Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions

Magdalena Wischnewski
magdalena.wischnewski@tu-dortmund.de
Research Center for Trustworthy Data Science and Security
Dortmund, Germany

Nicole Krämer
nicole.kraemer@uni-due.de
University of Duisburg-Essen
Duisburg, Germany
Research Center for Trustworthy Data Science and Security
Dortmund, Germany

Emmanuel Müller
emmanuel.mueller@cs.tu-dortmund.de
Technical University Dortmund
Dortmund, Germany
Research Center for Trustworthy Data Science and Security
Dortmund, Germany

ABSTRACT

Trust has been recognized as a central variable to explain the resistance to using automated systems (under-trust) and the overreliance on automated systems (over-trust). To achieve appropriate reliance, users' trust should be calibrated to reflect a system's capabilities. Studies from various disciplines have examined different interventions to attain such trust calibration. Based on a literature body of 1000+ papers, we identified 96 relevant publications which aimed to calibrate users' trust in automated systems. To provide an in-depth overview of the state-of-the-art, we reviewed and summarized measurements of the trust calibration, interventions, and results of these efforts. For the numerous promising calibration interventions, we extract common design choices and structure these into four dimensions of trust calibration interventions to guide future studies. Our findings indicate that the measurement of the trust calibration often limits the interpretation of the effects of different interventions. We suggest future directions for this problem.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **HCI design and evaluation methods**; **User studies**;

KEYWORDS

trust calibration, automation, empirical studies, warranted trust, survey

ACM Reference Format:
Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3544548.3581197

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581197

1 INTRODUCTION

Automation has become ubiquitous in our everyday lives. From basic automated processes which take over rudimentary, simple tasks, relying on repetitive or static rules to highly sophisticated systems which act intelligently based on available data, powered by deep learning algorithms, automation can guide credit scoring and risk assessments [42], transportation [52, 128], medical diagnostics [38], hiring decisions [78], and judicial sentencing [46] and recidivism [21]. Moreover, automated systems help people decide what to wear [125], what to watch [1] or listen to [115], which news to read [37], and whom to date [134]. As diverse as the application contexts are functionalities that automated systems hold. They can be used to discover patterns that would otherwise go unnoticed, assist and aid human judgments, and make sophisticated predictions and recommendations. The hopes and promises of such systems presage greater efficiency and accuracy, taking over routine tasks by overcoming individuals' fundamental physical and cognitive limitations.

Besides the functional and technical diversity, automated systems appear under a myriad of terminologies. To encompass a wide range of different systems which can, but do not have to, be powered by artificial intelligence, algorithms and/or machine learning, for this survey, we opted to examine trust calibrations for *automated* systems/automation with automation referring to any system that can operate autonomously. However, we also acknowledge that terminological ambiguity exists in previous work (e.g., [131]) and that choosing a particular terminology affects reader's expectations and perceptions of systems [72].

Going beyond questions of functionalities, application, and terminology, research from the field of human-machine interaction is interested in how human users interact with and employ such automated systems. Two central observations have been made examining how users follow, interact, and rely on automated technologies: (1) users sometimes resist using automated systems, while (2) users sometimes also display an overreliance on automation. To understand and explain these observations, special importance has been given to the role of *trust* in automation. Originating in early works [74, 85, 106], trust has become a central variable to explain both resistance to use automated systems (disuse) as well as overreliance on automated systems (misuse).

Hence, many studies aim to adjust trust bestowed in a system to reflect the trustworthiness of systems, arriving at the concept of *calibrated trust* [74, 106]. For calibrated trust, the differentiation

of the perceived trustworthiness of a system and the actual trustworthiness of a system is crucial, as systems can be more or less trustworthy due to their functionality and reliability, while users might perceive these systems differently. In calibrated trust, the perceived trustworthiness of a system matches the actual trustworthiness of a system.

To achieve an appropriate calibration of trust, different approaches have been taken. Through empirical human-subject studies, researchers from multiple scientific fields have implemented trust calibration strategies, adopting various methodologies. However, with calibrated trust constituting a central variable to the appropriate adoption of automated systems, a coherent overview of this field is currently lacking. Thus, our aim with this study is to provide an overview of the current state of the field by focusing on empirical human-subject studies for the appropriate calibration of trust in automated systems. In doing so, we first provide a broader overview of the current understanding of calibrated trust in automated systems, discussing common challenges in achieving such a state. We then report the results of our survey of current empirical human-subject studies, which aimed to achieve a trust calibration. We summarize the system contexts and the task, the employed calibration interventions (experimental design), the measurement strategies of the calibration and evaluation metrics, and the results of these efforts. After each section, we reflect on the current state and point to future challenges. Based on the trust calibrations from all survey papers, we develop four trust calibration dimensions that reflect the different calibration strategies and their key advantages and shortcomings: (1) exo versus endo trust calibrations, (2) warranted versus unwarranted trust calibrations, (3) static versus adaptive trus t calibrations, and (4) capabilities versus process-oriented trust calibrations. We close our paper by summarizing our observations, identifying potential gaps, and providing recommendations for future work.

2 FROM TRUST IN AUTOMATION TO CALIBRATED TRUST IN AUTOMATION

For this work, we follow Lee and See's [74] definition of trust in automation, which is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p. 54) . In other words, trust is an attitude that is relevant in situations that include (1) levels of uncertainty, (2) a cooperative relationship between at least two entities, and (3) some exchange.

Hence, central to the definition of trust is the notion of risk/uncertainty (see the 2 above). Trust is not needed in situations where the outcome is certain or irrelevant because neither risk nor vulnerability are involved. Consequently, the notion of risk or uncertainty allows for the possibility of wrongfully trusting someone or something, which is usually accompanied by some sort of loss or pain, as well as wrongfully distrusting, which can be detrimental to performance. To avoid any such loss due to over- or under-trust, trust needs to be calibrated [74] or warranted [53], which describes "how well trust matches the true capabilities of the automation" [74, p. 57] (see also [28]). In cases of over-trust, a person's perception exceeds the system's true capabilities, leading to the system's misuse. For example, a person might rely too much on the capabilities of a navigation system, even in cases where it is clearly not appropriate. In the case of under-trust, a person's perception of a system's capabilities falls short of the true capabilities, leading to the system's disuse. For example, a person might wrongly disregard a warning from an autopilot.

It has to be mentioned at this point that, while most studies do not explicitly differentiate between trust and reliance, both concepts are, albeit often equated in everyday language, normatively distinct. Reliance is used for inanimate objects [9], for which, when reliance fails, we do not feel emotions of betrayal [121]. In contrast, to speak of trust requires the trustor to assign intent and anthropomorphism to the trustee which results in feelings of betrayal when trust is failed. Although all automated systems are inherently inanimate, users oftentimes anthropomorphize these systems [53, 97], hence, shifting from reliance to trust in the systems.

Theoretical and empirical work shows that whether one accurately calibrates one's trust in a system depends on various factors. Lee and Moray [73] proposed, for example, that the three factors performance, process, and purpose of the system determine trust, where performance describes the general reliability of a system, process describes how well the inner workings of a system are understood, and purpose describes the intentions with which the system was built. While all three factors, performance, process, and purpose, relate to the system itself, in a recent meta-analysis, Kaplan et al. [60] extended these system-related antecedents of trust in artificial intelligence by adding human-related antecedents of trust (e.g., individual's disposition to trust and ability, attitudes), and context-related antecedents of trust (e.g., risk of the situation).

Consequently, related to all three antecedents of trust, many possible sources of trust miscalibration exist. Individuals might be, by disposition, less/more inclined to trust a system, independent of its capabilities. Similarly, following insights on algorithm aversion [30] and algorithm appreciation [80], individuals' attitudes towards automation likely influence trust levels. From a system perspective, adding explanations or increasing system transparency might allow individuals to better gauge the true capabilities of a system. However, some results suggest that the increased transparency can also be detrimental to trust calibrations. Poursabzi-Sangdeh et al. [108] found that providing users with more information (through increased interpretability and transparency) decreased the user's capabilities to detect and correct mistakes by the system. The authors propose that this might be the result of information overload.

In sum, trust in automation is necessary for situations of uncertainty and dependence. Miscalibrating one's trust can lead to wrongfully accepting an automated system in cases of over-trust and wrongfully rejecting an automated system in cases of under-trust. Such miscalibration can be caused by various factors related to the human user, the system, and/or the context.

3 METHODOLOGY

To achieve a calibrated level of trust, different approaches have been taken, varying in their operationalizations of calibration, contexts, systems used, experimental set-ups, subjects, and results. In this section, we define the scope of our survey and describe the paper selection criteria and search strategies. While we did not intend to conduct a systematic review but a scoping review, to ensure,

nevertheless, a high quality standard for the search and selection process as well as the report of our results, we follow the Statement on Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020 Statement) [105]. A systematic overview of the steps we pursued to ensure quality can be found in the online supplementary material under the PRISMA 2020 checklist.

