rosseel, 2012; version 0.6–2). To test the hypothesized relationships, bivariate Pearson correlations were inspected (H1, H2, H4, H5, and H6) and path modeling was utilized (H3 and H7). For the path models, as recommended by Preacher and Hayes (2008), the residuals of constructs at mediating variable levels (states) were allowed to covary to avoid biased standard errors and model misspecification. For testing the study hypotheses, direct and specific indirect effects were investigated. Direct effects test the relationships between adjacent variables and indicate how well a variable directly predicts another variable (e.g., personality variable → emotional state). Indirect effects test the hypothesized mediations. The two types of effects were investigated with the percentile bootstrap procedure, as recommended by Hayes (2018), providing a 95% confidence interval (CI) for the estimations. The null hypothesis was rejected if zero was not included in the CI (Taylor et al., 2008; Hayes, 2009). We used 5000 bootstrap samples as recommended by Hayes (2009). In the following, standardized direct and indirect effects are reported.

Due to the mentioned manipulation of an independent variable (transparency; not in the scope of this study) in between the measurement of the personality and state variables and the measurement of trust in automation, to rule out any biases by group effects (interaction of the traits/states and the independent variable), a series of general linear models was conducted. As no such interaction was found to be significant, all reported statistics and indices can be viewed as unbiased by the experimental manipulations and can therefore be interpreted without problems.

Concerning the preconditions of the applied statistical procedures, one participant was identified as a univariate outlier for both depressiveness and self-esteem ($|z| > 3.29$) as well as a multivariate outlier as indicated by Mahalanobis distance. A closer inspection of the answering pattern showed a BDI score of 37 and a self-esteem score of seven for the individual, both values in a valid range indicating depression. As no further reason for exclusion could be identified and an exclusion of this participant would lead to an invalid restriction of the sample variance we did not exclude the participant. Furthermore, several of the investigated variables were not normally distributed as indicated by significant Shapiro–Wilk tests (depression, self-esteem, self-efficacy, negative affect and trust). As the Pearson correlation coefficient has been shown to be robust to violations of normality (e.g., Havlicek and Peterson, 1976; Bishara and Hittner, 2012) the reported correlations the reported correlations are valid to be interpreted. In the same way, the reported bootstrapped effects of the path models are not affected by distribution assumptions.

RESULTS

Relationship Between Emotional States and Trust in Automation

Hypothesis 1 stated that higher levels of anxiety are related to lower trust in an automated driving system. Inspection of the bivariate correlation of state anxiety and trust in automation revealed a strong negative correlation of $r = -0.400$ ($p < 0.001$) between state anxiety and trust in automation in support of H1 (see **Table 1**). Also, in terms of the investigated path models, anxiety showed a significant negative direct effect on trust in automation, $\beta = -0.413$ [−0.697; −0.062] (CIs from the model with Depression, **Figure 3A**), likewise supporting H1. To evaluate the extent of incremental variance of anxiety in the explanation of trust in automation above those emotional states already investigated (Stokes et al., 2010), correlations between trust in automation and positive and negative affect were inspected (see **Table 1**). Although in the same direction as in earlier research, both correlations were found to be non-significant in this study. Neither did both states mediate the effect of the investigated personality variables on trust in automation as indicated by non-significant indirect effects in all analyses (see **Figures 3A–D**). Taken together, in this study anxiety predicts trust in automation better than both positive and negative affect.

Relationships Between Personality Variables and Trust in Automation

Also, hypotheses on the direct effects of the investigated personality variables on trust were tested with bivariate correlations (see **Table 1**). Regarding H2, the analysis revealed that depression was not directly related to trust ($r = -0.262$; $p > 0.05$). However, in the face of the comparably strong effect size and the significant indirect effect of depression on trust through anxiety (see below, **Figure 3A**), H2 gains some support in this study. Regarding the hypotheses on the relationships of

TABLE 1 | Aggregated scale values (e.g., means or sum scores according to respective scale logic) standard deviation, internal consistencies (Cronbach's α) and correlations of the investigated personality and state variables and trust in automation (correlations are based on scale means).

		Score	*SD*	α	1	2	3	4	5	6	7
1	BDI	4.98	5.92	0.881	–						
2	LOC	10.28	4.11	0.732	0.172	–					
3	Self-esteem	23.38	4.91	0.889	−0.622**	−0.179	–				
4	Self-efficacy	30.21	4.22	0.872	−0.481**	−0.446**	0.618**	–			
5	STAI	35.60	7.29	0.704	0.466**	0.315*	−0.606**	−0.489**	–		
6	PA	33.32	5.73	0.843	−0.055	−0.166	0.171	0.297*	−0.106	–	
7	NA	12.62	2.64	0.696	0.312*	0.212	−0.540**	−0.412**	0.400**	−0.135	–
8	Trust	5.26	0.99	0.880	−0.262	−0.218	0.370*	0.291*	−0.400**	0.229	−0.107

*BDI = depressiveness; LOC = locus of control; STAI = state anxiety; PA = positive affect; NA = negative affect; Higher values in LOC indicate an external LOC; *p < 0.05; **p < 0.01.*

traits reflecting negative self-evaluations and trust in automation, both self-esteem (H4; $r = 0.370$; $p < 0.05$) and self-efficacy (H5; $r = 0.291$; $p < 0.05$) showed the hypothesized significant positive correlation with trust (see **Table 1**). The correlation between LOC and trust, however, did not reach significance, thus contradicting H6 ($r = -0.218$; $p > 0.05$).

Mediation of the Effects of Personality Variables on Trust in Automation by State Anxiety

To test the hypothesized mediations of the effects of the investigated personality variables on trust in automation by state anxiety (H3 and H7), parallel-mediation models were computed separately for each trait (see **Figure 3**). Separate models were calculated to prevent underestimation of the true relationships as a result of intercorrelations of the investigated traits (Hayes, 2018). Positive and negative affect were included on the state level as control variables to investigate if state anxiety provides variance explanation in the outcome trust in automation additional to these two emotional states.

As a statistical test for H3 and H7, indirect effects were investigated. As shown in **Figure 3**, in line with the hypotheses the indirect effects of depression (H3; β = −0.193 [−0.373; −0.007]; **Figure 3A**), self-esteem (H7.1; β = 0.251 [0.025; 0.480]; **Figure 3C**) and self-efficacy (H7.2; β = 0.202 [0.017; 0.401]; **Figure 3D**) on trust in automation via state anxiety were found to be significant. For self-esteem and self-efficacy (see **Figures 3C,D**), this effect can be interpreted as a mediation, as the correlation between self-esteem and trust is significant on their own (see **Table 1**). For depression, interestingly (see **Figure 3A**), this finding has to be interpreted as a mere indirect effect as the direct correlation between the trait and trust was not significant (see **Table 1**). In the same way as LOC did not show any effect on state anxiety in the first place, no significant indirect effect on trust in automation could be observed, which is in contradiction to H7.3 (β = −0.130 [−0.363; 0.007]; see **Figure 3B**).

Relative Importance of the Investigated Mediation Paths via State Anxiety

In order to test the relative importance of the investigated personality variables, a mediation model considering all traits together was calculated (see **Figure 4**). The results revealed that only the direct effect of self-esteem on state anxiety remains significant when all investigated personality variables are considered together in one model (β = −0.459 [−0.800; −0.019]). As a consequence, only the indirect effect of self-esteem to trust in automation mediated by state anxiety was found to be significant in the simultaneous path model (β = 0.184 [0.002; 0.435]). While this finding does not restrict the findings of the separate analyses (as multicollinearity in related trait variables is to be expected), it underlines the special importance of lower levels of self-esteem for an enhanced experience of anxiety in contact with automated driving systems leading to decreased levels of trust in such an automated system.

DISCUSSION

This research tested the hypotheses that depression and personality variables related to depressiveness and self-evaluation influence the experience of anxiety in the face of a newly introduced highly automated driving system. Furthermore, it was investigated, if increased anxiety in this situation is associated with reduced trust in the system and if anxiety mediates the relationships of the investigated personality variables on trust.

Summary of Results

Taken together, five of the seven hypotheses proposed in this research were substantially supported by the findings of this study (H1, H3, H4, H5, and H7). Additionally, H2 was partly supported. Overall, the reported study provided substantial empirical evidence for a negative relationship of the extent of anxiety when being introduced to an automated system on subsequent levels of trust in automation (H1). Those who experienced more anxiety at the time when they were introduced to a new and unfamiliar automated driving system showed lower trust levels after some interaction with the automated driving system. Furthermore, self-esteem (H4) and self-efficacy (H5) have been found to significantly predict learned trust in an automated system. Most interestingly, these relationships are mediated by the experience of anxiety at the time when the automated system was introduced providing substantial support

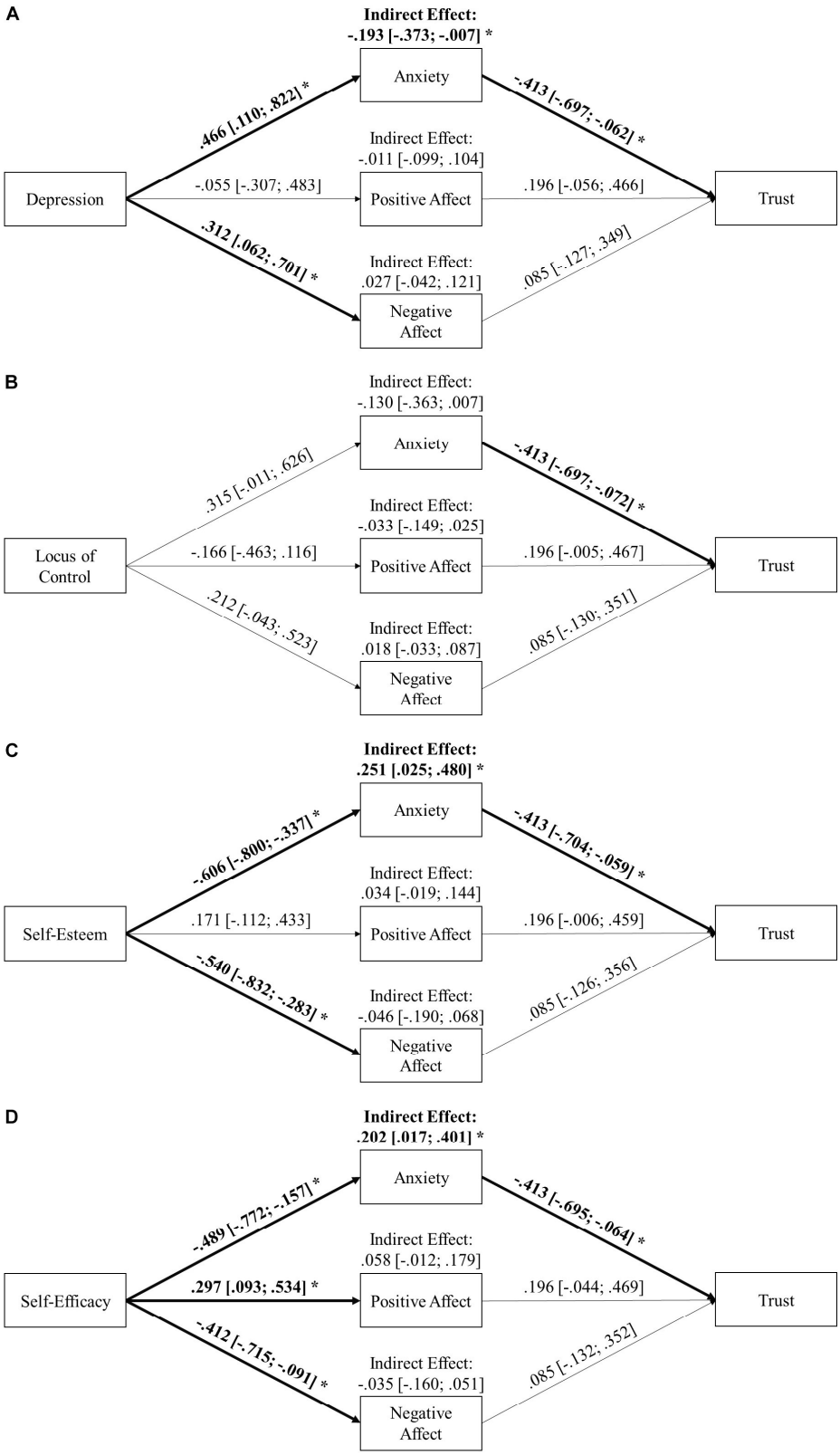


FIGURE 3 | Standardized direct and indirect effects of the personality variables on trust in automation mediated by the three emotional states. Separate path models were calculated for each personality variable (**A**: Depression, **B**: Locus of Control, **C**: Self-Esteem, **D**: Self-Efficacy). Significant effects ($\alpha = 0.05$) are indicated by an asterisk.