3.1 Scope and inclusion criteria

For this survey, we focused on studies that adhered to the following inclusion criteria:

(1) The paper must be peer-reviewed and published either within an academic journal or a conference proceeding. Preprints and theses were excluded.
(2) The paper must be empirical and include human subjects. We excluded studies that were solely based on simulation data or which solely focused on a system's perspective. Qualitative investigations were also excluded. This includes investigations which were empirical in nature, but data collected were qualitative. Papers which introduced a theory, a framework or guidelines without empirical data were also excluded.
(3) The paper must include a true trust calibration intervention. Thus, studies that merely intended to increase/decrease trust were excluded if they did not also include a matching operationalization of trust with the system's capabilities. This meant that it was necessary that three sub-criteria were fulfilled: first, trust was somehow assessed, i.e., through self-reports but also psychophysiological and behavioral measures, and, second, the system's ability had to be at least defined but ideally manipulated (by including, for example, different levels of reliability or stating a system's reliability). Only through this step the estimated level of system trustworthiness can be assessed which is needed in the third and final step. In a third step, to be included, studies had to match these two information with each other by assuming, for example, that high reliable systems were perceived more trustworthy than low reliability systems. This also meant that if a study, for example, only compared users' trust in a system after receiving an explanation versus after receiving no explanation, we excluded the study as there was no matching of the system's actual abilities with the perceived trustworthiness. Without information on the system's ability, the appropriateness of the potential trust increase due to an explanation cannot be judged.

We included the following keywords in our search to cover the central constructs:

- **Trust:** trust calibration OR reliability calibration OR trust adjustment OR reliability adjustment OR warranted trust OR unwarranted trust OR appropriate trust OR appropriate reliability OR trust repair OR overtrust OR undertrust.
- As **automated systems** can come under a myriad of systems or names, we extended this keyword search by: AND automat* OR autonomous OR algorithm* OR artificial intelligence OR machine learning OR robot* OR machine OR system OR agent OR computer.

3.2 Search Strategies and Selection Process

To select suitable papers, we proceeded in two search waves. During a first search, conducted on August 1, 2022 we searched for suitable papers using different platforms. We began our search on Google Scholar and extended the search to premier proceedings of human-computer interaction conferences such as the ACM CHI Conference on Human Factors in Computing Systems, ACM Conference on Computer-supported Cooperative Work and Social Computing, and ACM Conference on Fairness, Accountability, and Transparency. In a second wave, conducted on November 22, 2022, we extended our first search by including the full ACM library and the Scopus database. We finalized our search by examining and cross-referencing relevant citations from the papers we found in the first two steps. Hence, the time frame of our search ranged between the date of database inception and November 22, 2022. We only included English language publications.

This two-wave search resulted in over 1000 papers. A single coder removed duplicates and work that was not peer-reviewed (preprints & theses). Because the third inclusion criterion was more complex and required scanning the papers' abstracts and, if needed, the full texts, three coders reviewed 12% of the remaining papers to decide whether these were out-of-scope. The coders worked independently and reached a satisfactory inter-rater agreement of Krippendorff's alpha = .83 after a first round of coding. Disagreements were resolved via consensus discussions. Based on the discussion of the 12% of the paper corpus between the three independent reviewers, a single coder selected papers from the remaining body to be in- or excluded in the survey.

All articles selected for final inclusion were copied into a spreadsheet and coded by a single coder in reference to our primary categories of interest: the system, the task, the employed calibration intervention, the calibration measurement, trust measures, and calibration results. In addition to these categories, we also collected data to describe the datasets, such as study participants, number of participants, and set-up (online, lab, or field). In the second coding round, codes were grouped into related categories such as task domains, variables of interest, or calibration strategies. The spreadsheets containing (a) all papers found (including the coding of the three selection criteria), and (b) a list of all final papers is publicly accessible in the online supplementary material.

From the over 1000 papers found, 96 papers were included in this survey. For an overview of the selection process, see Figure 1. Studies were published during the years 1992 and 2023.

4 RESULTS

The central aim of our study is to provide an overview of the current state of the field by focusing on empirical human-subject studies for the appropriate calibration of trust in automated systems. To achieve this, in the next section, we summarize common strategies, advantages, and challenges. We start our survey by reviewing the studies' samples, and the tasks participants were asked to do. We further divided the task section into four subchapters—the role of automation, the degree of automation, the required expertise needed for the task, and the risk of the task—to reflect how the selected tasks influenced the results. This is followed by a section

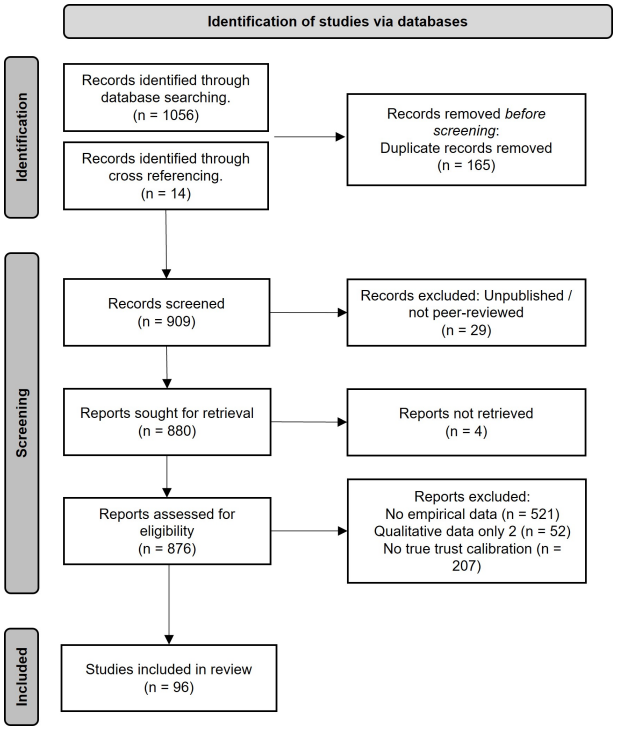

Figure 1: Flow chart of the paper selection process.

on the trust calibration interventions, and a final section which provides and overview of the results of these studies.

4.1 The Samples

Of the 96 articles included in the final selection, 40 were online studies, and 61 were conducted in a lab environment (some papers included multiple studies, hence the total number of studies exceeds the paper count of 96). The sample size of the online studies ranged from 9 to 1224 participants (M = 232, SD = 224). In contrast, the sample sizes of lab studies ranged from 8 to 865 [1] participants (M = 72, SD = 125). For the online studies, participants were all but one recruited via crowdsourcing panels such as Amazon Mechanical Turk or Prolific. One study used the researchers' social network. 29 of the lab studies explicitly recruited students, eight recruited experts, and three recruited staff members. The remaining study protocols did not disclose the recruitment process (some studies recruited a mix of participants, e.g., experts and non-experts).

In sum, the distribution of lab and online studies seems leveled, although skewed towards lab studies. Due to the increased effort of conducting field studies, not surprisingly, we did not find any in this set of studies. Hence, for future studies, we see great potential to test trust calibration interventions in the field.

4.2 The Tasks

We grouped all studies into different application domains to get an overview of the calibration tasks. We identified the following seven categories: security and safety, transportation, military, production, gaming, medicine, and others. Table 1 summarizes all seven domains, including the respective tasks. In the next section, we reflect on how these domain choices impact the results of the respective studies in terms of the role/function of automation, the required expertise needed for the task, and the risk of the task.

The role of automation. Overall, we identified two major roles of automated systems: cooperation and delegation. In turn, both cooperation and delegation were associated with different task clusters. We found that the cooperative role was mostly associated with automated decision aids which supported users in their decision-making process. We found this functionality in all seven task domains. For example, systems were used to aid users screen bags for dangerous objects [94], detect system malfunctions [89] or potholes [104], improve the production of orange juice pasteurization [73], play games [24], diagnose [100], discover number patterns [26], and classify plants and animals [110, 137].

The characteristic of these cooperative tasks is that users commonly perceive some sort of information or prompt from a system which the users then can decide to follow or reject. Hence, users retain a sense of autonomy. Drawing on the definition of trust, collaborative tasks should require less trust because the individual autonomy is greater and the dependency on another entity is less, reducing uncertainty and risk. However, need for trust in collaborative tasks might increase for situations of human-AI teaming, where the human team member is dependent on the performance of the system, such as [18, 24, 116] report. Such situations are characterized by increased dependency, possibly leading to an increased need for trust. In addition, when users become passive or complacent, the cooperative character might shift from cooperation to delegation. Passivity or complacency are usually reached in situations of over-trust. In turn, too much user engagement with the system might indicate that users trust the system too less, indicating the case of under-trust. Hence, too little or too much user engagement for cooperative tasks might be a good indicator to cue a trust calibration intervention.

In contrast to cooperation tasks, delegation tasks substitute human operators. Most predominantly, we found this functionality in the transportation domain for autonomous vehicles [2, 47, 61, 68–70, 84] but also in one study in which an automated player played the game Pong [50]. The characteristic of these delegative tasks is that the system operates primarily independent of the users. In addition, we found one particular case of automation substituting human operators in the case of a human-automation teaming task. In their paper, Johnson et al. [58] gave participants the task of gathering information (taking photos of ground target waypoints). The task was to be completed with another human and an autonomous teammate (see also [10]).

Unlike cooperative tasks, users are no longer in the position to choose whether they want to rely on or reject a system's decision/action but instead users are forced to accept or reject a system as a whole. When delegating, one could say that users must 'blindly trust' the system. In turn, this requires more trust as the dependency on another entity is higher; hence, uncertainty and risks increase.

Automation capabilities. Automation can refer to various different systems with varying capabilities, ranging from rather simple rule-based to sophisticated machine-learning algorithms. For most

[1] See [19]

Table 1: Task domains and the respective tasks of the studies included in this survey.

Domain	Task
Security & Safety	Screening of dangerous objects/subjects [4, 14, 15, 19, 36, 51, 54–56, 85, 94, 107, 117], detection of system malfunctions [89, 90], crime prevention [10], recidivism prediction [132], watch a video of a house search [66, 83]
Transportation	Responding to take-over requests [2, 6, 7, 47, 61, 68–70, 77, 84, 98], collusion avoidance [8], managing (air) traffic [31, 118], pedestrians interaction with AV [48], observe AVs [64, 119], drive in driving simulator [87, 96, 136]
Military	Screening tasks [17, 33, 44, 63, 81, 82, 127, 130, 135, 139, 140], gathering of information [58], mission planning [92], human-AI collaboration for search and destroy missions [116]
Production	Improving production [73, 133, 143], disassembly [5], moving objects [34, 35, 45], demand forecasting [39], harvesting [113], quality checks [142]
Gaming	trust game [3, 23], collaboration game [24], flanker task [25], image recognition [138, 144], finance game [62], first-person shooter game [67, 129], Pong [50]
Medicine	diagnostics [100], robot triage [102]
Others	Income prediction [40, 145], number pattern discovery [26, 27], pothole detection [104], classification of plants/animals/images [11, 75, 103, 110, 137], meal preparation [12], online dating [109, 141], hotel reviews [71], clearing tables [18], find housing [43, 124], navigation through buildings [101, 112], long-distance management of robots [126]

studies of this survey, it was difficult to gain insights into what powered the automated systems, as many authors merely described their systems as automated. While the development of many complex automated technologies is driven by artificial intelligence, we cannot make specific claims about these systems. For example, some papers examined robots (see e.g.,[3, 5, 6, 22, 24, 50, 130]). From previous work, we know that robots are increasingly equipped with artificial intelligence (see e.g., [111]).However, in most papers this information is missing. In contrast, most autonomous driving studies included information on the automation capabilities such as conditional automation levels (e.g., Level 3 - roughly 75% automation), except for one study, which included a highly automated system (Level 4) [61]. However, we found five studies that employed automation empowered by machine learning [71, 109, 110, 132, 141] and eight which specifically referred to artificial intelligence [11, 12, 43, 83, 100, 118, 144, 145]. Some studies also referred to fictitious automated systems [94] and others to a Wizard-of-Oz manipulations [58]. McGuirl and Sarter [90] used a cover story and told participants they were interaction with a system that is based on a neural net.

Required expertise. Following previous research, we differentiate between tasks that require task expertise and those that do not. This might overlap to some degree with the risk of the task (see next paragraph) as some high-risk domains (e.g., military, baggage screening, or diagnostics) inherently require task expertise. Others, such as driving autonomous cars or identifying animals and plants, do not require explicit expertise (other than a driver's license). However, we noticed that even though some tasks required task expertise (baggage screening), the study participants were non-experts (e.g., students). In fact, we noticed that only a few studies recruited experts to test their hypotheses [31, 66, 89, 90, 100, 113, 137, 140, 142].

We find this, albeit understandable due to limited resources and recruitment possibilities, problematic as experts react differently to systems than lay people (see, e.g., [59]). However, some studies compensated for the reduced expertise by introducing training sets before the intervention [47, 70, 73, 145]. Moreover, Doshi-Velez and Kim [32] argue that when the target group is challenging to attain, recruiting lay users allows for a "human-grounded evaluation" which can be understood as a proxy for the general behavior.

Risk of the task. To qualify for a task that requires trust, each task must involve a certain level of risk or uncertainty (see 2). Tasks that involve risks have also been described as high-stakes scenarios which can hold potential societal but also individual stakes. The task domains security and safety, military, medicine, and transportation inherently qualify as high-stakes scenarios. Tasks in the production domain can technically involve risks in terms of financial losses or physical risks at the workplace due to heavy machinery. The remaining tasks (such as income prediction, number pattern discovery, pothole detection, prescription screening, classification of plants and animals, playing a game, prize prediction) involve fewer risks and might be less suitable for triggering trust. However, some studies increased vulnerabilities by introducing gamification strategies in which participants were (financially) incentivized to perform well [10, 55]

Interestingly, we noticed that a large domain, medicine/healthcare, commonly characterized as a high-stakes domain, was represented by only two papers. As it is important to learn more about trust in high-stakes domains, this is unfortunate. Revisiting the papers throughout the selection process, we noticed, however, that various papers from the health domain did not meet our criteria of trust calibration in the sense of a matching of system capabilities and trustworthiness perceptions. Instead, many papers aimed to increase users' trust independent of the actual system capabilities. We speculate that this focus on trust increases might be the result of users generally under-trusting automated systems in the health domain. In contrast, we noticed that task domain automated driving has resulted in many efforts to assess and reduce unwarranted trust calibrations.

4.3 The trust calibration intervention

In this section, we review the trust calibration interventions. We divided this section into four subchapters. In the beginning, we briefly discuss the study design choices and how these likely affect the results of trust calibration interventions. Because any trust calibration consists of the matching of system capabilities and the perceived trustworthiness, we then review how the system capabilities were assessed/manipulated and how trust was measured. Lastly, for the assessment of the trust calibration, these two information (the system abilities and users' trust) have to be matched. In the subsection 'Measurement of calibration effects', we summarize and discuss how the different papers accomplish this matching task.

Study design choices. The selected papers differed in terms of the experimental design choices. In between-subjects design experiments, participants are randomly assigned to only one treatment condition. In contrast, in within-subjects (or repeated measures) designs, participants are exposed to all treatment conditions, whereas mixed-design experiments employ both strategies. All design choices hold their specific advantages and disadvantages. Between-subjects studies are usually shorter, reducing participants' workload, and study or learning effects can be avoided. In contrast, within-subjects designs raise awareness for variables of interest but also allow to control for the variability between participants. In the context of trust, this point might be of particular importance as previous studies have shown that individuals differ in their propensity to trust and their attitudes towards automation, both factors which affect trust [30, 41]. Despite this possible advantage of within- over between-subjects designs, most papers (49) reported results of between-subjects designs, followed by 25 papers using within-subjects designs, and 22 papers with mixed-design (between- and within-subjects manipulations). Many of the mixed-study designs aimed to test different calibration strategies over time - hence, introducing a within (repeated) subjects component. This also means that, if researchers are interested in the development of trust over time, researchers have to apply within-subjects designs.

System capabilities. Only few studies kept the system capabilities fixed (see, e.g., [71]), or measured a system's capabilities. Most studies *manipulated* the system's capabilities. The most common strategy to achieve this was to vary the levels of system reliability (see, e.g., [4, 10, 14, 15, 20, 26, 27, 55, 69, 81, 82, 89, 94, 104, 107, 117, 133, 137, 139, 140, 143]). For example, Chen et al. [15] presented participants with either 60%, 70%, 80%, or 90% reliable systems. Likewise, but choosing a within-subjects experimental design, de Visser et al. [27] let participants interact with a system that changed its reliability from 100% to 67%, 50%, and, finally, 0%.

In a similar manner, some studies manipulated changes in the system's reliability to represent either increases, decreases or a constant system reliability. For example, Pop et al. (2015) divided participants into three reliability groups with one group of participants experiencing a decrease in reliability (100%-80%), a second group experiencing an increase in reliability (60%-80%), and a third group experiencing no change (80%-80%). Some studies also investigated the effects of stated reliability and observable reliability [141].

While manipulating the system's reliability by controlling the exact reliability, some authors varied the system's capabilities by contextual factors such as the weather. For example, Helldin et al. [47] let participants drive an autonomous vehicle in a driving simulator but manipulated the weather conditions (sight was affected by snow) or the terrain of the course (steep climbs & tight turns), which, in turn, affected the system's reliability.

Other strategies included introducing system failures or malfunctions [68, 84], variations in the system's confidence [89, 90, 145], the system's accuracy [19, 94], the system's ability [132], and the system's credibility [20]. One subclass of system failures was the manipulation of error types by, for example, Huang et al. [50], who showed participants either a correct action, false-negatives, weak false-positives, false positives, or an incorrect action.