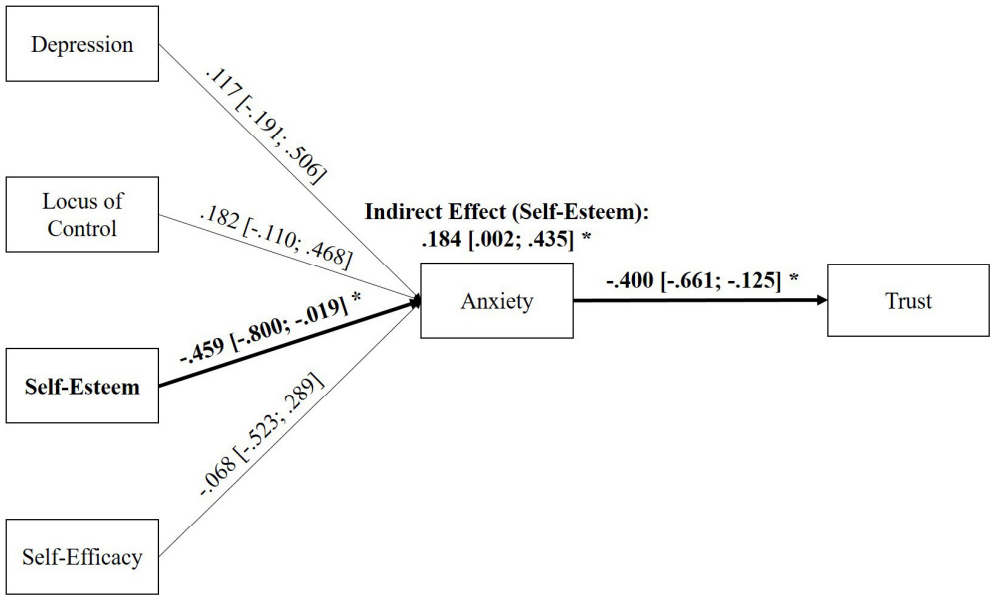


FIGURE 4 | Standardized direct and indirect effects of the personality variables on trust in automation mediated by state anxiety if all personality variables are considered in one path model together. Significant effects ($\alpha = 0.05$) are indicated by an asterisk. Only the indirect effect from self-esteem on Trust was significant while the indirect effects for the remaining personality variables were not.

for a role of anxiety for the formation of trust in automated driving systems. The prediction of trust by anxiety and the associated mediated indirect effects by anxiety were substantially higher than those by positive or negative affect. While there was no direct correlation between depression and trust in automation, a significant indirect effect supports the role of the individual tendency to experience depressive symptoms for trusting a technological system (H2). On the other hand, for LOC neither a direct nor an indirect effect on trust in automation could be found (H6). In the combined path model with all investigated personality traits, the special importance of self-esteem became apparent. Before we discuss our findings in detail, in summation of the reported findings, it can be concluded that anxiety is an important predictor for differences in trust in automation and that an enhanced experience of anxiety is well predicted by personality variables related to feelings of depression and self-evaluation. This finding underlines the importance of individual differences in users in the perception of their capability to use an automated system on the one hand for the experience of emotions when being confronted with unfamiliar automated technology and on the other hand for their subsequent trust in this system. The psychological mechanism supported in the presented study (personality → emotional states → trust in automation) helps to gain a better theoretical understanding for trust formation and calibration processes and provides starting points for the design of measures to facilitate trust calibration during system familiarization and early system use.

Theoretical Implications

The presented study supports a mediation of the effect of individual differences of users to learned trust after the first interaction with an automated driving system by the individual experience of anxiety when being introduced to the system at first.

Role of Anxiety for Trust in Automation

In line with models proposing an influence of emotional states on judgments (Forgas, 1995; Dunn and Schweitzer, 2005), this study provides empirical support for a negative relationship of anxiety to trust in an automated driving system. Trust as an attitude that guides decisions in situations of risk and uncertainty is influenced by the individual level of anxiety in this new and unfamiliar situation. This research enhances the understanding of emotional processes at play in the phase of getting to know and understand a newly introduced automated system. The moderate to high effect size of the relationship between anxiety and trust in automation, explaining 16% of the variance in trust, further highlights the importance of internal situational factors, as proposed in trust in automation models (e.g., Hoff and Bashir, 2015; Kraus et al., 2019b). The role of anxiety for the formation of trust toward a new and unfamiliar automated system underlines the affective facet of trust in automation (e.g., Lee and See, 2004). The psychological mechanism relating the experience of anxiety to trust in automation might be further investigated in future research. Interesting starting points might be processes associated with the subjective representation of risks and dangers in relation to advantages and gains provided by the automated system. Interestingly, the relationship of anxiety with trust in the system was stronger than the relationship of positive or negative affect with trust, which were investigated before (Stokes et al., 2010; Merritt, 2011). This strong relationship of anxiety and trust might be rooted in the theoretical relation between anxiety, the individual disposition to evaluate negative

outcomes like risks and losses (risk/loss aversion), and trust (e.g., Thielmann and Hilbig, 2015). Future studies might investigate if the influence of anxiety also diminishes after time like it has been found for positive and negative mood (Stokes et al., 2010).

Role of Personality for the Experience of Anxiety and Trust in Automated Technology

Another goal of the presented study was an investigation of the influence of depressiveness and a tendency for negative self-evaluations (self-esteem, self-efficacy, and LOC) in the formation of trust in automation.

In this study, while depressiveness did not directly predict trust in automation, an indirect effect via the experience of anxiety was supported. Participants who experience more depressive symptoms experienced more anxiety before being introduced to an automated driving system. In this regard, the reported association of depression and anxiety generalized to the interaction with the automated system in the study. The higher experience of anxiety in depressed persons might be in part rooted in distorted self-confidence, mood and concentration (e.g., Fried et al., 2016; Fried, 2017; whereas the latter is related to attentional processes) – variables that have in turn been related to the formation of trust in automation (Hoff and Bashir, 2015). While this research provides a first investigation of the role of depression for trust in automation, future studies should further investigate this relationship.

Furthermore, this study supports the role of negative self-evaluations for both the experience of anxiety before being introduced to a new automated system and trust in the system after some initial interaction. The association of self-esteem and self-efficacy to an increased experience of anxiety resembles the findings of studies in other domains (Chamorro-Premuzic et al., 2008; Xiao et al., 2014). For these two personality characteristics (when investigated individually) also the mediation of their effect on trust by state anxiety was significant. People, who perceive themselves to be worthy and capable of performing well in a broad array of tasks, experience a lower level of anxiety when being about to interact with a new automated system. This lower level of anxiety is associated to higher trust in the system subsequently. On the contrary, people with lower regard for themselves also doubt their ability to manage new challenges well and therefore are more scared in unfamiliar situations. This experience of anxiety in turn translates to lower trust in an automated system that has been interacted with for a while. On the other hand, unlike hypothesized, a direct association between LOC and trust in the system was not supported by the data. This could be due to the fact that LOC refers to a general attribution of an outcome, not to its valence (e.g., Ajzen, 2002). In LOC, the attribution who is responsible for an outcome is more important as in self-esteem and self-efficacy. Moreover, when all investigated psychological variables were considered in one model, self-esteem showed the strongest association with anxiety. Self-esteem, when considered alone, explains 36% of the variance in state anxiety. In the reported study, the global evaluation of one's self-worth was a more prominent predictor for the experience of anxiety and individual differences in trust in automation than all other considered personality variables, also compared to self-efficacy which is related to one's beliefs to be able to successfully perform in a new task. How people evaluate their own abilities in general thus has far-reaching implications on their experience of anxiety in unfamiliar situations, which in turn mediates the effect of this generalized self-evaluations on the individual level of trust in automation.

Taken together, these findings underline the role of personality characteristics associated with the perception of one's self-worth and the assessment of one's capability to perform for the experience of anxiety in the adoption of automated driving systems. These psychological mechanisms are important to be considered in the design and dissemination of automated driving technology.

Strengths, Limitations and Future Research

This study is the first of its kind to investigate a mediation of personality effects on the evaluation of an automated system by anxiety. An essential strength of this research is that predictors and criteria variables were assessed at different points in time. This allows for an approximation of causality in a correlative design. Also, the proposed direction of effects from personality via states to trust in automation is in line with theory and empirical findings on the associations between personality and emotional states and judgments in general and trust in specific (Schwarz and Clore, 1988; Forgas, 1995; Dunn and Schweitzer, 2005; Forgas and East, 2008). Any other order of these constructs in the mediation process would not be meaningful from a theoretical perspective. Furthermore, this study was conducted in a driving simulator laboratory, which allows for a certain realism and immersion as compared to other set-ups typically used in correlative designs (e.g., online studies). Also, the study used established validated scales for all measurements, which further increases the external validity of the findings.

Besides from these strengths, the presented study comes with some limitations to be considered, and that might be addressed in future studies. First, although most hypothesized relationships were supported by the data, some of the non-significant results might be due to the restricted power of this study and could be investigated in studies with larger samples. Second, the sample consisted in the mere majority of not depressed and subclinical depressed persons. The relationship of depression, anxiety, and trust might be further elevated in persons with a clinical MDD diagnosis. Future studies might further investigate the reported relationships in clinical samples. Third, the included psychological variables only reflect a selection of theoretically relevant variables for the emergence of individual differences in situational trust appraisals in the interaction with automated (driving) systems. Also, in this regard, future studies could investigate the relationship of other emotional states besides positive, negative affect, and anxiety to trust in automation. Furthermore, potential moderating variables of the mediation cascade from traits via emotional states to trust or automation-related behavior might be investigated. Additionally, future studies could manipulate the relevant system or situational

characteristics that trigger the experience of anxiety to better understand the relationship of anxiety and trust in automated systems. For example, the perception of risk and uncertainty of using an automated system might be manipulated in future studies. It has to be noted that the reported findings neither restrict the hypothetical relevance of LOC nor that of positive and negative affect for the formation of trust as the theoretically expectable directions of effects could be observed and might only not been significant due to this study's relatively small sample. Whatsoever, as the remaining effects were significant with this sample size, at least it might be concluded that the predictive power of those variables with supported relationships seems to be somewhat higher under the circumstances of the reported study.

Practical Implications

The relationship of anxiety and trust in automation underlines the importance of paying respect to users' emotional states, feelings of uncertainty and risk perception in people, who initially get to know an automated system and interact with it for the first time. As a consequence, practitioners might pay attention to consider in which situations and in which mindset users are when they are first introduced to an unfamiliar system. For example, it seems worthwhile to reduce the perception of uncertainty and risks in this situation and rather provide reassuring information in order to reduce anxiety and increase initial trust in the system. Thereby it should be kept in mind that the goal of the process of introducing the user to a new system is always a calibrated level of trust rather than maximizing trust irrespective of the actual capabilities of the system.

Also, the reported relationships of individual feelings of depression and negative self-evaluations in regard to one's worth or capability to effectively manage new tasks with increased anxiety levels and decreased trust in automated systems allow for the derivation of implications for practice. The reported study findings point into the direction of individualized communication and training of system functionality as well as personalized design of user interfaces in orientation to personality differences to facilitate an appropriate level of trust in different user groups. For example, people with lower self-esteem and lower self-efficacy could be provided with a specific driver training that bolsters against anxiety in the first (independent) use of the automated driving system. Such training could explicitly assess which unrealistic fears exist and address them by providing factual information about the capabilities and functioning of the respective system. On the other hand, persons with high levels of self-esteem and self-efficacy might tend to experience low levels of anxiety potentially leading to overtrust in the system. To prevent them from using the system in situations it is not designed for a personalized driver training might stress potential dangers and risks in the interaction with the automated driving system under consideration and thereby enhancing a more realistic assessment of their own and the system's abilities. Regarding the design of automated systems, user groups with different levels of positive or negative self-evaluation will likely react very differently to certain information in different phases during system familiarization. Practitioners might therefore carefully balance out the provided information about system functioning and system reliability in automated driving systems under consideration of both their costumers' individual level of anxiety and their self-assessment of the capability to use the system adaptively. Above this, similar to the mentioned personalized driver training, personalized design of user interfaces might be useful for users with different levels of a positive or negative self-evaluation to facilitate individual trust calibration.

The role of depressive symptoms for the experience of anxiety and the formation of trust in automated systems underlines that besides subclinical personality traits also clinical diagnoses should be considered in the dissemination phase of driving automation technology. While this is on the one hand important for manufacturers it is also essential to be addressed on a societal and legislative level in the face of the growing prevalence of MDD. The reported findings of an association of the experience of depressive symptoms and anxiety in turn associated with decreased levels of trust in automation are of special interest as depressive symptoms and the intake of respective medication have been related to negative driving behavior like risky driving (e.g., McDonald et al., 2014), increased reaction times (e.g., Bulmash et al., 2006) and increased risk of accidents (e.g., Bulmash et al., 2006; Hilton et al., 2009; Cameron and Rapoport, 2016), for which driving automation might provide a solution. Therefore, people with depression symptoms are among those who could benefit most from driving automation, while at the same time being especially prone to the experience of anxiety in the face of unfamiliar and potentially risky situations (like the interaction with a new automated driving system). This underlines the importance of explicitly addressing their fears both on an individual and societal level. Additionally, for example, the acquisition of automated driving aids might be funded by governments or health insurances to provide a broader access to automated driving aids for depressed persons. Hereby, it should be carefully taken into consideration to not discriminate against people with MDD but rather to empower them to fully benefit from the significant alleviation by this technology for their mobility.

Conclusion

Taken together, this research underlines the importance of considering anxiety and other emotional states in trust formation when people first get to know automated systems. Furthermore, the experience of anxiety when being introduced to the new technology was found to be rooted in the individual tendency to experience depressive symptoms and have negative self-evaluations. While a certain degree of anxiety is a normal and adaptive reaction in the face of risky and potentially dangerous new technology, in order to facilitate a calibrated level of trust, it should be taken care that anxiety levels do neither fall below nor exceed the actual level of dangers and risks posed by this new technology as otherwise, this might lead to over- and distrust. To facilitate a calibrated level of trust, the

investigated psychological mechanisms in this study underline that both situational and personality characteristics associated with an experience of anxiety in the face of driving automation should be considered.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The study was carried out in accordance with the Declaration of Helsinki. Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JK generated the project idea, collected the data, performed the analyses, and led the manuscript write-up. DS assisted with collecting the data, data analyses, and project write-up. E-MM assisted with the data analyses and project write-up. MM assisted with generating the project idea and project write-up. MB assisted with the project write-up.

FUNDING

The article processing fee was paid by the Open Access Fund of Ulm University.