Trust measures. Throughout the papers included in this survey, two major strategies of assessing trust could be observed: (1) perceived or subjective trust measures, and (2) demonstrated or objective trust measures. Encompassing subjective trust measures, one common way to assess whether users perceived a system as trustworthy were validated self-report scales. We found 16 studies (see, e.g., [26, 68, 69, 84, 107]) which used the Trust in Automation Scale by Jian et al. [57], at least two [10, 94] studies used a scale developed by Merritt et al. [93], and at least five studies (see e.g., [4, 130]) used the trust predisposition scale by McKnight et al. [91]. We say "at least" as some studies also used very short ad-hoc 1-item self-report measures of trust (see, e.g.[15, 143]) which might be inspired by some of these validated scales. Moreover, Seong et al. [117] used the trust questionnaire by Llinas et al. [79], and de Visser et al. [27] used a combination of the scales by de Vries et al. [29], Lee & Moray [73], and Lewandowsky et al. [76]. Likewise, some studies [22, 68, 98, 100], used the Human-Computer Trust Questionnaire by Madsen and Gregor [86], whereas others [3, 56, 83] used a scale developed by Mayer and colleagues [95]. Other scales that were used are Holthausen's et al. [49] Situational Trust Scale for Automated Driving, a scale used in Muir [99], the Usability and Trust Questionnaire developed by Chen [16], the Robot Trust Questionnaire by Schäfer [114], a scale to assess trust in human-robot collaborations [13], and Körber's trust in automation scale [65]. In addition, some studies also employed linear trust meters [137], or visual analog scales [139].

While many of these scales differentiate between the different trust subconstructs performance, process, and purpose, only few authors differentiate between these subconstructs in their analysis. For example, Fahim and colleagues [36] examined the effects of different emotions such as hostility on each trust subconstruct separately. Similarly, Esterwood and Robert [35] investigated the effects of different trust repair strategies on perceived performance, process, and purpose. Results of both papers indicate that interventions affected the trust subconstructs differently. Hence, the missing differentiation between them might be critical.

In contrast to such subjective measures, many authors [89, 90, 104, 145] opted to only measure participants' behavior, indicated, for example, as responses to or compliance with the system and task performance [104, 145]. Many argued that such behavioral measures displayed a more objective trust measure than subjective self-reports (see, e.g., [14]). In addition to compliance and task performance, other performance measures were, for example, error rates and response times [81]. Most often, we found, however, that researchers used a combination of both self-report measures and

behavioral measures (see, e.g.,[4, 14, 19, 26, 47, 50, 55, 61, 84, 94, 117, 130, 139, 140]). In two studies, researchers also added qualitative questionnaires to advance their quantitative trust assessments [84, 90]. In rare cases, psychophysiological measures were implemented to assess trust, such as EEG [20, 25], EMG [58], electrocardiograms [58], and eye-tracking [70, 81].

In many cases, trust measures were accompanied by subconstructs of trust such as perceptions of system accuracy [90, 143], predictability [68, 73], and reliability [15, 68, 107, 133], or trust related measures such as perceived usability [4], frustration [4], perceived humanness [26, 27, 55, 58], perceived intelligence [27], workload [58], and understanding of the system [132].

In sum, we see a wide range of different methodologies. Especially, the increased use of validated scales ensures the quality of the measurements. In contrast, the use of ad-hoc 1 item measures makes these measures less reliable. Yet, their use was often justified by increased usability and lowered intrusiveness (see e.g., [104]). In these cases, concerns regarding the measurement quality were met by combining self-reports with behavioral measures or psychophysiological assessments. However, we also noticed that only a few studies differentiated between different subconstructs of trust (performance, process, purpose) which is inherently included in many scales (e.g., [91]) but authors did not make use of this information.

Measurement of calibration effects. Most commonly, the trust calibration was measured as the statistical differences of trustworthiness perceptions of systems with high reliability and trustworthiness perceptions of systems with low-reliability groups, examined through analysis of variance (see, e.g.,[4, 15, 26, 47, 58, 61, 68–70, 82, 84, 89, 90, 107, 117, 130, 140, 145]). The result is some sort of ordinal matching, which assumes that we can speak of calibration when systems with higher capabilities elicit higher trustworthiness perceptions than systems with lower capabilities.

While this operationalization is easy to implement and will allow us to compare the perceived trustworthiness of different systems, we think it comes with a significant limitation: It is a relative measure. This method allows us to infer that users rightfully differentiate between different systems, perceiving one as more trustworthy than the other. However, we do not know whether, for example, the system with lower capabilities is perceived as too (little) trustworthy, inducing over-trust (under-trust). The perceived trustworthiness might be reduced in comparison with a high capability system, but we cannot out rule that users still overestimate (underestimate) its trustworthiness. Likewise, a high capability system might elicit too much (less) perceived trustworthiness, inducing over-trust (under-trust).

Some authors work with correlations to get a better sense of perceived trustworthiness and system capabilities [5, 12, 24, 39, 66, 67, 81, 94, 130, 139, 142]. Here the larger the correlation between perceived trustworthiness and system capabilities, the better the calibration. For example, Merritt et al. [94] operationalized the trust calibration through trust sensitivity, which "reflects the extent to which a user's trust changes as the automation's actual reliability level changes." (p. 36). The authors achieve this through a repeated-measures experimental design for which they varied the actual reliabilities. While this operationalization is closer to the actual definition of trust calibration, Merritt et al. [94] also discuss the possible limitation of this procedure: "Individuals with a great deal of trust sensitivity (meaning that their trust changes dramatically as actual reliability changes) might actually either overreact or underreact to the changes, thereby resulting in suboptimal automation reliance decisions." (p. 36)

A third way to measure the trust calibration was to measure over-, under-, and appropriate trust as the behavioral deviation (reliance) from the ideal behavior [27, 50, 55, 104, 110, 132, 137]. For example, Jensen et al. [55] asked participants to identify dangerous vehicles in a set of images. Participants could manually check the images or let an automated aid do it for them. Participants were informed that they would be credited for speed and accuracy. Depending on the reliability condition, the authors could then define the ideal reliance on the aid in a way that delegating 18 images to the aid would identify as over-trust but delegating only 10 images would identify as under-trust. However, it must be noted that behavior can but does not have to be the result of users' trust. It is, for example, possible that users simply wanted to reduce their own workload and, hence, relied more on the automation. Overall, we found 18 papers which included such a behavioral reliance measure.

4.4 The calibrations dimensions

Because we identified many different trust calibration interventions, we extracted their commonalities and differences with the aim to allow us to summarize and abstract these various strategies. As a final result, we abstracted four different dimensions of trust calibration: (1) exo versus endo trust calibration, (2) warranted versus unwarranted trust calibration, (3) static versus adaptive trust calibration, and (4) capabilities versus process-oriented trust calibrations.

(1) Exo versus endo trust calibrations. With the dimensions'exo' and 'endo', we refer to the point in time when trust calibrations occur. While 'exo' refers to a point in time outside the interaction with a system (hence, before or after the interaction), 'endo' refers to an interventions that occurs while participants interact with a system.

Exo calibrations aim to align the users' trustworthiness perception of the system with the actual trustworthiness as early as possible in the interaction (initial) or try to achieve alignment after the interaction. One way to achieve this is, for example, to inform users about a system's capabilities and limitations prior to interacting with the system. To that end, Khastgir et al. [61] found that increasing knowledge about a system's capabilities and limitations increased trust in high and low capability-autonomous vehicles and reduced the number of accidents for low capability-autonomous vehicles. Similar results were obtained by Kraus et al. [69], who provided information about the system's reliability and reputation. Moreover, Kraus and colleagues [69] found that providing prior information affected users differently depending on their need for cognition. To achieve an alignment after an interaction, some studies provide feedback about the user and system performance (see e.g., [5, 14, 82]).

Like exo trust calibrations, endo calibrations offer insights into a system's capabilities and limitations. In contrast to exo trust calibration strategies, endo trust calibrations allow users to gain these insights *while* engaging with a system. This is achieved by, for example, presenting information about a system's confidence throughout

the interaction [90, 104, 145] using cues or warning signals or by increasing a system's transparency through explanations [108].

The papers of this review offer a wide range of different calibration strategies, aiming to calibrate trustworthiness perceptions at different stages (exo = prior & after; endo = during) of the interaction with systems. An overview of both exo and endo calibrations can be found in Table 2.

(2) Warranted versus unwarranted trust calibration. In its actual sense, warranted trust refers to the accurate calibration of trust, reflecting a system's trustworthiness [53, 74]. Hence, providing knowledge about a system's capabilities and limitations (as suggested in the previous section) should increase warranted trust. However, other factors have also been identified to increase the trust users bestow in a system, such as a system's reputation [69] or anthropomorphism [26, 55]. Yet, while a system's reputation or a human-like appearance might serve as a cue/heuristic to assess whether to trust or not to trust a system, neither of these factors genuinely affect the system's reliability. Hence, such factors induce unwarranted trust and can potentially be misused to make unreliable systems appear more reliable.