REFERENCES

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. *J. Appl. Soc. Psychol.* 32, 665–683. doi: 10.1111/j.1559-1816.2002.tb00236.x

Bandura, A. (1988). Self-efficacy conception of anxiety. *Anxiety Res.* 1, 77–98. doi: 10.1080/10615808808248222

Bandura, A. (1991). Social cognitive theory of self-regulation. *Org. Behav. Hum. Decis. Process.* 50, 248–287. doi: 10.1016/0749-5978(91)90022-L

Baumeister, H., Nowoczin, L., Lin, J., Seifferth, H., Seufert, J., Laubner, K., et al. (2014). Impact of an acceptance facilitating intervention on diabetes patients' acceptance of Internet-based interventions for depression: a randomized controlled trial. *Diabetes Res. Clin. Pract.* 105, 30–39. doi: 10.1016/j.diabres.2014.04.031

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). An inventory for measuring depression. *Arch. Gen. Psychiatry* 4, 561–571. doi: 10.1001/archpsyc.1961.01710120031004

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

Biros, D. P., Daly, M., and Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decis. Negot.* 13, 173–189. doi: 10.1023/b:grup.0000021840.85686.57

Bishara, A. J., and Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychol. Methods* 17, 399–417. doi: 10.1037/a0028087

Briggs, P., Burford, B., and Dracup, C. (1998). Modelling self-confidence in users of a computer-based system showing unrepresentative design. *Int. J. Hum. Comput. Stud.* 49, 717–742. doi: 10.1006/ijhc.1998.0224

Bulmash, E. L., Moller, H. J., Kayumov, L., Shen, J., Wang, X., and Shapiro, C. M. (2006). Psychomotor disturbance in depression: assessment using a driving simulator paradigm. *J. Affect. Disord.* 93, 213–218. doi: 10.1016/j.jad.2006.01.015

Cameron, D. H., and Rapoport, M. J. (2016). Antidepressants and driving in older adults: a systematic review. *Can. J. Aging* 35, 7–14. doi: 10.1017/S0714980816000064

Chamorro-Premuzic, T., Ahmetoglu, G., and Furnham, A. (2008). Little more than personality: dispositional determinants of test anxiety (the Big Five, core self-evaluations, and self-assessed intelligence). *Learn. Individ. Differ.* 18, 258–263. doi: 10.1016/j.lindif.2007.09.002

Cohen, S., Rothbart, M., and Phillips, S. (1976). Locus of control and the generality of learned helplessness in humans. *J. Pers. Soc. Psychol.* 34, 1049–1056. doi: 10.1037/0022-3514.34.6.1049

Coopersmith, S. (1981). *Coopersmith Self-Esteem Inventory (School Form).* San Diego: Consulting Psychology Press.

Cuijpers, P., De Graaf, R., and Van Dorsselaer, S. (2004). Minor depression: risk profiles, functional disability, health care use and risk of developing major depression. *J. Affect. Disord.* 79, 71–79. doi: 10.1016/S0165-0327(02)00348-8

Cuijpers, P., Koole, S. L., van Dijke, A., Roca, M., Li, J., and Reynolds, C. F. (2014). Psychotherapy for subclinical depression: meta-analysis. *Br. J. Psychiatry* 205, 268–274. doi: 10.1192/bjp.bp.113.138784

Cummings, C. M., Caporino, N. E., and Kendall, P. C. (2014). Comorbidity of anxiety and depression in children and adolescents: 20 years after. *Psychol. Bull.* 140, 816–845. doi: 10.1037/a0034733

Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A., Yang, X., et al. (2019). Look who's talking now: implications of AV's explanations on driver's trust, AV preference, anxiety, and mental workload. *Transport. Res. Part C* 104, 428–442. doi: 10.1016/j.trc.2019.05.025

DuCette, J., and Wolk, S. (1973). Cognitive and motivational correlates of generalized expectancies for control. *J. Pers. Soc. Psychol.* 26, 420–426. doi: 10.1037/h0034486

Dunn, J. R., and Schweitzer, M. E. (2005). Feeling and believing: the influence of emotion on trust. *J. Pers. Soc. Psychol.* 88, 736–748. doi: 10.1037/0022-3514.88.5.736

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *Int. J. Hum. Comput. Stud.* 58, 697–718. doi: 10.1016/S1071-5819(03)00038-7

Ebert, D. D., Berking, M., Cuijpers, P., Lehr, D., Pörtner, M., and Baumeister, H. (2015). Increasing the acceptance of internet-based mental health interventions in primary care patients with depressive symptoms. A randomized controlled trial. *J. Affect. Disord.* 176, 9–17. doi: 10.1016/j.jad.2015.01.056

Elhai, J. D., Levine, J. C., Dvorak, R. D., and Hall, B. J. (2016). Fear of missing out, need for touch, anxiety and depression are related to problematic smartphone use. *Comput. Hum. Behav.* 63, 509–516. doi: 10.1016/j.chb.2016.05.079

Ferring, D., and Filipp, S.-H. (1996). messung des selbstwertgefühls: befunde zu reliabilität, validität und stabilität der rosenberg-skala. *Diagnostica* 42, 284–292.

Forgas, J. P. (1995). Mood and judgment: the affect infusion model (AIM). *Psychol. Bull.* 117, 39–66. doi: 10.1037/0033-2909.117.1.39

Forgas, J. P., and East, R. (2008). On being happy and gullible. Mood effects on skepticism and the detection of deception. *J. Exp. Soc. Psychol.* 44, 1362–1367. doi: 10.1016/j.jesp.2008.04.010

Forgas, J. P., and George, J. M. (2001). Affective influences on judgments and behavior in organizations: an information processing perspective. *Org. Behav. Hum. Decis. Process.* 86, 3–34. doi: 10.1006/obhd.2001.2971

Fried, E. I. (2017). The 52 symptoms of major depression: lack of content overlap among seven common depression scales. *J. Affect. Disord.* 208, 191–197. doi: 10.1016/j.jad.2016.10.019

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., and Borsboom, D. (2016). What are 'good' depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *J. Affect. Disord.* 189, 314–320. doi: 10.1016/j.jad.2015.09.005

Glöckner, A., and Hilbig, B. E. (2012). Risk is relative: risk aversion yields cooperation rather than defection in cooperation-friendly environments. *Psychonom. Bull. Rev.* 19, 546–553. doi: 10.3758/s13423-012-0224-z

Greenberg, J., Pyszczynski, T., and Solomon, S. (1986). "The causes and consequences of a need for self-esteem: a terror management theory," in *Public Self and Private Self*, ed. R. Baumeister, (New York, NY: Springer), 189–212. doi: 10.1007/978-1-4613-9564-5_10

Greenberg, J., Solomon, S., Pyszczynski, T., Rosenblatt, A., Burling, J., Lyon, D., et al. (1992). Why do people need self-esteem? Converging evidence that self-esteem serves an anxiety-buffering function. *J. Pers. Soc. Psychol.* 63, 913–922. doi: 10.1037//0022-3514.63.6.913

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

Harter, S. (1993). "Causes and consequences of low self-esteem in children and adolescents," in *Self-Esteem: The Puzzle of Low Self-Regard*, ed. R. Baumeister, (New York, NY: Plenum), 87–116. doi: 10.1007/978-1-4684-8956-9_5

Hautzinger, M., Keller, F., and Kühner, C. (2009). *Beck-Depressions-Inventar: BDI-II; Manual [Beck Depression-Inventory: BDI II; Manual]*, 2nd Edn. Frankfurt am Main: PearsonAssessment.

Havlicek, L. L., and Peterson, N. L. (1976). Robustness of the Pearson correlation against violations of assumptions. *Percept. Mot. Skills* 43, 1319–1334. doi: 10.2466/pms.1976.43.3f.1319

Hayes, A. F. (2009). Beyond Baron and Kenny: statistical mediation analysis in the new millennium. *Commun. Monogr.* 76, 408–420. doi: 10.1080/03637750903310360

Hayes, A. F. (2018). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. New York, NY: Guilford Publications.

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

Herz, L., and Gullone, E. (1999). The relationship between self-esteem and parenting style: a cross-cultural comparison of Australian and Vietnamese Australian adolescents. *J. Cross Cult. Psychol.* 30, 742–761. doi: 10.1177/0022022199030006005

Hilton, M. F., Staddon, Z., Sheridan, J., and Whiteford, H. A. (2009). The impact of mental health symptoms on heavy goods vehicle drivers' performance. *Accid. Anal. Prevent.* 41, 453–461. doi: 10.1016/j.aap.2009.01.012

Hiroto, D. S. (1974). Locus of control and learned helplessness. *J. Exp. Psychol.* 102, 187–193. doi: 10.1037/h0035910

Hock, P., Kraus, J., Babel, F., Walch, M., Rukzio, E., and Baumann, M. (2018). "How to design valid simulator studies for investigating user experience in automated driving: review and hands-on considerations," in *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, (New York, NY: ACM), 105–117. doi: 10.1145/3239060.3239066

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

Horikawa, M., and Yagi, A. (2012). The relationships among trait anxiety, state anxiety and the goal performance of penalty shoot-out by university soccer players. *PLoS One* 7:e35727. doi: 10.1371/journal.pone.0035727

Hörl, S., Ciari, F., and Axhausen, K. W. (2016). "Recent perspectives on the impact of autonomous vehicles," in *Working paper Institute for Transport Planning and Systems,10XX, Institute for Transport Planning and Systems (IVT)*, (Zurich: ETH Zurich), doi: 10.3929/ethz-b-0001213

Horwath, E., Johnson, J., Klerman, G. L., and Weissman, M. M. (1992). Depressive symptoms as relative and attributable risk factors for first-onset major depression. *Arch. Gen. Psychiatry* 49, 817–823. doi: 10.1001/archpsyc.1992.01820100061011

Jacques, P. H., Garger, J., Brown, C. A., and Deale, C. S. (2009). Personality and virtual reality team candidates: the roles of personality traits, technology anxiety and trust as predictors of perceptions of virtual reality teams. *J. Bus. Manag.* 15, 143–156.

Jerusalem, M., and Schwarzer, R. (1999). "Allgemeine Selbstwirksamkeitserwartung," in *Skalen zur Erfassung von Lehrer- und Schülermerkmalen*, eds R. Schwarzer, and M. Jerusalem, (Berlin: Freie Universität Berlin), 57–59.

Jian, J. Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergonom.* 4, 53–71. doi: 10.1207/S15327566IJCE0401_04

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., and Walters, E. E. (2005a). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* 62, 593–602. doi: 10.1001/archpsyc.62.6.593

Kessler, R. C., Chiu, W. T., Demler, O., and Walters, E. E. (2005b). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* 62, 617–627. doi: 10.1001/archpsyc.62.6.617

Kim, S. S., Chung, Y., Perry, M. J., Kawachi, I., and Subramanian, S. V. (2012). Association between interpersonal trust, reciprocity, and depression in South Korea: a prospective analysis. *PLoS One* 7:e30602. doi: 10.1371/journal.pone.0030602

Kraus, J. M., Forster, Y., Hergeth, S., and Baumann, M. (2019a). Two routes to trust calibration: effects of reliability and brand information on trust in automation. *Int. J. Mob. Hum. Comput. Interact.* 11, 1–17. doi: 10.4018/IJMHCI.2019070101

Kraus, J., Scholz, D., Stiegemeier, D., and Baumann, M. (2019b). The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Hum. Fact.* doi: 10.1177/0018720819853686 [Epub ahead of print]

Krohne, H. W., Egloff, B., Kohlmann, C.-W., and Tausch, A. (1992). Untersuchung mit einer deutschen version der „positive and negative affect schedule" (PANAS). *Diagnostica* 42, 139–156.

Kühn, M., and Hannawald, L. (2015). "Verkehrssicherheit und Potenziale von Fahrerassistenzsystemen," in *Handbuch Fahrerassistenzsysteme*, eds H. Winner, S. Hakuli, F. Lotz, and C. Singer (Wiesbaden: Springer), 55–70. doi: 10.1007/978-3-658-05734-3_4

Larsen, R. J., and Buss, D. M. (2005). *Personality Psychology: Domains of Knowledge About Human Nature*. New York, NY: McGraw Hill.

Larsen, R. J., and Ketelaar, T. (1991). Personality and susceptibility to positive and negative emotional states. *J. Pers. Soc. Psychol.* 61, 132–140. doi: 10.1037/0022-3514.61.1.132

Lee, J., and 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., and Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *Int. J. Hum. Comput. Stud.* 40, 153–184. doi: 10.1006/ijhc.1994.1007

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

Lester, D., and Gatto, J. L. (1990). Interpersonal trust, depression, and suicidal ideation in teenagers. *Psychol. Rep.* 67, 786–786. doi: 10.2466/pr0.1990.67.3.786

Letzring, T. D., and Adamcik, L. A. (2015). Personality traits and affective states: relationships with and without affect induction. *Pers. Individ. Differ.* 75, 114–120. doi: 10.1016/j.paid.2014.11.011

Lewandowsky, S., Mundy, M., and Tan, G. (2000). The dynamics of trust: comparing humans to automation. *J. Exp. Psychol.* 6, 104–123. doi: 10.1037/1076-898X.6.2.104

Locke, E. A., Frederick, E., Lee, C., and Bobko, P. (1984). Effect of self-efficacy, goals, and task strategies on task performance. *J. Appl. Psychol.* 69, 241–251. doi: 10.1037/0021-9010.69.2.241

Locke, E. A., and Latham, G. P. (1994). "Goal setting theory," in *Motivation: Theory and Research*, eds H. F. O'Neil, and M. Drillings (Hillsdale: Lawrence), 13–30.

Lovibond, P. F., and Lovibond, S. H. (1995). The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the

Beck Depression and Anxiety Inventories. *Behav. Res. Ther.* 33, 335–343. doi: 10.1016/0005-7967(94)00075-u

Marteau, T. M., and Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). *Br. J. Clini. Psycbhol.* 31, 301–306. doi: 10.1111/j.2044-8260.1992.tb00997.x

Mathews, A. (1990). Why worry? The cognitive function of anxiety. *Behav. Res. Ther.* 28, 455–468. doi: 10.1016/0005-7967(90)90132-3

Mathews, A., and Mackintosh, B. (1998). A cognitive model of selective processing in anxiety. *Cognit. Ther. Res.* 22, 539–560. doi: 10.1023/A:1018738019346

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manag. Rev.* 20, 709–734. doi: 10.5465/amr.1995.9508080335

McDonald, C. C., Sommers, M. S., and Fargo, J. D. (2014). Risky driving, mental health, and health-compromising behaviours: risk clustering in late adolescents and adults. *Inj. Prev.* 20, 365–372. doi: 10.1136/injuryprev-2014-041150

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Hum. Factors* 53, 356–370. doi: 10.1177/0018720811411912

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

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

Moray, N., and Inagaki, T. (2000). Attention and complacency. *Theor. Issues Ergonom. Sci.* 1, 354–365. doi: 10.1080/14639220052399159

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

Muir, B. M., and 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

Murphy, J. M., Horton, N. J., Laird, N. M., Monson, R. R., Sobol, A. M., and Leighton, A. H. (2004). Anxiety and depression: a 40-year perspective on relationships regarding prevalence, distribution, and comorbidity. *Acta Psychiatr. Scand.* 109, 355–375. doi: 10.1111/j.1600-0447.2003.00286.x

Nomura, T., Kanda, T., Suzuki, T., and Kato, K. (2008). Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE Trans. Robot.* 24, 442–451. doi: 10.1109/TRO.2007.914004

Nutt, D. J. (1999). Care of depressed patients with anxiety symptoms. *J. Clin. Psychiatry* 60, 23–27.

Osswald, S., Wurhofer, D., Trösterer, S., Beck, E., and Tscheligi, M. (2012). "Predicting information technology usage in the car: towards a car technology acceptance model," in *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (New York, NY: ACM), 51–58.

Owens, T. J. (1993). Accentuate the positive-and the negative: rethinking the use of self-esteem, self-deprecation, and self-confidence. *Soc. Psychol. Q.* 56, 288–299. doi: 10.2307/2786665

Özkan, T., Lajunen, T., and Kaistinen, J. (2005). "Traffic locus of control, driving skills and attitudes towards in-vehicle technologies (ISA & ACC)," in *Proceedings of the 18th International Cooperation on Theories and Concepts in Traffic Safety (ICTCT)*, Helsinki, Finland.

Parasuraman, R., and Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Hum. Factors* 39, 230–253. doi: 10.1518/001872097778543886

Payre, W., Cestac, J., and Delhomme, P. (2016). Fully automated driving: impact of trust and practice on manual control recovery. *Hum. Factors* 58, 229–241. doi: 10.1177/0018720815612319

Peters, E., Västfjäll, D., Gärling, T., and Slovic, P. (2006). Affect and decision making: a "hot" topic. *J. Behav. Decis. Making* 19, 79–85. doi: 10.1002/bdm.528

Phares, E. J. (1976). *Locus of Control in Personality*. Morris-town, NJ: General Learning Press.

Preacher, K. J., and Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav. Res. Methods* 40, 879–891. doi: 10.3758/BRM.40.3.879

Rajaonah, B., Anceaux, F., and Vienne, F. (2006). Trust and the use of adaptive cruise control: a study of a cut-in situation. *Cogn. Technol. Work* 8, 146–155. doi: 10.1007/s10111-006-0030-3

Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling and more. Version 0.5–12 (BETA). *J. Stat. Softw.* 48, 1–36.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of 1130 reinforcement. *Psychol. Monogr.* 80, 1–28. doi: 10.1037/h0092976

Rotter, J. B. (1990). Internal versus external control of reinforcement: a case history of a variable. *Am. Psychol.* 45, 489–493. doi: 10.1037/0003-066X.45.4.489

Sanchez, J., Fisk, A. D., and Rogers, W. A. (2004). "Reliability and age-related effects on trust and reliance of a decision support aid," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 48 (Los Angeles, CA: Sage Publications), 586–589. doi: 10.1177/154193120404800366

Schachter, S., and Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychol. Rev.* 69, 379–399. doi: 10.1037/h0046234

Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T. L., Chen, J. Y., et al. (2014). *A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Human-Robot Interaction*. Aberdeen: Army Research Laboratory.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., and Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. *Hum. Factors* 58, 377–400. doi: 10.1177/0018720816634228

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Edu. Psychol.* 26, 207–231. doi: 10.1080/00461520.1991.9653133

Schwarz, N., and Clore, G. L. (1988). "How do I feel about it? Informative functions of affective states," in *Affect, Cognition, and Social Behavior*, eds K. Fiedler, and J. Forgas (Toronto, ON: Hogrefe), 44–62.

Seligman, M. E. (1972). Learned helplessness. *Annu. Rev. Med.* 23, 407–412.

Sowislo, J. F., and Orth, U. (2013). Does low self-esteem predict depression and anxiety? A meta-analysis of longitudinal studies. *Psychol. Bull.* 139, 213–240. doi: 10.1037/a0028931

Spector, P. E. (1982). Behavior in organizations as a function of employee's locus of control. *Psychol. Bull.* 91, 482–497. doi: 10.1037/0033-2909.91.3.482

Spielberger, C. D. (1966). *Anxiety and Behavior*. New York, NY: Academic Press.

Stokes, C. K., Lyons, J. B., Littlejohn, K., Natarian, J., Case, E., and Speranza, N. (2010). "Accounting for the human in cyberspace: effects of mood on trust in automation," in *Proceedings of the International Symposium on Collaborative*, Chicago, IL, 180–187.

Szalma, J. L., and Taylor, G. S. (2011). Individual differences in response to automation: the five factor model of personality. *J. Exp. Psychol.* 17, 71–96. doi: 10.1037/a0024170

Taylor, A. B., MacKinnon, D. P., and Tein, J. Y. (2008). Tests of the three-path mediated effect. *Organ. Res. Methods* 11, 241–269. doi: 10.1177/1094428107300344

Thatcher, J. B., Loughry, M. L., Lim, J., and McKnight, D. H. (2007). Internet anxiety: an empirical study of the effects of personality, beliefs, and social support. *Inform. Manag.* 44, 353–363. doi: 10.1016/j.im.2006.11.007

Thielmann, I., and Hilbig, B. E. (2015). Trust: an integrative review from a person–situation perspective. *Rev. Gen. Psychol.* 19, 249–277. doi: 10.1037/gpr0000046

Winkle, T. (2015). "Sicherheitspotenzial automatisierter Fahrzeuge: Erkenntnisse aus der Unfallforschung," in *Autonomes Fahren*, eds H. Winner, S. Hakuli, and G. Wolf (Heidelberg: Springer), 351–376. doi: 10.1007/978-3-662-45854-9_17

Xiao, W., Zhou, L., Wu, Q., Zhang, Y., Miao, D., Zhang, J., et al. (2014). Effects of person-vocation fit and core self-evaluation on career commitment of medical university students: the mediator roles of anxiety and career satisfaction. *Int. J. Mental Health Syst.* 8:8. doi: 10.1186/1752-4458-8-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Kraus, Scholz, Messner, Messner and Baumann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Article III: Two Routes to Trust Calibration: Effects of Reliability and Brand Information on Trust in Automation

Kraus, J. M., Forster, Y., Hergeth, S., & Baumann, M. (2019). Two routes to trust calibration: Effects of reliability and brand information on trust in automation. *International Journal of Mobile Human Computer Interaction, 11*(3), 1-17. https://doi.org/10.4018/IJMHCI.2019070101

This paper appears in the International Journal of Mobile Human Computer Interaction authored by Johannes Kraus, Yannick Forster, Sebastian Hergeth and Martin Baumann

Copyright 2019, IGI Global, www.igi-global.com. Posted by permission of the publisher.

International Journal of Mobile Human Computer Interaction
Volume 11 • Issue 3 • July-September 2019

Two Routes to Trust Calibration:
Effects of Reliability and Brand Information on Trust in Automation

Johannes Maria Kraus, Ulm University, Ulm, Germany
Yannick Forster, BMW Group, Bayern, Germany
Sebastian Hergeth, BMW Group, Bayern, Germany
Martin Baumann, Ulm University, Ulm, Germany

ABSTRACT

Trust calibration takes place prior to and during system interaction along the available information. In an online study N = 519 participants were introduced to a conditionally automated driving (CAD) system and received different a priori information about the automation's reliability (low vs high) and brand of the CAD system (below average vs average vs above average reputation). Trust was measured three times during the study. Additionally, need for cognition (NFC) and other personality traits were assessed. Both heuristic brand information and reliability information influenced trust in automation. In line with the Elaboration Likelihood Model (ELM), participants with high NFC relied on the reliability information more than those with lower NFC. In terms of personality traits, materialism, the regulatory focus and the perfect automation scheme predicted trust in automation. These findings show that a priori information can influence a driver's trust in CAD and that such information is interpreted individually.

KEYWORDS

Attitude Formation, Automated Driving, Driver-Vehicle Interaction, Elaboration Likelihood Model, Human-Machine Interaction, Online Research, Personality, Reliability, Trust, Trust Calibration

INTRODUCTION

The complex interaction pattern in CAD at level 3 automation as defined by the Society of Automotive Engineers (SAE) in the SAE J3016R technical report (SAE, 2016) requires an elaborate design of the user interface to guarantee efficient and comfortable coordination of driving tasks between driver and automation. In this regard psychological concepts like an appropriate level of trust support a human-centred design of interaction concepts. Trust in automation has been found to play a particularly important role in the decision if and how drivers use automated systems (e.g. Hoff & Bashir, 2015). Thus, the nature of trust formation holds important implications for the design of safe, acceptable and efficient driving automation (e.g. Walch et al., 2017). Among others, the information about the automated system available to the driver before the first system use may determine if the system

DOI: 10.4018/IJMHCI.2019070101

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

is used despite initial suspicion. To overcome this obstacle – and at the same time preventing an overly optimistic impression of the automation - the information provided prior to actual use has to be designed in a way that a general positive attitude towards the automation is generated while at the same time an appropriate understanding about its actual capabilities and limitations is facilitated. As a consequence, an important challenge for designing safe and efficient CAD systems is it to increase the knowledge of what kind of information drivers use for trust calibration under which conditions.

At this point, little research has been conducted on the mechanisms of trust formation and the nature and timing of information fostering an adequate mental model and efficient trust calibration (Lee & See, 2004; Kraus, Scholz, Stiegemeier, & Baumann, 2018). Also, the contextual factors playing a role in this process have not been investigated to a sufficient degree. The current study provides an investigation of the effects of information on a priori trust levels towards automated driving. On the theoretical basis of the Elaboration Likelihood Model (ELM) (Petty & Cacioppo, 1986) two types of a priori information (OEM brand information and reliability information) were manipulated in an online experiment. Based on a cluster analysis, six OEM brands were grouped in three brand reputation groups (below average/average/above average) and as a manipulation for system reliability, participants read one-pagers, in which the results of an imaginary function test were reported (low vs high reliability). Additionally, several personality traits that are hypothesized to influence the first impression of an automated vehicle were investigated. Furthermore, the role of need for cognition (NFC) (Cacioppo, Petty, Kao, & Rodriguez, 1986) in the utilization of different types of information was investigated.

THEORETICAL BACKGROUND

Trust in Automation

Trust in automation is defined as "[…] the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p. 51). With the increasing complexity of driving automation and the dynamic character of the driving task, for users, a thorough understanding of its functional scope and its operation sequence becomes more and more difficult. Consequently, trust in automation gains importance for the decision to use or not use an automation (e.g. Lee & See, 2004) and thus it is crucial that users' trust is calibrated in line with a system's actual capabilities and performance (Muir, 1987). An appropriate level of trust is the key to a balanced and safe usage of driving automation and an essential goal for the design of interaction concepts for driving automation. This is best illustrated by the two situations in which trust is wrongly calibrated and either falls below (distrust) or exceeds an appropriate trust level (overtrust; Parasuraman & Riley, 1997).

Trust calibration is the result of a complex psychological process prior, during and after the first interaction with a system, in which the available information is combined into a trust judgement (e.g., Castelfranchi, 1998; Hoff & Bashir, 2015; Lee & Moray, 1992). In line with the model by Hoff and Bashir (2015) it can be differentiated between trust processes prior to and during system interaction. It can be assumed that in different stages trust calibration builds on different information (Hoff & Bashir, 2015; Lee & Moray, 1992; Lee & See, 2004). In reference to the conceptualization of trust as an attitude (Lee & See, 2004), available information at both stages might be used according to an attitude formation process to build up trust. Accordingly, it seems fruitful to apply models of attitude change to the area of trust in automation to better understand trust formation processes and the associated information use. In their model, Lee and See (2004) used the theory of planned behaviour (PLABE theory; Ajzen & Fishbein, 1980; Ajzen, 1991) to explain how trust is formed in a cascade from norms and belief and how trust in turn affects intentions and behavioural tendencies in a dynamic process. Additionally, mental models (Carol & Olson, 1987) could provide a cognitive structure for the representation of available information of the system and could be integrated into

the PLABE cascade feeding into the formation of beliefs, attitudes and behavioural intentions. In line with this, it has been shown that the mental model of a level 1 driving automation is influenced by a priori information and related to trust development over time (Begiatto & Krems, 2013).

Influence of a Priori Information on Trust in Automation

The formation of attitudes towards automated driving with information provided prior to actual system use and their role for trust calibration was both addressed in theoretical discourse (e.g. Hoff & Bashir, 2015) and empirical investigations (e.g. Hergeth, Lorenz, & Krems, 2017) recently. Regarding empirical evidence, in an experiment, in which participants were primed either with a correct, an incorrect or an incomplete mental model of the ACC function, Beggiato and Krems (2013) report that only a realistic mental model leads to a steadily increasing calibrated level of trust. In accordance with Lee and See (2004), they found that for the evolution of trust predictability of errors plays a more important role than error magnitude. Furthermore, Hergeth and colleagues (2017) found that prior information led to decrease of a priori trust levels and an improvement of driver decisions in take-over request (TOR) scenarios – thus potentially decreasing trust in these situations. Concerning the theoretical discussion of how available information is used to form attitudes about an automated driving system, Lee and See (2004) hypothesize three routes through which information affects trust. First, in the analytical route trust is built up in a process similar to rational decision-making under uncertainty (e.g. Lewicki & Bunker, 1996; Cohen, 2000), in which gains and losses are evaluated against each other. Second, in the analogical route trust is transferred from characteristics of the user interface or information of the environment along category membership of the attitude object. Third, in the affective route emotions towards the system under consideration or the interaction environment build the attitude (e.g. Kramer, 1999). It is important to note that these routes are not mutually exclusive, but available information can be processed via a combination of these routes and their relative importance depends, e.g. on the available information and how it is displayed. The proposed routes by Lee and See (2004) correspond well with dual process attitude change theories (Chaiken & Trope, 1999; Chen, Shechter, & Chaiken, 1996). The current study applies the Elaboration Likelihood Model (ELM) (Petty & Cacioppo, 1986) to the context of automated driving. In the ELM, two routes are proposed through which information can lead to attitude changes. First, the central route builds on the thorough elaboration of arguments and their significance, and thus attitude changes via this route are rather stable and possess high predictability for behaviour (e.g. Cialdini, Petty, & Cacioppo, 1981). As this route is characterized by thorough elaboration it is proposed that information via this route is evaluated in an analytical process (Lee & See, 2004). Second, the peripheral route builds on heuristic processes and a rather cognitively undemanding analysis of surface characteristics of information, which results in more instable attitude changes. The peripheral route is proposed to be mainly established by both attributional and affective processes.