(3) Static versus adaptive trust calibration. Static versus adaptive trust calibration refers to a system's capability to assess whether the user is currently under- or over-relying on the system. Most exo (i.e.,initial and later) trust calibrations surveyed in this paper, for example, are inherently static. They inform all users equally about a system's capabilities and limitations at the beginning or the end of the interaction and do not offer insights while users are using a system. While some users, for example lay users, might initially need more information, others, for example expert users, likely do not require the same information. Similarly, after using a system multiple times, static trust calibrations likely become redundant for users.

However, an exo trust calibration does not have to be static. For example, if a system is provided with information about its user (e.g., level of experience as a heuristic to likely over- or under-reliance) and the system can adapt to this information, trust calibration is adaptive. Moreover, during the interaction, endo calibrations which increase a system's transparency through explanations or by providing model internals would only qualify as adaptive trust calibration when this information is given in cases when users display under- or overreliance.

Two examples of adaptive trust calibrations comes from Okamura and Yamada [104] as well as Chen et al. [18]. Okamura and Yamada [104] implemented specific trust calibration cues, which were displayed when participants relied too little and too much on the system. For example, cues were shown whenever a user did not rely on the system, although the system's current prediction accuracy was high. In turn, cues were shown whenever a user relied on the system, although the system's current prediction accuracy was low. Okamura and Yamada [104] could show that participants who received adaptive trust calibration cues outperformed those who continuously received information about the system's performance.

Similary, Chen and colleagues [18] developed a computational model, the Partially observable Markov decision process (POMDP), which integrates the extend users trust the system into the system's decision making by, among others, inferring users' trust through interactions. Through an experiment Chen et al. could show that the system adopted to users' level of trust which increased the overall human-robot team performance.

Importantly, adaptive trust calibration strategies signify a change of adjustments. While most calibration strategies are directed to change the users' attitudes and behavior (i.e., adjusting the user to the system), in adaptive strategies the system also adjusts to the users.

(4) Performance versus process-oriented trust calibrations. With these dimensions, we follow the in the introduction presented trust model by Lee and See [74] which includes the dimensions performance, process, and purpose. Performance trust calibrations inform users about the specific capabilities and limitations of a system, such as information about a system's confidence [90, 104, 145] or general information about a system's reliability [61, 69]. Such performance-oriented trust calibrations do not inform users how the system arrived at a decision.

In contrast, process-oriented trust calibrations such as explanations clarify the inner workings of a system. The task of translating information about the process into information about the performance of a system lies with the users. In turn, users vary, among other factors, in their literacy and experience which makes it likely that not all users will profit equally from such process-oriented calibrations.

4.5 Results of the trust calibration interventions

The following paragraphs provide an overview of the trust calibration results. In doing so, it is not our intention to provide an extensive insight into all 96 studies and their theoretical implications and limitations. Instead, we intend to provide a rather general overview of common trends and possible conflicting results.

Unsupported trust calibrations. Many of the studies in this survey included control conditions, allowing us to make assumptions about user's unsupported, natural trust calibrations. Results of these efforts support the notion that users can and do differentiate between low and highly reliable systems (see e.g.,[19, 20, 69, 130]), with more reliable systems being perceived as more trustworthy than less reliable systems. Yet, some studies also found that low-reliability systems tended to be perceived as more trustworthy than they were, leading to over-trust, whereas high-reliability systems tended to be perceived as less trustworthy than they were, leading to under-trust (see, e.g., [55]). Various studies also reported that trust gradually developed over time and that special attention should be paid to the early interaction processes, which tend to be decisive for trust development, and, hence, trust calibration [143].

While the users' perceptions might change and develop, reliability changes in the system also affected users' trustworthiness perceptions. One study found, for example, that users reacted rather slowly to changes in reliability and that such changes depended on the direction of reliability change [133]. Decreases in reliability from perfect to 80% were perceived as less reliable than a constant reliability of 80%, whereas increases in reliability did not benefit as much, and reliability perceptions remained lower than the increase. In addition, Lu and Sarter [82] found that participants trusted systems more following the recovery from small and short reliability changes than large and short reliability drops. If system failures or malfunctions due to lower reliability occur, two studies showed

Table 2: Overview of the different calibration strategies sorted by time of intervention.

Point of intervention	Intervention
Prior to the interaction	Prior information about the system such as capabilities, reliability, consistency, consensus, distinctiveness of the automated aid's performance, or malfunctions; training phase; cognitive forcing
During the interaction	Cues (alarms, warning signals, augmented reality cues, confidence cues or updates); uncertainty displays; communication style (command vs. status); explanations; interactivity (e.g., promise to perform better in the future)
After the interaction	Feedback type (automation feedback independent of user, feedback on trust behavior of user); performance feedback

that the error type also affected trust calibration differently. For example, Chen et al. [14] found that false alarms reduced perceived trustworthiness more than misses.

Similarly, Lee and Moray [73] found that after transient malfunctions, performance and trust dropped but recovered quickly. In contrast, chronic malfunctions only decreased trust but not performance. The authors argue that users can adapt to chronic low reliability and compensate with their skills at the cost of increased workload. In turn, Guznov et al. [44] found that increased workload led to more reliance on a system. To assess trust, one paper found that trust is best reflected at each moment in the interaction rather than at the end of the interaction [139]. This might be explained by results from Lee and Moray, who found that immediately preceding events affected trust the most.

Moreover, two papers [48, 127] examined whether malfunctions of subsystems affected the perceived reliability of the whole system. Holĺander and colleagues [48] found that users generally trusted a subsystem more than the system as a whole, and if a subsystem failed, this malfunction reduced trust in the whole system. These results are supported by Walliser et al. [127] who found that a single inaccurate system lead to more verifications and lower subjective trust for similar but independent systems.

To conclude, the reviewed studies indicate that users are generally sensitive to a system's reliability as well as to changes in reliability. The results also suggest that this sensitivity is oftentimes not enough and comes at the cost of over- and undertrusting as well as other coping mechanisms such as increasing one's attention and workload to surveille the system. Because of these limitations, the results also confirm that trust calibrations should be supported to assist the users.

Supported trust calibrations. While many studies included control conditions to make assumptions about trust calibrations in the wild, the main focus of most studies was, however, to support users through different interventions to appropriately calibrate their trust in a system. The aim of many these interventions was to make systems more transparent. Results of these efforts were mixed. For example, it was found that uncertainty information [47], the display of confidence levels [145], and reliability updates [90] facilitated an appropriate trust calibration and performance. Yet, uncertainty cues also increased workload in some cases [70] but not in others [92] and led to instances of over-trust when the cues were not reliable [140]. Ribeiro et al. [110] found that displaying class probabilities did not help users to appropriately assess a system's trustworthiness (see also [6]). However, Rechkemmer and colleauges [109] investigated the effects of false transparency interventions by providing users with a wrong reliabiltiy information. The authors found that users quickly disregarded the wrong information and adapted to the observable reliability.

Timing of the support. The effectiveness of interventions also varied depending on how and when they were shown. Yang and colleagues [139] found that likelihood alarms took more time to adjust trust calibration than binary warning signals. Also, updated system reliability cues improved trust calibration better than a static display of system reliability [90]. Similarly, adapting the cue display to users' trust behavior was more beneficial than a static display of the system's reliability [104]. Informing users about the reliability of a system, for example, helped to appropriately adjust trustworthiness perceptions [61] which the authors coined as "informed safety". These results are in line with findings by Johnson et al. [58] who found that one way to ensure a calibrated level of perceived trustworthiness was to include a training phase. Yet, results by Leichtmann et al. [75] indicated that merely educating users about how a system functions was not as helpful as including explanations. Moreover, providing cognitive feedback information[117] or forcing users to critically assess the trustworthiness of a system, [12] improved the trust calibration. Yet, Alhaji et al. [5] could show that system feedback was redundant when the system performed well.

After the interaction, some studies provided feedback to the users about the users' performance or the systems' performance. Again, the results were mixed. Lu and Sarter [82] found that neither performance feedback, users' or system's performance, affected the trust calibration. Yet, others found positive effects of feedback [14, 15, 94].

Explanations. Another way to adjust the perceived trustworthiness was through explanations. The results of these studies were mixed. For example, Zhang et al. [145] found that local explanations did not affect the trust calibration. In contrast, Yang et al. [137] found that visual explanations supported an appropriate trust calibration. Testing different explanations, Wang and colleagues [132] found that feature importance and feature contribution explanations supported the users, whereas the effects of nearest neighbors and counterfactual explanations were inconclusive. Naiseh and colleagues [100] found that example-based and counterfactual explanations were more understandable to their participants than local, global, and no explanation conditions. In addition, the effects of explanations might also be dependent on the systems' reliability. Wang et al. [130] found, for example, that explanations were only beneficial for low reliable systems and that very reliable systems did not benefit from the explanation. Moreover, Lai and Tan [71]

found that increased task difficulty reduced the positive effects of explanations.