According to the preceding reasoning, it can be concluded that information drivers perceive prior to the interaction with an automated vehicle affect their perception and their attitudes towards these vehicles. In line with the ELM, we follow that both routes of information processing influence trust in automated vehicles. This study investigates the impact of two kinds of a priori information on the formation of trust attitudes towards a CAD system: system reliability information which is hypothesized to be mainly used in an analytical process via the central route of the ELM and manufacturer brand information that is hypothesized to be used in an analogical (and to a certain extent affective) process via the peripheral route.

STUDY HYPOTHESES

Körber, Baseler, and Bengler (2018) showed an effect of different reliability information prior to system use on trust and performance in TOR situations. In accordance with the findings by Hergeth and colleagues (2017) lower trust levels were found to be associated with increased performance

when driver intervention was necessary. These findings point into the direction that information about system reliability and limitations is necessary to build up a realistic mental model about a system's functionality. This is further supported by findings of an online study by Biassoni, Ruscio, and Ciceri (2016), who found an interaction effect of reliability information and level of automation on trust: A listing of more system limitations lead to a decrease in trust towards the system, which was enhanced for SAE level 4 compared to level 2. In line with this we hypothesize:

Hypothesis 1: An automation that is described as reliable is trusted more than one that is described as unreliable.

While analytical processes (reliability information) have been more frequently investigated, this study takes an extended approach and investigates analogical processing (of brand information) additionally. First evidence for the role of analogical processing stems from an online study by Abraham, Seppelt, Mehler, and Reimer (2017), who found an effect of the naming of an automated driving function on the assessment of the perceived level of automation and thus an effect on the mental model of the automation, which was in turn found to influence trust and acceptance of an automated system (Beggiato & Krems, 2013). At this point, to our knowledge no research on the influence of manufacturer brand reputation on the formation of trust in an automated driving system is available. We, therefore, hypothesize, that the priming of a brand leads to the activation of brand attitudes, which are in turn analogically transferred to the attitude towards the system:

Hypothesis 2: A better OEM brand reputation of an automated system leads to higher trust in the system.

At this point, the mechanisms responsible for route selection in information processing in the attitude formation process leading to a specific level of trust in automation are not sufficiently investigated. To provide a clearer idea, which processing mechanism utilizes which information in a given situation, the research tradition of the ELM provides some directions. The route, in which available information is processed depends on a person's motivation and the available cognitive resources for an effortful analysis of the provided information (Petty & Cacioppo, 1986). Therefore, the effortful elaboration of arguments via the central route is only likely if a person has enough motivation and at the same time sufficient cognitive capacity to do so. One key variable that has been found to influence the motivation for an effortful elaboration of information is the personality trait need for cognition (NFC; Cacioppo et al., 1986). NFC has been defined as a "tendency to engage in and enjoy effortful cognitive endeavors" (Cacioppo, Petty, & Kao, 1984; p. 306) and describes the individual general motivation for investing cognitive effort in a task as well as a positive evaluation of one's own cognitive ability (Bless, Wänke, Bohner, Fellhauer, & Schwarz, 1994). In terms of the ELM, it was shown, that people with higher NFC scores used the central route more as compared to those with lower NFC scores (Cacioppo, Petty, & Morris, 1983). In line with this, we hypothesize:

Hypothesis 3: Need for cognition moderates the usage of either information in the following manner:
H3.1: People with higher NFC use the reliability information more than people with lower NFC.
H3.2: People with lower NFC use the brand reputation information more than people with higher NFC.

As personality is commonly hypothesized to influence trust formation in automated driving and other technological domains (e.g. Hock, Kraus, Babel, Walch, Rukzio, & Baumann, 2018; Schaefer, Chen, Szalma, & Hancock, 2016; Hoff & Bashir, 2015) three further personality traits were investigated: regulatory focus, materialism, and the perfect automation scheme. The selected

personality traits reflect different levels of situational specificity at which personality might influence the interpretation and representation of provided information in the process of trust formation (e.g. Mowen, 2000). Regulatory focus (Higgins, 1997) differentiates the factors promotion and prevention focus. While promotion-oriented persons are stronger occupied with winning than losing in situation representation, focus more on the fulfilment of one's own hopes and wishes as well as self-actualization, prevention-focused persons try to avoid mistakes and to fulfil security needs (Fellner, Holler, Kirchler, & Schabmann, 2007). Taken together, the regulatory focus should guide how information on risks and possible positive outcomes are represented in trust formation. Second, materialism was defined as the degree of "importance a consumer attaches to worldly possessions" (Belk, 1984, p. 291) as reflected in three characteristics: acquisition centrality, acquisition as the pursuit of happiness and possession-defined success (Richins & Dawson, 1992). Materialism should strongly influence the perception of communication about new technology, especially in regard to automotive innovations as for their meaning for materialistic self-representation. Third, the perfect automation scheme entails beliefs about the performance of automated technology and includes high expectations towards automated systems to perform nearly perfectly as well as a tendency to be unforgiving towards system malfunction (Merritt, Unnerstall, Lee, & Huber, 2015). Dispositional individual differences in this cognitive scheme should strongly affect how provided information on driving automation is interpreted.

Hypothesis 4: Regulatory focus, materialism, as well as the perfect automation scheme, predict a priori trust in an automated system.

PILOT STUDY

In order to determine which brands to include in the main study to reflect different levels of brand reputation, an online pilot study was conducted. For this study, eight brands (BMW, Tesla, Volvo, Opel, Skoda, Toyota, Kia and Dacia) were selected based on an expert discussion and included in a within-design, in which participant rated all familiar brands. If a participant indicated not to be familiar with a brand, the brand was skipped. The dependent variables were divided in general (i.e., safety, reliability, quality, sympathy, features, comfort) and automation-related indicators (i.e., brand's competence in CAD development). All items were worded as statements and were to be answered on a 7-point Likert scale. The sample of the pilot study consisted of $N = 53$ participants, who were recruited at Ulm University. The sample consisted of 42 females, 10 males and one participant without further gender indication. The average age was $M = 23.42$ ($SD = 6.57$). For the identification of OEM reputation groups, cluster analysis with a Ward-Linkage hierarchical procedure was conducted. Group intervals were calculated with squared Euclidian distance. As a final result, BMW and Tesla were both clustered into an "above average reputation group" while all other OEM were grouped in a second group. Earlier steps in the statistical iteration separated Opel, Skoda and Volvo from Dacia, Kia and Toyota. On this basis, we chose Opel and Skoda as representative OEMs for an average reputation group and Kia and Dacia for the group with a below average reputation for the main study.

MAIN STUDY

Method

Design and Procedure

The study followed a three-factorial between design with the two between factors reliability information (two levels) and OEM brand reputation (three levels). Participants were randomly assigned to one combination of the two between factors. As an additional quasi-experimental between factor NFC was included in the study design. Participants had to use desktop computers to answer the survey.

The study procedure is depicted in Figure 1. Participants had to indicate which of the included car manufacturers they were familiar with and that they hold a driver license for at least three years. In the following, participants were welcomed, informed about the survey and had to read and sign an informed consent form. All participants read the same description of system characteristics (translated from German): "With this system a vehicle can drive independently in public traffic by executing steering, accelerating, distance management and certain manoeuvres. The driver has the possibility to detach from the driving task (hands off the steering wheel, feet off the pedals) and can engage in non-driving related tasks. However, he/she has to be ready to take back vehicle control if the automated system encounters technical boundaries and cannot handle a situation by itself."

Subsequently, a set of questionnaires on attitudes about automated driving (not in the scope of this paper) and trust in automation were assessed (t_{pre}). In the second part, a one-pager was presented with a description of a CAD system, which included the manipulation of both experimental variables. The OEM brand was presented at two prominent positions of the one-pager. For manipulating system reliability, the result of a fictional function test was reported, which read in the low-reliability condition: "The present system was evaluated in an extensive functionality test. There were 9 malfunctions in 981km, which required a control transition from the system to the driver. Reasons for the malfunctions were hardware discrepancy (n=2), software discrepancy (n=1), unwanted manoeuvre (n=3), system limitation due to perception discrepancy (n=2) and incorrect behaviour prediction of other traffic participants (n=1)." Additionally, a ranking of 22 systems was provided. In the high-reliability condition the system ranked second and in the low-reliability condition 16th. In the low-reliability condition, all errors were multiplied by 10. After the one-pagers, participants rated their trust in the system a second time (t_1). Then, they watched four videos that were said to show a driver using the automated system in the driving simulator (Figure 2). After the videos, participants had to answer the trust questionnaire again (t_2) and subsequently questionnaires for several personality variables were provided (see below). In the end, participants were informed that the reports on the function tests were of fictional nature, thanked and dismissed. The survey was programmed in SoSci survey (Leiner, 2014). At all times when trust was assessed, also acceptance was measured, which is not in the scope of this paper (reported in Forster, Kraus, Feinauer, & Baumann, 2018). Data was collected between March 2nd and 9th 2018. Participants who completed the survey were given financial compensation in return for their participation.

Video Material

Each participant watched four videos of an automated driving system in the static driving simulator of the Human Factors Department at Ulm University (SILAB; WIVW GmbH). The HMI consisted of a center touchscreen showing system status and take-over requests (TORs; button press). The following scenarios were shown (each about 30 s):

1. **Activation:** The driver initiates the automation by button press;
2. **Overtaking:** Automated take-over and TOR due to work zone;

Figure 1. Study procedure

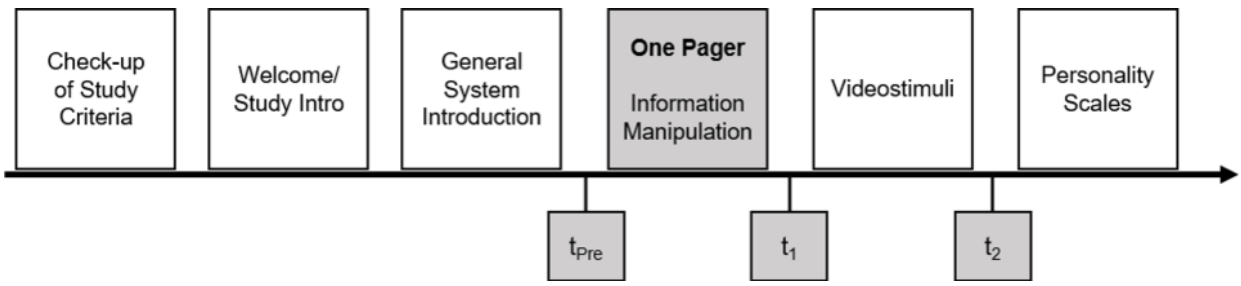


Figure 2. Setup of the driving simulator at Ulm University in the video stimuli. On the left, the driver activates the CAD function. On the right, the CAD issues a TOR.



3. **Missing lane markings:** System-initiated TOR and driver takes over control;
4. **System deactivation:** Driver deactivates the system.

Materials and Questionnaires

All questionnaires were based on validated psychological scales and were answered with 7-point Likert scales from 1 ("not at all") to 7 ("very much"). Trust in automation was assessed with the Automation Trust Scale (ATS; Jian, Bisantz, & Drury, 2000) with 12 items, which were averaged into one scale mean. For measuring NFC, a shortened 8-item version of the German scale by Bless and colleagues (1994) was used (item selection by item difficulty and factor loadings). Regulatory focus was assessed with the 10 items by Fellner and colleagues (2007), embedded in the context of 22 distractor items. Scale means for promotion and prevention focus were calculated. Materialism was measured with the scale by Richins and Dawson (1992). For each of the three factors, three items were selected on the basis of factor loadings. The perfect automation scheme was assessed with a translated German version of the scale by Merritt and colleagues (2015). For the factor "high expectations" four and for the factor "all-or-none thinking" three items were selected on the basis of factor loadings. In order to check for successful manipulation of independent variables, manipulation checks were included. For the OEM brand, participants were asked to select the brand from a dropdown list. For reliability information participants had to rate how reliable the presented system was compared to other systems (5-point Likert from 1 ("worse") via 3 ("comparable") to 5 ("better")).

Participants

The study sample was recruited online with a German online panel. Altogether, $N = 623$ participants completed the questionnaire. Participants were excluded if the manipulation checks indicated poor memory for the study stimuli (wrong OEM brand selected; for low reliability a value of ≥ 3; for high reliability a value of ≤ 3) or if their processing time indicated speeding (half the time of the sample median (< 24 min)). After exclusion the final study sample consisted of $n = 519$ participants (55.3% female, 44.7% male). The average age of participants was 45.5 years ranging from 18 to 80 ($SD = 14.35$). Gender distribution and mean age of the study groups were in a comparable range to the overall sample (41.19 to 50.22 years; see Forster et al., 2018). No outliers were detected for the three times of trust measurement with the criteria of extreme z scores exceeding +/-3 (e.g. Aggarwal, 2013).

RESULTS

Data was analyzed with Matlab 7.11.1 (R2010b) and IBM SPSS Statistics Version 23.

Data Preparation and Manipulation Checks

All means, standard deviations and Cronbach's α are shown in Table 6. Due to the exclusion of participants (see above), sample sizes for the study groups were not balanced (between n = 35-45), and thus we calculated type-III sum of squares (Howell, 2007). For the trust pretest (t_{pre}), the study groups did not differ in their trust ratings as indicated by a two-way ANOVA with reliability and the single OEM brands as between factors (F_{max} = 1.08; all $p > .05$). Additionally, Table 1 presents the trust means of the ATS scale for the study groups at t_1 and t_2.

For data analyses of the study hypotheses on the effects of the information manipulations, according to the findings of the pilot study the brands were grouped into three OEM brand reputation groups (below average/average/above average) and the baseline condition was left out of the analysis resulting in a n = 435. As a manipulation check for the brand reputation groups, a subset of 13 items of the brand image questionnaire from the pilot study (see Forster et al., 2018) referencing to the presented brand was used. A one-factorial ANOVA on mean image with the OEM brand reputation groups as between factor revealed a significant main effect for brand reputation, $F(2, 432) = 27.49$, $p < .01$, $\eta p^2 = .113$ ($M_{below\ average} = 4.03$, $SD = 1.09$; $M_{average} = 4.41$, $SD = 1.16$; $M_{above\ average} = 4.95$, $SD = 0.94$) indicating a successful manipulation of the brand reputation groups. For the analyses on the interrelationships of the personality traits and trust in automation, the full sample with n = 519 was used.

In Table 2 means and *SD*s for trust at t_1 and t_2 for the remaining six study groups are depicted. In order to test hypothesis 1 and 2, two three factorial ANOVAs with the between factors reliability (low vs high reliability), OEM brand reputation (below average vs average vs above average) and NFC (median-split) with trust at t_1 and t_2 as dependent measure was conducted. The median for NFC in the sample was $Mdn = 4.75$ (low NFC: $n = 240$, $M_{NFC\ low} = 4.10$ ($SD = 0.54$); high NFC: $n = 195$, $M_{NFC\ high} = 5.44$ ($SD = 0.48$).

Effect of Reliability and OEM Brand

For NFC neither at t_1 nor at t_2 a significant main effect was found. For reliability ANOVAs revealed significant main effects for trust both at t_1 and t_2 (see Figure 3). At t_2 participants of the low reliability group showed less trust in the system ($M = 3.23$, $SD = 1.06$) as compared to the high reliability group ($M = 3.91$; $SD = 1.07$), $F(1, 423) = 48.14$, $p < .001$, $\eta p^2 = .102$). The same pattern was found at t_2, $F(1, 423) = 16.34$, $p < .001$, $\eta p^2 = .037$ (see Table 3). These findings conclusively support H1.

In terms of the provided OEM brand reputation information, a significant main effect was found at t_1, $F(2, 423) = 3,17$, $p < .05$, $\eta p^2 = .015$. A post hoc test with Tukey's HSD revealed that the only significant group difference was between the above average reputation (M = 3.71; SD = 1.14) and the

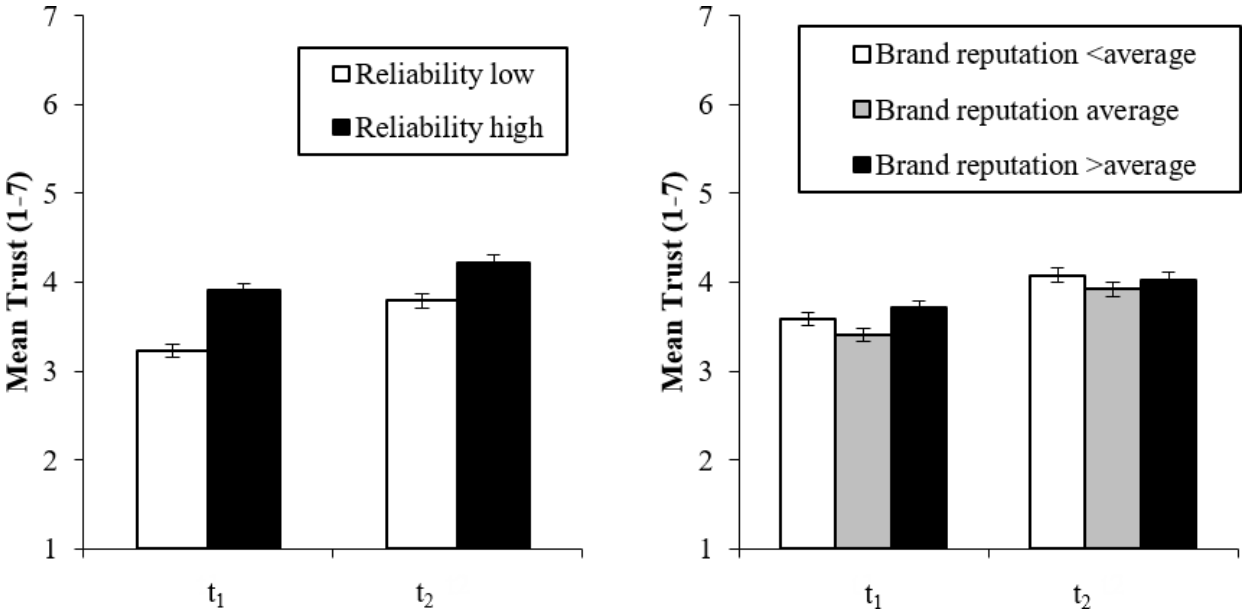
Table 1. Mean trust ratings and standard deviations (SD) for all study conditions

	Reliability		BL	Dacia	Kia	Skoda	Opel	Tesla	BMW
Trust t_1	low	M (SD)	3.54 (1.07)	3.30 (1.01)	3.27 (1.02)	2.87 (1.05)	3.12 (1.06)	3.51 (1.00)	3.32 (1.17)
	high	M (SD)	3.62 (1.25)	3.71 (0.91)	4.05 (0.96)	3.86 (1.25)	3.82 (1.07)	3.93 (0.98)	4.08 (1.25)
Trust t_2	low	M (SD)	4.02 (0.98)	3.88 (1.12)	3.82 (1.36)	3.58 (1.22)	3.74 (1.18)	3.82 (1.20)	3.93 (1.27)
	high	M (SD)	3.95 (1.28)	4.07 (0.93)	4.53 (1.01)	4.28 (1.35)	4.14 (1.05)	4.07 (1.07)	4.26 (1.35)

Table 2. Mean trust ratings and standard deviations (SD)at t_1 and t_2 for all study conditions

	Reliability		Brand Reputation		
			<Average	Average	>Average
Trust t_1	low	M (SD)	3.29 (1.01)	2.99 (1.06)	3.41 (1.09)
	high	M (SD)	3.88 (0.94)	3.84 (1.15)	4.00 (1.12)
Trust t_2	low	M (SD)	3.85 (1.23)	3.66 (1.19)	3.87 (1.23)
	high	M (SD)	4.30 (1.00)	4.20 (1.19)	4.16 (1.21)

Figure 3. Bar charts indicate pooled group means for the reliability information groups (left side) and the OEM brand reputation information groups (right side). Error bars indicate 1 standard error.



average reputation group (M = 3.40; SD = 1.18). At t2 no significant main effect for OEM reputation was found indicating that the effect of reputation information diminishes over time. These findings provide partial support for H2 that the mere presentation of an OEM brand leads to attributional processes calibrating trust in the direction of a brand's reputation.

Interaction Between Information Use and Need for Cognition

In terms of differential information usage, the two-way and three-way interactions of NFC and the respective information manipulations were assessed. Neither the three-way interactions nor the interactions between OEM brand reputation and NFC were significant at both times of measurement. Thus, H3.2 had to be rejected. On the other hand, H3.1 was supported by a significant interaction between reliability information and NFC at both times of measurement (t_1: $F(1, 423) = 6.84$; $p < .01$; $\eta p^2 = .016$; t_2: $F(1, 423) = 4.541$; $p < .05$; $\eta p^2 = .011$; Table 3 and 4). Figure 4 illustrates that while systems described as highly reliable are trusted more in both groups, this difference is more extreme in the group with high NFC. Thus, people with higher NFC seem to calibrate their trust more strongly in accordance to reliability information. In order to gain a more detailed picture of how

Table 3. Results of the three-way ANOVAs for trust at t_1 and t_2 with need for cognition median-split (NFC_ms)

Dependent Variable	Factor	*df* (Error)	*F*	*p*	η_p^2
Trust t_1	Reliability	1 (423)	48.14	.000	0.102
	Brand Reputation	2	3.17	.043	0.015
	NFC_ms	1	1.72	.191	0.004
	Reliability*Brand Reputation	2	0.68	.510	0.003
	Reliability*NFC_ms	1	6.84	.009	0.016
	Brand Reputation*NFC_ms	2	0.55	.579	0.003
	Reliability*Brand Reputation*NFC_ms	2	0.67	.511	0.003
Trust t_2	Reliability	1 (423)	16.34	.015	0.037
	Brand Reputation	2	0.58	.559	0.003
	NFC_ms	1	2.72	.100	0.006
	Reliability*Brand Reputation	2	0.30	.743	0.001
	Reliability*NFC_ms	1	4.54	.034	0.011
	Brand Reputation*NFC_ms	2	0.19	.826	0.001
	Reliability*Brand Reputation*NFC_ms	2	0.05	.953	0.000

different levels of NFC are related to information used in trust formation in CAD systems, NFC was additionally split into the four quartile groups (Q_1 = [min, 4.125]; Q_2 = [4.125, 4.75]; Q_3 = [4.75, 5.25]; Q_4 = [5.25, max]; Table 5). Three-way ANOVAS for trust at t_1 and t_2 were conducted. Again, no main effects for NFC was found. For the interaction between NFC and reliability information the F-test was significant at t_1, $F(3, 411) = 2.82$, $p < .05$, $\eta p^2 = .020$, but missed significance at t_2, $F(3, 411) = 1.69$, $p = .168$, $\eta p^2 = .012$, while the effect size ranges in a similar area as for the median-split. Figure 4 shows plots of the mean trust scores for the reliability X NFC quartile interaction for each point of measurement. Both plots resemble each other well and indicate a non-linear relationship between NFC and utilization of the reliability information. Differences are especially emphasized between Q_2 and Q_3.

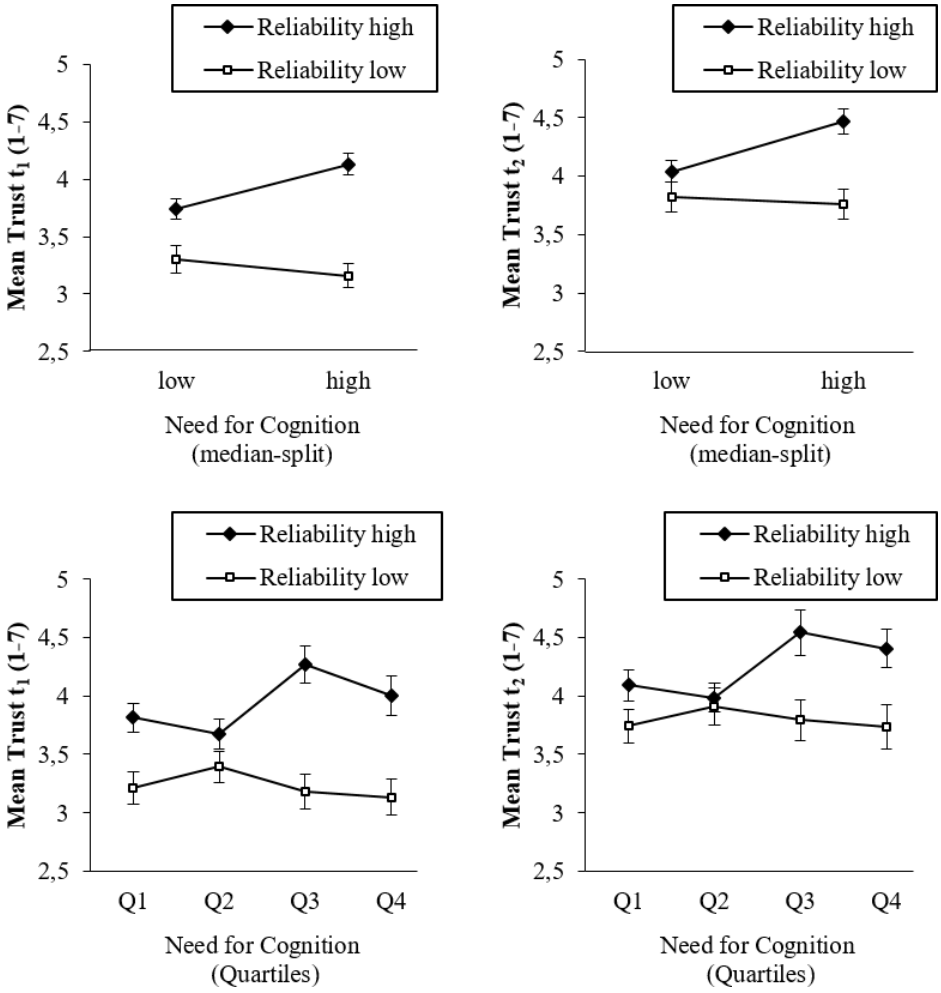
Relationships of Trust in Automation and the Personality Scales

In Table 6 correlations of the personality traits with trust in automation are depicted. While trust in automation is not predicted by NFC, for the remaining traits significant relationships with trust in automation were found. Promotion and prevention focus showed small and materialism medium significant correlations to trust at all times of measurement. For the perfect automation scheme facet "high expectations" high correlations were found with trust at all times of measurement. For the other facet "all or none", small negative significant correlations were found. Taken together, H4 was supported.

DISCUSSION

The presented online study investigated the influence of two kinds of a priori information (CAD reliability and OEM brand) on attitude formation processes associated with trust calibration towards a CAD system. Additionally, the role of several personality traits for trust in automation at the pre-usage

Figure 4. Line graphs indicate pooled group means for the reliability information and NFC-groups interaction (upper part: median-split; lower part: quartiles). Error bars indicate 1 standard error.



phase was investigated - particularly, the moderating influence of NFC for the relative consideration of different information types in trust formation.