Special case: Trust repair strategies. A whole section of papers from this survey had the designated aim to repair trust after a system malfunction. One common repair strategy was to include some sort of system reaction into the interaction with the users such as an apology or a promise. To that end, Kohn and colleagues [64] found that apologies worked better to regain trust than a system denying an error. However, Esterwood and Robert [34] could show that promising no errors in the future as well as explanations for why the error happened showed greater effects than mere apologies (see also results by [136]). In contrast to that, Robinette et al. [112] and Schelble et al. [116] found no effects of apologies on trust repair.

Examining possible moderating variables, Kox et al. [67] found that trust repair strategies profited the most when accompanied by an expression of regret, while including different levels of uncertainty had no effect on the trust repair [67]. In a similar vain, Kim and Seong [62] examined the effects of anthropomorphism and error attribution. The authors found that machine-like agents profited more from external attributions of blames than internal attributions, whereas it was the opposite case for human-like agents which profited the most when acknowledging a mistake (i.e., internal attributions).

In addition, some scholars also examined whether trust repair strategies affect different trust subconstructs differently. For example, Esterwood and Robert [35] found that perceptions of a system's abilities were generally unaffected by any repair strategy. In contrast, repair strategies increased the perceived benevolence and integrity of a system.

Calibration Moderators. Some studies also examined moderating variables which affect the trust calibration. Usually, these moderating variables are not directly related to the system but describe variations between different people or contexts. For example, Kraus et al. [69] examined how the system reputation, individuals' need for cognition disposition, and personality affected trust calibration. The authors found that a lower perceived reputation also decreased individuals' trustworthiness perceptions, and the individuals with a high need for cognition relied more on reliability information in their trust calibration. In terms of personality, Kraus et al. [69] found that higher levels of materialism, as well as a regulatory focus, were associated with higher levels of trust. Pop and colleagues [107] found that individuals who held high expectations towards systems were more sensitive to changes in reliability and calibrated their trust levels more appropriately for reliability increases but not decreases as compared to individuals who had lower expectations. In addition, individuals' level of experience affected the trustworthiness perceptions [103], and one study even found that experts were generally less willing to rely on automation than lay users [113].

In addition to these factors, which relate to the individual user, others found differences depending on the system's appearance. For example, Jensen et al. [55] found that the more human-like a system was, the more it was perceived as benevolent, which did not translate into differences in behavior trust. This is in line with findings by de Visser et al. [27], who found that more human-like systems were also perceived as more trustworthy – a finding we interpret as the elicitation of unwarranted trust (see 4.4). Contrasting these results, Christoforakos et al. [22] found no effects of anthropomorphism.

To achieve human-likeness strategies also varied. De Visser and colleagues [27], for example, compared a human agent with an avatar and a computer. In turn, Jensen et al. [55] compared variations of the communication style and Gupta et al. [43] varied the interfaces between conversational interfaces and web-based graphical interfaces. In addition to the anthropomorphism of a system, the perceived expertise of a system had no effects on trust in cases when the system was inaccurate [25]. Yet, Madhavan and Wickens [85] found that the perceived expertise was also inconclusive if the system operated well.

Focusing more on the context than on the user or the system, Chen et al. [19] found that trustworthiness perceptions also depended on the severity of the outcome, with more severe outcomes reducing the perceived trustworthiness. This finding aligns well with the definition of trust as a process in which one party becomes vulnerable to another party. The greater the vulnerability of one party, the more trust is required. In relation to that, the authors also found that trust calibration might vary between different domains (here: pharmacy versus online banking). However, the results are challenged by findings from Yin et al. [141] who found no effects of stakes in trust.

Mediators of trust calibrations. One rather small portion of papers examined the psychological processes of the trust calibration. For example, Fahim et al. [36] suggested that emotional reactions towards a system can partly explain the resulting trust attitude. Interestingly, their results showed that hostility mediated the relationship of a system's reliability and the perceived system integrity but did not so for the perceived abilities and benevolence of a system. The results are not only one of the few which highlight the importance of affective processes but also which make it evident to differentiate between different subconstructs of trust.

In sum, we found that results were mixed. Generally, users were perceptive of a system's capabilities and trustworthiness, and the perceived trustworthiness developed over time. However, users did not react appropriately to changes in reliabilities. Decreases in reliability resulted in even steeper trust decay, whereas increases did not result in a similar trust increment. As it is likely that system's capabilities will vary depending on the context, these results are critical for trust calibration interventions. Future studies should find ways to buffer such decreases in reliability and boost increases. One way to achieve this might be an explicit training phase.

To facilitate the trust calibration, one broader strategy was to implemented interventions to increase transparency (e.g., cues, warning signals, explanations). However, increases in transparency were no panacea for appropriate trust calibrations. System uncertainty information, confidence levels, and reliability updates were helpful, but also increased users' workload. Especially, the results for explanations were mixed with findings indicating advantages but also null findings.

5 SUMMARY AND TAKEAWAYS

In this survey, based on a literature body of 1000+ papers, we reviewed 96 empirical studies which aimed to calibrate users' trust in automated systems. In doing so, we provide an extensive overview

of the study design choices (samples, tasks, systems, measures, interventions), the measurement of the trust calibration, and an overview of the studies' results. For each section, we reflect on how these choices affected the results and elaborate on how to overcome possible limitations. The critical points of these reflections are summarized below.

5.1 Current trends

- **Moving from trust to calibrated trust.** Overall, we found that many studies have recognized that merely increasing users' trust is not sufficient but that we need to arrive at calibrated levels of users' trust.
- **Diverse domains.** Studies in this survey come from various domains such as security and safety, transportation, military, production, gaming, medicine, and others. However, we noticed that only very few of the surveyed studies came from the medical/health domain (n=2). With the increased application of automated systems powered by artificial intelligence, we see great potential for future studies to fill this gap.
- **Fewer experts and less risk.** As a result of the often difficult recruiting process, we found that many papers chose systems that require expert knowledge but did not recruit experts. Similarly, we found that the individual risk of users, a prerequisite for trust, was relatively low. However, for both problems, we also found adequate solutions. The lack of expertise was compensated by including training sessions, and risk was increased by introducing (financial) incentives for task performance.
- **Moving forward to complex automation.** The automated system was not well defined for most of the studies surveyed in this work. While automation was previously predominantly used as a tool to support human workflow, recent algorithm- and artificial intelligence-based systems increasing assume agency. Such machine agency competes with human agency and is likely perceived differently from less agentic systems [120]. Moreover, such a sense of agency is critical in the context of trust as increases in agency imply increases in vulnerability. For static automated systems, the outcome becomes more predictable than outcomes of systems that are powered by, for example, deep learning algorithms.
- **Measuring trust.** We saw many ways to measure trust, such as self-reports, behavioral and psychophysiological measures, such as EEG, EMG, or eye-tracking. This wide spectrum offers researchers various methods to choose from, depending on what is most appropriate for their task/system. In situations where long self-report scales are less feasible (e.g., situations that require a lot of workload), researchers can rely on short one-item measures, behavioral indicators, and psychophysiology.
- **Understanding or trusting.** Many calibration strategies increased the transparency of the systems either before the interaction with the system through prior information about the systems' capabilities or during the interaction through reliability cues, confidence updates, explanations, and alarms. This trend points to an interesting interaction of trust and understanding. As proposed earlier, if all a system's doing were understandable and predictable, trust would not be needed due to lacking vulnerability and risk. However, some understanding seems to be necessary to promote trust—the question is, then, how much understanding is necessary? As McAllister [88] suggested, "the amount of knowledge necessary for trust is somewhat in between total knowledge and total ignorance" (p. 26). We think this question poses great potential for the human-machine interaction community.

5.2 Challenges

- **What makes a good calibration measurement?** We identified three different strategies to measure the trust calibration: a relative measure, a correlational measure, and a behavioral measure. The relative measure does not quantify the calibration per se. Instead, it allows us to compare different groups, for example, a group that received an intervention, with a group that did not. This is problematic as it does not inform about the actual matching of a system's capabilities with its perceived trustworthiness and, consequently, does not diagnose under- or over-trust. A better way to connect a system's capabilities with its perceived trustworthiness is through correlational measures, as it quantifies the actual relation of both measures. The definition of under- or over-trust might vary, however, depending on the selected measures, and should be defined prior to the data collection through, for example, pre-tests. One straightforward way to assess the trust calibration was to measure behavior and conclude whether users relied too much or too less on the system. However, this is only feasible for cooperative systems where users can choose to reject or accept a system's behavior/decision. For delegative systems, it becomes more challenging to assess under- or over-trust. At the most, one could observe users' take-over behavior.
  To conclude, it remains complicated to operationalize the trust calibration as trust operates on a different latent scale than the also latent system's capability measure, making it impossible to compare both directly. Hence, we advise future researchers to carefully consider the operationalization and avoid using especially relative measures. It should be ideally guided by the question of identifying under- and over-trust.
- **Interventions as a second level of trust calibration.** Especially the results of one study [140] showed that adding a trust calibration intervention (here by including uncertainty cues) resulted in participants trusting the cues too much. This case exemplifies well that by adding more information, we also add another layer of trust. Similar cases have been found for explanations. Results by Poursabzi-Sangdeh et al. [108] suggest that providing users with more information (through increased interpretability and transparency) decreased the user's capabilities to detect and correct mistakes by the system. Just as the system can lead to over-trust, introducing calibration measures can also backfire. To avoid such miscalibrations, any intervention needs to go through thorough testing. Moreover, grounded in dual-processing

theories of cognition, adding additional information might just serve as a heuristic cue, leading to quick reliance merely on the basis simply *any* additional information being present.