Prior information both on the reliability of an automated driving system as well as on the OEM brand influenced trust in the system prior to any actual interaction. Additionally, in line with the ELM, the individual disposition for NFC influences how the provided information is used to establish a trust attitude towards the automation (H3). In regard to H1, a CAD function introduced as unreliable with a higher number of system failures was trusted significantly less as compared to one that was described as reliable with a small number of failures (see Table 4, Figure 3). The same holds true for the mere presentation of a brand in the description of the system. Study participants significantly trusted OEM brands with a below average reputation less than those with a reputation in the average range (Figure 3). Interestingly, no significant trust differences were found for the above average reputation group. While this study provides initial support for the role of brand image in the adoption of CAD, the exact mechanism needs further investigation. Taken together, both types of information – heuristic brand information and analytic reliability information – seem to play a role in trust calibration corroborating earlier findings on the importance of prior information for the adaption of automated technology (e.g. Körber et al., 2018; Biassoni et al., 2016; Beggiato & Krems, 2013; Hergeth et al., 2017). While both kinds of information lead to a short-term attitude change in the direction of the provided information at t_1, only the reliability information still had an effect at t_2. It can be concluded that the effect of the OEM brand information is weaker and more

Table 4. Mean trust ratings and standard deviations (SD)at t_1 and t_2 for the study conditions with NFC median-split

	Reliability		NFC Low	NFC High
Trust t_1	low	M (SD)	3.30 (1.04)	3.16 (1.09)
	high	M (SD)	3.74 (0.99)	4.13 (1.13)
Trust t_2	low	M (SD)	3.82 (1.14)	3.76 (1.31)
	high	M (SD)	4.04 (0.99)	4.47 (1.26)

Table 5. Mean trust ratings and standard deviations (SD) at t1 and t2 for the study conditions with NFC quartile groups

	Reliability		Need for Cognition			
			Q_1	Q_2	Q_3	Q_4
Trust t_1	low	M (SD)	3.21 (1.08)	3.39 (0.99)	3.18 (1.02)	3.13 (1.15)
	high	M (SD)	3.81 (0.96)	3.68 (1.02)	4.27 (1.07)	4.00 (1.17)
Trust t_2	low	M (SD)	3.74 (1.12)	3.91 (1.16)	3.79 (1.21)	3.74 (1.40)
	high	M (SD)	4.09 (1.01)	3.99 (0.97)	4.54 (1.24)	4.40 (1.29)

Table 6. Number of items, means, standard deviations (SD), Cronbach's α and bivariate correlations for the investigated personality traits and trust in automation

		# Items	M	SD	Alpha	1	2	3	4	5	6	7	8
1	Need for Cognition	8	4.70	0.86	.761	-							
2	Promotion Focus	5	4.92	0.73	.595	.369**	-						
3	Prevention Focus	5	4.55	0.72	.390	-.110*	.258**	-					
4	Materialism	9	3.31	1.02	.817	-.196**	-.027	.281**	-				
5	PAS - High Expectations	4	2.29	0.95	.892	-.143**	.128**	.271**	.233**	-			
6	PAS - All or none	3	2.88	0.86	.657	-.121**	.061	.123**	.037	.154**	-		
7	Trust t_{pre}	12	4.06	1.15	.933	.049	.183**	.129**	.238**	.599**	-.090*	-	
8	Trust t_1	12	3.57	1.12	.928	.029	.108*	.127**	.212**	.613**	-.131**	.712**	-
9	Trust t_2	12	4.00	1.18	.931	.033	.143**	.139**	.191**	.605**	-.159**	.737**	.786**

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

prone to change over time. This is line with the respective characteristic of the provided information as conceptualized in the ELM. Also in line with the ELM, NFC and thus the motivation to process provided information thoroughly, was found to moderate the use of the central route of processing and to influence the evaluation of the automation in direction of the provided arguments (Cacioppo et al., 1986) as indicated by a significant interaction between system reliability and NFC (median-

split; Table 4, Figure 4). Accordingly, the group of participants with high NFC values more strongly calibrated their trust in the system in line with the reliability information provided. This interaction was still found to be significant at the second time of measurement (for the median-split, but not for the quartiles), thus supporting the notion of a more stable attitude change as a result of deeper processing. As indicated by an additional analysis of the quartiles of NFC, the relationship between NFC and the extent of consideration of argument strength might not be of linear nature (Table 5, Figure 4). The exact interrelation between NFC and information use in trust formation processes in the CAD domain might be investigated further in future studies.

This pattern of effects supports a process of attitude change through a priori information that largely aligns with the ELM. Both, arguments on the central and peripheral route seem to influence trust in automated vehicles. This is in line with Lee and See (2004) who argue for both analytical and analogical processes playing a role in attitude formation before actual system use and trust calibration (see also Forster et al., 2018) and thus provides further theoretical support for the conceptualization of trust in automation as an attitude (e.g. Lee & See, 2004).

Additionally, in line with earlier studies (e.g. Merrit & Illgen, 2008; Szalma & Taylor, 2011; Payre, Cestac, & Delhomme, 2014) the present research provides further evidence for the role of personality traits in trust formation. There is a positive relationship between materialism and trust in automation (Table 6), which may be originated in the higher importance of luxury goods for materialists (Belk, 1984). Second, both factors of regulatory focus (Fellner et al., 2007) are positively correlated to trust in automation prior to system use. This underlines the role of individual representations of risks and gains of automated driving in trust formation, which should be further investigated in the future. Third, both factors of the perfect automation scheme (Merritt et al., 2015) were found to predict trust in automation. The factor "high expectations", resembling a priori acceptance towards automated driving systems, showed the highest correlation with trust. This provides further evidence for the role of specific attitudes in the adaption of automated technology (e.g. Merritt, Heimbaugh, LaChapell, & Lee, 2013). Taken together, study findings underline the importance of personality traits of different specificity for the formation of trust in automated systems and should thus be considered in theory and practice.

In conclusion, the reported findings support an interaction of the provided information and individual perception and expectations as reflected in personality traits during trust calibration.

Practical Implications

The reported findings underline the importance of the information provided prior to actual system use for trust calibration and the adoption of CAD. This relates to every kind of a priori information available to drivers: word-of-mouth, news, blogs, marketing communication, driver education and tutorials as well as documentation of CADs in operation manuals. Therefore, automated driving development not only needs to consider function and HMI design but also to focus on consumer education (e.g. Forster, Hergeth, Naujoks, Krems, & Keinath, 2019 for L3 automated driving). The NHTSA's AV policy (2017), for example, states that consumer education approaches should be developed as these constitute an "imperative for increased safety during the deployment of ADS" (p. 15).

The presented study findings suggest that the character and boundary conditions of information processing in attitude formation might be fruitful to be considered in designing information available to drivers prior (and during) system use. First of all, the kind of provided information is important and should be suitable for the situation it is likely to be processed. Thereby, small changes in the information representation might be of high impact on attitude formation. While more heuristic information is most effective in situations, in which people are not able or willing to invest in costly cognitive demanding information processing, they also tend to produce rather weak attitude changes that do not persist over time. In order to evoke a stable and robust trust level reflecting the actual capabilities of the automation, it seems recommendable to use mainly information with high argument quality to be processed via the central route. This entails realistic information on functionality, conditions and

limitations of the automation with the goal to produce an adequate mental model and a calibrated level of trust preventing both mis- and disuse of the automated driving system. The presented findings especially indicate an opportunity in this strategy for OEMs that are not perceived overly positive as they can use the central route to safeguard their systems against a negative attitude transfer from their brand to the system. In this regard, it is recommendable to consider the boundary conditions for central information processing as defined in the ELM. These are besides others sufficient time and cognitive resources for message processing, motivation to do so as well as an easy-to-process manner of information presentation (for a more detailed discussion see Petty & Cacioppo, 1986; Lien, 2001).

Furthermore, study findings on the relationships of personality traits and trust in automation, show that not only the character of provided information affects information use and attitude formation in trust calibration, but also that individual processes are at play, that should be respected in both information and UI design for CAD. For practitioners it is important to know who they are designing for as different types of CAD users may exist and respective communication might be necessary.

In regard to the role of persuasion in trust calibration in CAD, it is of utmost importance to always aim at a calibrated rather than a maximum level of trust as the goal of driver education. In this, the mechanisms of information processing and attitude formation in trust calibration might help to design effective communication with drivers before they actually use an automated driving system for the first time in order to facilitate safe and efficient system use right from the start.

Limitations and Future Research

The present study has some limitations. First, while providing a large sample, the online survey setting does not support realistic interaction and experience for the participants with a CAD function. The present study setup enhanced realism by providing not only a description but also video footage of a CAD function. Future research might investigate trust formation in a more realistic setup. Second, the present approach ordered two OEM brands into each of the three OEM brand reputation groups (i.e., above average, average, below average) on the basis of findings of a cluster analysis from a pilot study as described in Forster et al. (2018). In future studies, different brands could be used, and further research might be conducted to get a better understanding of the relative impact of different dimensions of brand characteristics on trust calibration. Third, the routes of the ELM are not exclusive and future research might investigate the relative importance of the routes in trust calibration and how attributes of the provided information influence trust formation. Furthermore, besides NFC further variables hypothesized to influence the selection of the processing route (cognitive demands, motivation, etc.) could be investigated in future studies.

CONCLUSION

The current study strongly supports the importance of a priori information on trust formation towards CAD. Thereby, theories on psychological processes involved in attitude formation and persuasion provide a promising direction for theoretical progress, research and design in regard to trust calibration. For these endeavours, informed consideration of different information characteristics, situational and cognitive boundary factors, as well as personality factors, seems fruitful.

REFERENCES

Abraham, H., Seppelt, B., Mehler, B., & Reimer, B. (2017). What's in a Name: Vehicle Technology Branding & Consumer Expectations for Automation. In S. Boll, B. Pfleging, I. Politis, D. Large, & B. Domnez (Eds.), *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 226-234). New York: ACM. doi:10.1145/3122986.3123018

Aggarwal, C. C. (2013). *Outlier analysis*. New York: Springer. doi:10.1007/978-1-4614-6396-2

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 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

Belk, R. W. (1984). Three scales to measure constructs related to materialism: Reliability, validity, and relationships to measures of happiness. In T. C. Kinnear (Ed.), *Advances in consumer research* (pp. 291–297). Provo: Association for Consumer Research.

Biassoni, F., Ruscio, D., & Ciceri, R. (2016). Limitations and automation. The role of information about device-specific features in ADAS acceptability. *Safety Science*, *85*, 179–186. doi:10.1016/j.ssci.2016.01.017

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F., & Schwarz, N. (1994). Need for Cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben [Need for cognition: A scale measuring engagement and happiness in cognitive tasks]. *Zeitschrift für Sozialpsychologie*, *25*(2), 147–154.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307. doi:10.1207/s15327752jpa4803_13 PMID:16367530

Cacioppo, J. T., Petty, R. E., Kao, C. F., & Rodriguez, R. (1986). Central and peripheral routes to persuasion: An individual difference perspective. *Journal of Personality and Social Psychology*, *51*(5), 1032–1043. doi:10.1037/0022-3514.51.5.1032

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology*, *45*(4), 805–818. doi:10.1037/0022-3514.45.4.805

Carroll, J. M., & Olson, J. R. (1987). Mental models in human-computer interaction. In M. Helander (Ed.), *Handbook of Human-Computer Interaction* (pp. 45–65). Amsterdam: North Holland.

Castelfranchi, C. (1998). Modelling social action for AI agents. *Artificial Intelligence*, *103*(1-2), 157–182. doi:10.1016/S0004-3702(98)00056-3

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York: Guilford Press.

Chen, S., Shechter, D., & Chaiken, S. (1996). Getting at the truth or getting along: Accuracy-versus impression-motivated heuristic and systematic processing. *Journal of Personality and Social Psychology*, *71*(2), 262–275. doi:10.1037/0022-3514.71.2.262

Cialdini, R. B., Petty, R. E., & Cacioppo, J. T. (1981). Attitude and attitude change. *Annual Review of Psychology*, *32*(1), 357–404. doi:10.1146/annurev.ps.32.020181.002041

Cohen, M. S. (2000). Training for trust in decision aids, with applications to the rotorcraft's pilot's associate. *Presented at the International Conference on Human Performance, Situation Awareness, and Automation*, Savannah, GA. Retrieved from http://www.cog-tech.com/papers/Trust/Adaptive%20Automation%20abstract%20final.pdf

Fellner, B., Holler, M., Kirchler, E., & Schabmann, A. (2007). Regulatory Focus Scale (RFS): Development of a Scale to Record Dispositional Regulatory Focus. *Swiss Journal of Psychology*, *66*(2), 109–116. doi:10.1024/1421-0185.66.2.109

Forster, Y., Hergeth, S., Naujoks, F., Krems, J., & Keinath, A. (2019). User Education in Automated Driving: Owner's Manual and Interactive Tutorial Support Mental Model Formation and Human-Automation Interaction. Information, 10(4), 143.

Forster, Y., Kraus, J., Feinauer, S., & Baumann, M. (2018). Calibration of Trust Expectancies in Conditionally Automated Driving by Brand, Reliability Information and Introductionary Videos: An Online Study. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 118-128). New York: ACM. doi:10.1145/3239060.3239070

Hergeth, S., Lorenz, L., & Krems, J. F. (2017). Prior familiarization with takeover requests affects drivers' takeover performance and automation trust. *Human Factors*, *59*(3), 457–470. doi:10.1177/0018720816678714 PMID:27923886

Higgins, E. T. (1997). Beyond pleasure and pain. *The American Psychologist*, *52*(12), 1280–1300. doi:10.1037/0003-066X.52.12.1280 PMID:9414606

Hock, P., Kraus, J., Babel, F., Walch, M., Rukzio, E., & Baumann, M. (2018). How to Design Valid Simulator Studies for Investigating User Experience in Automated Driving: Review and Hands-On Considerations. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 105-117). New York: ACM. doi:10.1145/3239060.3239066

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. doi:10.1177/0018720814547570 PMID:25875432

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. doi:10.1207/S15327566IJCE0401_04

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 PMID:28958427

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, *50*(1), 569–598. doi:10.1146/annurev.psych.50.1.569 PMID:15012464

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (in press). The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction and System Transparency.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. doi:10.1080/00140139208967392 PMID:1516577

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. doi:10.1518/hfes.46.1.50.30392 PMID:15151155

Leiner, D. J. (2014). SoSci survey. Retrieved from https://www.soscisurvey.de

Lewicki, R. J., & Bunker, B. B. (1996). Developing and Maintaining Trust in Work Relationships. In R. M. Kramer & T. R. Tyler (Eds.), *Trust in Organizations: Frontiers of Theory and Research* (pp. 114–139). Thousand Oaks, CA: SAGE Publications. doi:10.4135/9781452243610.n7

Lien, N. H. (2001). Elaboration Likelihood Model in Consumer Research: A Review. *Proceedings of the National Science Council Part C*, *11*(4), 301–310.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, *55*(3), 520–534. doi:10.1177/0018720812465081 PMID:23829027

Merritt, S. M., & Illgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194–210. doi:10.1518/001872008X288574 PMID:18516832

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors*, *57*(5), 740–753. doi:10.1177/0018720815581247 PMID:25882304

Mowen, J. C. (2000). *The 3M model of motivation and personality: Theory and empirical applications to consumer behavior*. Berlin: Springer. doi:10.1007/978-1-4757-6708-7

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5-6), 527–539. doi:10.1016/S0020-7373(87)80013-5

NHTSA. (2017). *Automated Driving Systems 2.0. A Vision for Safety*. U.S. Department of Transportation.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. doi:10.1518/001872097778543886

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*, 252–263. doi:10.1016/j.trf.2014.04.009

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), (pp. 1–24). Advances in experimental Social Psychology New York: Academic Press.