- **Is it ok to induce unwarranted trust?** We found that some studies examined trust calibration strategies that did not relate to the system's capabilities, for example, the system's reputation [69] or anthropomorphism [26, 55]. However, according to the definition of trust calibration, we can only speak of calibrated trust when the users' perceived trustworthiness reflects the system's actual trustworthiness. Hence, the effects of reputation or anthropomorphism induce, strictly speaking, unwarranted trust. Is this justified? While we cannot provide a normative answer to this question, we also want to point out that this is not a matter of choice for some systems. Wherever users interact with an embodied agent or a robot, anthropomorphism will come into play. Likewise, (brand) reputation cannot be excluded from a system. By pointing toward these issues, we hope to foster future debates.
- **From static to adaptive calibration strategies.** In the third dimension of trust calibrations (see 4.4), we differentiate between static versus adaptive calibrations. While static calibrations (e.g., providing prior information before the interaction or providing system updates during the interaction) do not adapt to the individual user, adaptive calibrations can be individualized to assess users' momentary under- or overtrust. Static interventions like providing prior information about a system's capabilities expect the user to adapt to the system. Miscalibrations are then the result of users failing to calibrate. In contrast, for adaptive calibrations, the system adapts to the users' needs and miscalibrations become the result of the system failing to calibrate to the user.
  We observed that a majority of interventions employed a static calibration strategy which treats all users equally, expecting that users adapt to the system. We see great potential for system developers to change this imbalance and to strive for adaptive calibration strategies which shift the burden of calibrations away from the users and more towards the system itself. Two great examples of how a system can adapt to the users' level of trust come from Okamura and Yamada [104]mand Chen et al. [18].

5.3 Limitations

As with every work, this survey comes with specific theoretical and methodological limitations. First, from a theoretical perspective, we follow Lee and See's [74] definition of calibrated trust as "trust [that] matches the true capabilities of the automation" [74, p. 57]. However, limiting the trust calibration process to matching only the capabilities of systems to trustworthiness perceptions arguably limits one's definition of trust. As we delineated in section 2 of this paper, a system's capabilities (i.e., it's performance) is only one of three factors which contributes to trust. Besides performance, a system's process (the inner workings of a system) and purpose (the intentions with which a system is build) shape the trust process. For example, Tolmeijer et al. [123] found that ethical decision making system were perceived as more capable (catering to our definition of trust) but also less moral (catering to the perceived purpose of a system). Similarly,Textor and colleagues [122] presented with two similarly capable system which differed in their ethical behavior. To conclude, we have limited our survey to investigations which cater to only of three possible factors in the trust calibration process.

Second, in the calibration dimensions section (4.4), we have discussed warranted and unwarranted trust calibrations, using the example of anthropomorphism. We argued that a human-like appearance might serve as a heuristic to increase trust in a system without genuinely affecting the system's capabilities, leading to unwarranted trust. However, unlike Lee and See [74], Jacovi and colleagues [53] have argued that calibrated trust "is trust that is caused by trustworthiness (to some contract)" (p.632). By that definition, if the contract includes a human-like appearance of a system, anthropomorphism would lead to warranted trust as anthropomorphism has been included in the a priori system specifications.

From a methodological perspective, our results are limited to English language texts. Moreover, the largest part of the papers we found was screened by a single coder. While this might appear less thorough than a screening process which includes more coders, we think that the high inter-rater agreement (Krippendorff's alpha of .83) of the training phase justifies our procedure. The single coder should have been sufficiently trained to follow our selection criteria as closely as possible.

5.4 Recommendations for Practical implications

To advance future studies on trust calibrations, in this section we discussed in-depth the different decision stages of designing such interventions. We highlighted the four dimensions of trust calibration—initial versus dynamic trust calibrations, warranted versus unwarranted trust calibrations, static versus adaptive trust calibrations, and capabilities versus process-oriented trust calibrations—to assist in the planning and interpretation of future studies, and pointed to current trends and challenges of the trust calibration task. As a last and final step, we developed a list of recommendations as a hands-on guide for practitioners and academics alike:

(1) Do not just try to increase trust but try to **reach levels of calibrated/appropriate trust.** Operationalizing the trust calibration (i.e., matching the system capabilities with users' trustworthiness perceptions) is difficult. Try to think of ways how to identify that a user is currently over- or under-trusting.
(2) Be aware whether your task is characterized by **cooperation or delegation**. Do users work together with the system (e.g., as with most decision aids) or does the system substitute the users' action (e.g., as with Level 3 autonomous driving)? Be aware that delegation tasks likely require users to trust the system more than for cooperative tasks.
(3) If **task expertise** is needed, try to recruit experts when testing your system. If this is not possible include **training phases**.
(4) Use **validated scales** to measures the perceived trustworthiness. If possible include a variety of measures. Combine, for example, self-reports with behavioral measures (e.g., compliance with the system or reaction times).

(5) Know of the **different ways to calibrate** users' trustworthiness perceptions. Ask yourself when the calibration should/can happen: before, during or after the interaction? Is the calibration warranted? For example, anthropomorphizing might increase users' trust but does the resulting trust perception reflect the system's capabilities? Also, ask yourself if you can implement ways in which the system can adapt to the users. The four dimensions which we developed in the section 4.4 will help in your decision-making.

ACKNOWLEDGMENTS

This work has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the https://uaruhr.de.

REFERENCES

[1] Shreya Agrawal and Pooja Jain. 2017. An improved approach for movie recommendation system. In *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. IEEE, 336–342.

[2] Kumar Akash, Neera Jain, and Teruhisa Misu. 2020. Toward adaptive trust calibration for level 2 driving automation. In *Proceedings of the 2020 international conference on multimodal interaction*. 538–547.

[3] Gene M Alarcon, Anthony M Gibson, and Sarah A Jessup. 2020. Trust repair in performance, process, and purpose factors of human-robot ttust. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. IEEE, 1–6.

[4] Yusuf Albayram, Theodore Jensen, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, Ross Buck, and Emil Coman. 2020. Investigating the effects of (empty) promises on human-automation interaction and trust repair. In *Proceedings of the 8th International Conference on Human-Agent Interaction*. 6–14.

[5] Basel Alhaji, Michael Prilla, and Andreas Rausch. 2021. Trust Dynamics and Verbal Assurances in Human Robot Physical Collaboration. *Frontiers in Artificial Intelligence* (2021), 103.

[6] Kamilla Egedal Andersen, Simon Köslich, Bjarke Kristian Maigaard Kjær Pedersen, Bente Charlotte Weigelin, and Lars Christian Jensen. 2017. Do we blindly trust self-driving cars. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-robot Interaction*. 67–68.

[7] Jackie Ayoub, Lilit Avetisyan, Mustapha Makki, and Feng Zhou. 2021. An Investigation of Drivers' Dynamic Situational Trust in Conditionally Automated Driving. *IEEE Transactions on Human-Machine Systems* 52, 3 (2021), 501–511.

[8] Hebert Azevedo-Sa, Huajing Zhao, Connor Esterwood, X Jessie Yang, Dawn M Tilbury, and Lionel P Robert Jr. 2021. How internal and external risks affect the relationships between trust and driver behavior in automated driving systems. *Transportation research part C: emerging technologies* 123 (2021), 102973.

[9] Annette Baier. 1986. Trust and antitrust. *ethics* 96, 2 (1986), 231–260.

[10] Philip Bobko, Leanne Hirshfield, Lucca Eloy, Cara Spencer, Emily Doherty, Jack Driscoll, and Hannah Obolsky. 2022. Human-agent teaming and trust calibration: a theoretical framework, configurable testbed, empirical illustration, and implications for the development of adaptive systems. *Theoretical Issues in Ergonomics Science* (2022), 1–25.

[11] Michelle Brachman, Zahra Ashktorab, Michael Desmond, Evelyn Duesterwald, Casey Dugan, Narendra Nath Joshi, Qian Pan, and Aabhas Sharma. 2022. Reliance and Automation for Human-AI Collaborative Data Labeling Conflict Resolution. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–27.

[12] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[13] George Charalambous, Sarah Fletcher, and Philip Webb. 2016. The development of a scale to evaluate trust in industrial human-robot collaboration. *International Journal of Social Robotics* 8, 2 (2016), 193–209.