Richins, M. L., & Dawson, S. (1992). A consumer values orientation for materialism and its measurement: Scale development and validation. *The Journal of Consumer Research*, *19*(3), 303–315. doi:10.1086/209304

SAE. (2016). *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems, J3016R*. Warrendale: SAE International.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors*, *58*(3), 377–400. doi:10.1177/0018720816634228 PMID:27005902

Szalma, J. L., & Taylor, G. S. (2011). Individual Differences in Response to Automation: The Five Factor Model of Personality. *Journal of Experimental Psychology. Applied*, *17*(2), 71–96. doi:10.1037/a0024170 PMID:21688932

Walch, M., Mühl, K., Kraus, J., Stoll, T., Baumann, M., & Weber, M. (2017). From car-driver-handovers to cooperative interfaces. In G. Meixner & C. Müller (Eds.), *Automotive User Interfaces: Creating Interactive Experiences in the Car* (pp. 273–294). Cham: Springer. doi:10.1007/978-3-319-49448-7_10

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

Yannick Forster received his master's degree in Psychology from the University of Wuerzburg in 2016. In 2017, he started a PhD position at Chemnitz University of Technology in cooperation with the BMW Group working towards methodological development for Human-Machine Interface evaluation in the area of automated driving.

Sebastian Hergeth received his doctoral degree in Psychology from the Technical University Chemnitz in 2016. In 2016 he was hired as a Usability Specialist by the BMW Group and is currently working on the evaluation of automotive human-machine interfaces.

Martin Baumann is head of the department Human Factors at Ulm University. His main research interests are the psychological basis of human-machine interaction in different domains, mainly traffic, human-robot interaction, interaction with intelligent systems, and the development and validation of concepts of cooperative human-machine systems. Specifically, his main research topics are trust, comprehension of dynamic situations and intelligent systems, aspects of cooperative human-machine interaction, and persuasive technologies.

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

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

Copyright: ©2019 Human Factors and Ergonomics Society

Check for updates

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. XX, No. X, Month XXXX, pp. 1–19
DOI: 10.1177/0018720819853686
Article reuse guidelines: sagepub.com/journals-permissions
Copyright © 2019, Human Factors and Ergonomics Society.

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.

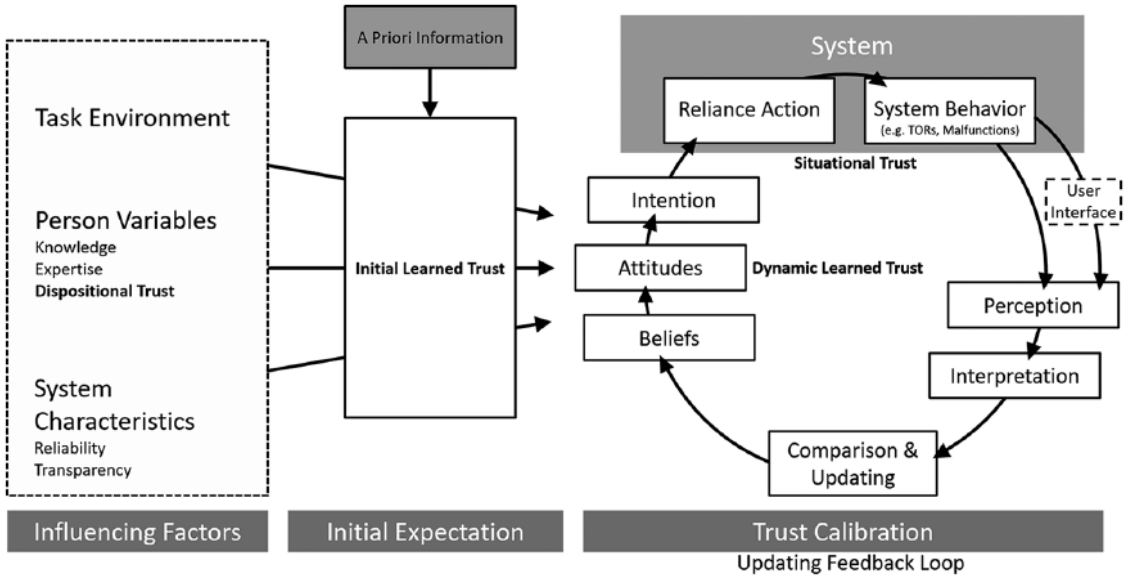


Figure 1. Model of dynamic trust calibration integrating different stages of trust attitude formation.
Note. Different trust forms reflect attitudes at different stages of this process. Initial learned trust is built up before the actual interaction with a system along the available information. In the 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

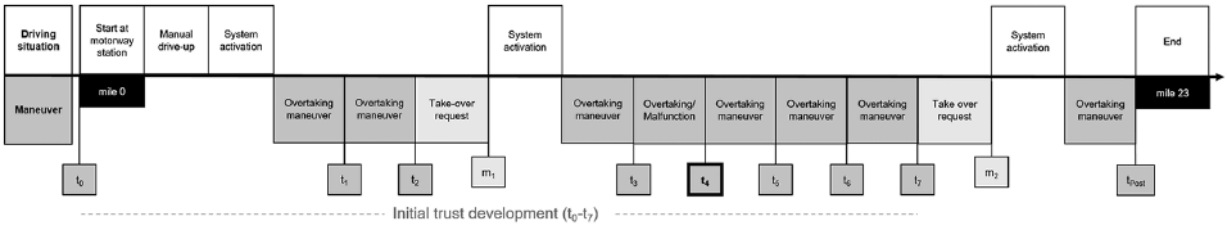


Figure 2. Experimental drive of Study 1.



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

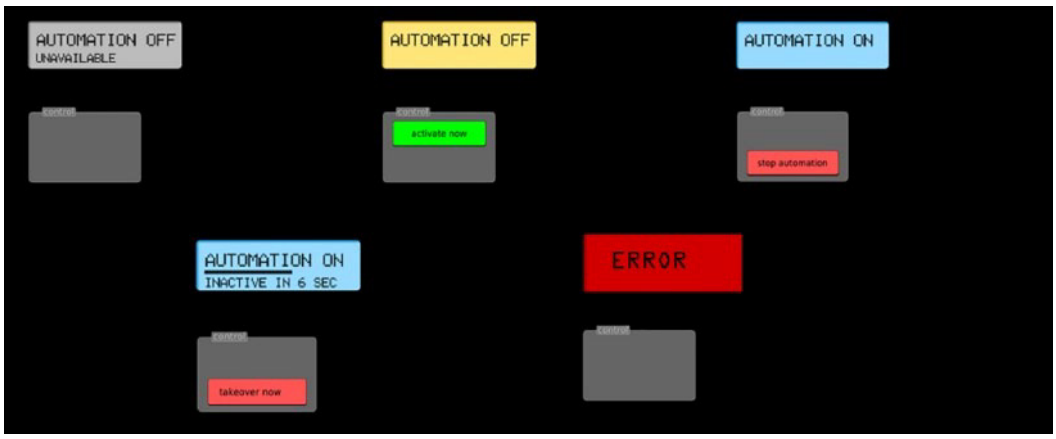


Figure 4. Interface for enabling and disabling the automation located in the center console.
Note. Each display consisted of information on the automation status in the upper part and a dynamic control panel with touch icons to turn the automation on or off in the lower part. In the upper row, the different 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.

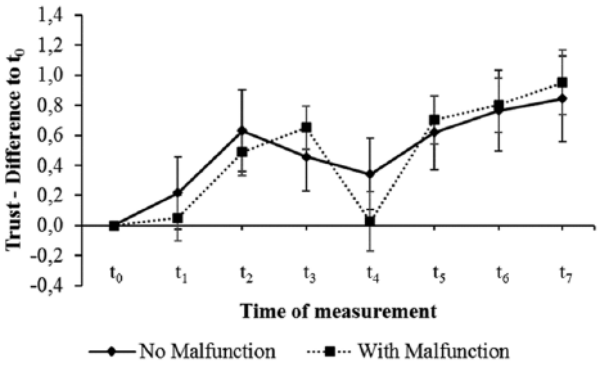


Figure 5. Trust development over the course of the experiment in the groups with and without malfunction.
Note. Before t_4 the malfunction occurred in the respective condition. t_0 serves as a baseline to facilitate comparison of the trust curves. Error bars indicate ±1 *SE*.

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.

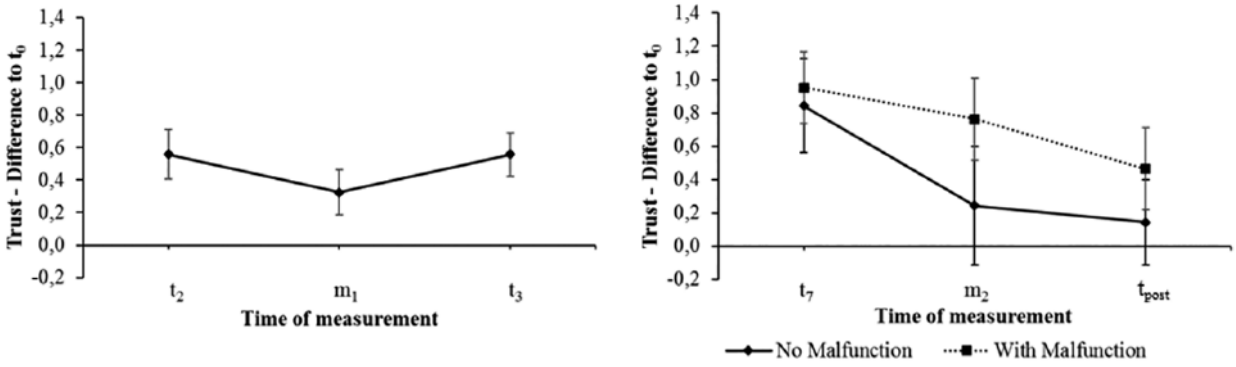


Figure 6. Trust development prior to, during, and after the first TOR (left) and the second TOR (right).
Note. t_0 serves as a baseline to facilitate comparison of the trust curves. Error bars indicate ±1 *SE*.
TOR = take-over request.

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 (0.57)	5.66 (0.73)	5.72 (0.94)	5.91 (0.62)	5.98 (0.73)	6.09 (0.79)	6.11 (0.76)	6.17 (0.71)	6.22 (0.66)	6.22 (0.67)	6.03 (0.66)
MF–/HT	4.47 (0.69)	5.13 (1.20)	5.60 (0.83)	5.67 (0.92)	5.73 (1.11)	5.61 (0.97)	5.93 (0.91)	5.85 (1.00)	6.04 (0.78)	5.98 (0.91)	5.81 (1.09)
MF+/LT	4.64 (1.23)	4.95 (1.14)	5.27 (1.13)	5.42 (1.06)	5.60 (0.98)	5.13 (1.22)	5.55 (1.09)	5.75 (1.06)	5.84 (1.04)	5.96 (0.92)	5.45 (1.12)
MF+/HT	5.20 (0.87)	5.25 (0.84)	5.53 (0.88)	5.75 (0.85)	5.75 (0.64)	5.67 (0.61)	5.83 (0.58)	5.82 (0.59)	5.90 (0.59)	5.90 (0.56)	5.81 (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

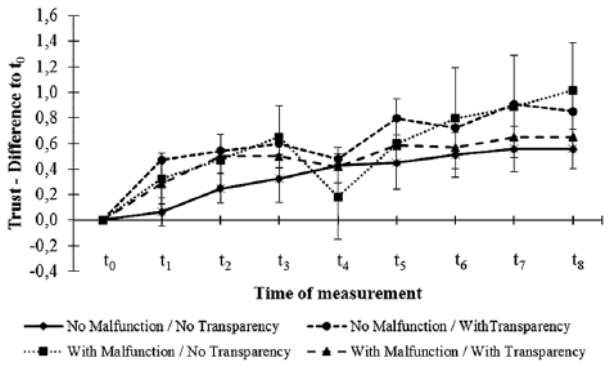


Figure 7. Trust development over the course of the experiment in the different groups with and without malfunction and transparency.
Note. Before t_4 the malfunction occurred in the respective condition. t_0 serves as a baseline to facilitate comparison of the trust curves. Error bars indicate ±1 *SE*.

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

Erklärung

Ich versichere hiermit, dass ich die Arbeit selbständig angefertigt habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die zur Zeit gültige Satzung der Universität Ulm zur Sicherung guter wissenschaftlicher Praxis beachtet habe (§ 8 Abs. 1 Nr. 5 Rahmenpromotionsordnung).

Ich bin damit einverstanden, dass die Dissertation auch zum Zweck der Überprüfung der Einhaltung allgemein geltender wissenschaftlicher Standards benutzt wird, insbesondere auch unter Verwendung elektronischer Datenverarbeitungsprogramme (§ 8 Abs. 1 Nr. 8 Rahmenpromotionsordnung).

Ulm, den 12.05.2020 ……

(Unterschrift)