[14] Jing Chen, Scott Mishler, and Bin Hu. 2021. Automation error type and methods of communicating automation reliability affect trust and performance: An empirical study in the cyber domain. *IEEE Transactions on Human-Machine Systems* 51, 5 (2021), 463–473.

[15] Jing Chen, Scott Mishler, Bin Hu, Ninghui Li, and Robert W Proctor. 2018. The description-experience gap in the effect of warning reliability on user trust and performance in a phishing-detection context. *International Journal of Human-Computer Studies* 119 (2018), 35–47.

[16] Jessie YC Chen. 2009. Concurrent performance of military tasks and robotics tasks: Effects of automation unreliability and individual differences. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 181–188.

[17] Jessie YC Chen, Michael J Barnes, and Caitlin Kenny. 2011. Effects of unreliable automation and individual differences on supervisory control of multiple ground robots. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 371–378.

[18] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 307–315.

[19] Yan Chen, Fatemeh Mariam Zahedi, Ahmed Abbasi, and David Dobolyi. 2021. Trust calibration of automated security IT artifacts: A multi-domain study of phishing-website detection tools. *Information & Management* 58, 1 (2021), 103394.

[20] Sanghyun Choo and Chang S Nam. 2022. Detecting Human Trust Calibration in Automation: A Convolutional Neural Network Approach. *IEEE Transactions on Human-Machine Systems* (2022).

[21] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[22] Lara Christoforakos, Alessio Gallucci, Tinatini Surmava-Große, Daniel Ullrich, and Sarah Diefenbach. 2021. Can robots earn our trust the same way humans do? A systematic exploration of competence, warmth, and anthropomorphism as determinants of trust development in HRI. *Frontiers in Robotics and AI* 8 (2021), 640444.

[23] Michael G Collins and Ion Juvina. 2021. Trust Miscalibration Is Sometimes Necessary: An Empirical Study and a Computational Model. *Frontiers in Psychology* 12 (2021).

[24] Sylvain Daronnat, Leif Azzopardi, Martin Halvey, and Mateusz Dubiel. 2021. Inferring Trust From Users' Behaviours; Agents' Predictability Positively Affects Trust, Task Performance and Cognitive Load in Human-Agent Real-Time Collaboration. *Frontiers in Robotics and AI* 8 (2021), 194.

[25] Ewart J De Visser, Paul J Beatty, Justin R Estepp, Spencer Kohn, Abdulaziz Abubshait, John R Fedota, and Craig G McDonald. 2018. Learning from the slips of others: Neural correlates of trust in automated agents. *Frontiers in human neuroscience* 12 (2018), 309.

[26] Ewart J de Visser, Frank Krueger, Patrick McKnight, Steven Scheid, Melissa Smith, Stephanie Chalk, and Raja Parasuraman. 2012. The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 56. Sage Publications Sage CA: Los Angeles, CA, 263–267.

[27] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22, 3 (2016), 331.

[28] Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics* 12, 2 (2020), 459–478.

[29] Peter De Vries, Cees Midden, and Don Bouwhuis. 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 (2003), 719–735.

[30] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[31] Michael C Dorneich, Rachel Dudley, Emmanuel Letsu-Dake, William Rogers, Stephen D Whitlow, Michael C Dillard, and Erik Nelson. 2017. Interaction of automation visibility and information quality in flight deck information automation. *IEEE Transactions on Human-Machine Systems* 47, 6 (2017), 915–926.

[32] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning.

[33] Na Du, Kevin Y Huang, and X Jessie Yang. 2020. Not all information is equal: effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in human-automation teaming. *Human factors* 62, 6 (2020), 987–1001.

[34] Connor Esterwood, Lionel Robert, et al. 2022. Having The Right Attitude: How Attitude Impacts Trust Repair in Human-Robot Interaction. (2022).

[35] Connor Esterwood and Lionel P Robert. 2021. Do you still trust me? human-robot trust repair strategies. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 183–188.

[36] Md Abdullah Al Fahim, Mohammad Maifi Hasan Khan, Theodore Jensen, Yusuf Albayram, and Emil Coman. 2021. Do integral emotions affect trust? The mediating effect of emotions on trust in the context of human-agent interaction. In *Designing Interactive Systems Conference 2021*. 1492–1503.

[37] Chong Feng, Muzammil Khan, Arif Ur Rahman, and Arshad Ahmad. 2020. News recommendation systems-accomplishments, challenges & future directions. *IEEE Access* 8 (2020), 16702–16725.

[38] Hiroshi Fujita. 2020. AI-based computer-aided diagnosis (AI-CAD): the latest review to read first. *Radiological physics and technology* 13, 1 (2020), 6–19.

[39] Ji Gao and John D Lee. 2006. Effect of shared information on trust and reliance in a demand forecasting task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. SAGE Publications Sage CA: Los Angeles, CA, 215–219.

[40] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[41] Harjinder Gill, Kathleen Boies, Joan E Finegan, and Jeffrey McNally. 2005. Antecedents of trust: Establishing a boundary condition for the relation between propensity to trust and intention to trust. *Journal of business and psychology* 19, 3 (2005), 287–302.

[42] Rui Ying Goh and Lai Soon Lee. 2019. Credit scoring: a review on support vector machines and metaheuristic approaches. *Advances in Operations Research* 2019 (2019).

[43] Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. 2022. To Trust or Not To Trust: How a Conversational Interface Affects Trust in a Decision Support System. In *Proceedings of the ACM Web Conference 2022*. 3531–3540.

[44] Svyatoslav Guznov, Alexander Nelson, Joseph Lyons, and David Dycus. 2015. The effects of automation reliability and multi-tasking on trust and reliance in a simulated unmanned system control task. In *International Conference on Human-Computer Interaction*. Springer, 616–621.

[45] Kasper Hald, Matthias Rehm, and Thomas B Moeslund. 2021. Human-Robot Trust Assessment Using Top-Down Visual Tracking After Robot Task Execution Mistakes. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 892–898.

[46] Yugo Hayashi and Kosuke Wakabayashi. 2017. Can AI become reliable source to support human decision making in a court scene?. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 195–198.

[47] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 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.

[48] Kai Holländer, Philipp Wintersberger, and Andreas Butz. 2019. Overtrust in external cues of automated vehicles: an experimental investigation. In *Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications*. 211–221.

[49] Brittany E Holthausen, Philipp Wintersberger, Bruce N Walker, and Andreas Riener. 2020. Situational trust scale for automated driving (STS-AD): Development and initial validation. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 40–47.

[50] Sandy H Huang, Kush Bhatia, Pieter Abbeel, and Anca D Dragan. 2018. Establishing appropriate trust via critical states. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3929–3936.

[51] Aya Hussein, Sondoss Elsawah, and Hussein A Abbass. 2020. Trust mediating reliability–reliance relationship in supervisory control of human–swarm interactions. *Human Factors* 62, 8 (2020), 1237–1248.

[52] Lakshmi Shankar Iyer. 2021. AI enabled applications towards intelligent transportation. *Transportation Engineering* 5 (2021), 100083.

[53] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. , 624–635 pages.

[54] Theodore Jensen, Yusuf Albayram, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, Ross Buck, and Emil Coman. 2019. The apple does fall far from the tree: user separation of a system from its developers in human-automation trust repair. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 1071–1082.

[55] Theodore Jensen, Mohammad Maifi Hasan Khan, and Yusuf Albayram. 2020. The role of behavioral anthropomorphism in human-automation trust calibration. In *International Conference on Human-Computer Interaction*. Springer, 33–53.

[56] Theodore Jensen, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, and Yusuf Albayram. 2021. Trust and Anthropomorphism in Tandem: The Interrelated Nature of Automated Agent Appearance and Reliability in Trustworthiness Perceptions. In *Designing Interactive Systems Conference 2021*. 1470–1480.

[57] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.

[58] Craig J Johnson, Mustafa Demir, Nathan J McNeese, Jamie C Gorman, Alexandra T Wolff, and Nancy J Cooke. 2021. The Impact of Training on Human–Autonomy Team Communications and Trust Calibration. *Human factors* (2021), 00187208211047323.

[59] Stefanie Maria Jungmann, Timo Klan, Sebastian Kuhn, and Florian Jungmann. 2019. Accuracy of a Chatbot (ADA) in the diagnosis of mental disorders: comparative case study with lay and expert users. *JMIR formative research* 3, 4 (2019), e13863.

[60] Alexandra D Kaplan, Theresa T Kessler, J Christopher Brill, and PA Hancock. 2021. Trust in artificial intelligence: Meta-analytic findings. *Human Factors* (2021), 00187208211013988.

[61] Siddartha Khastgir, Stewart Birrell, Gunwant Dhadyalla, and Paul Jennings. 2018. Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation research part C: emerging technologies* 96 (2018), 290–303.

[62] Taenyun Kim and Hayeon Song. 2021. How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics* 61 (2021), 101595.

[63] Molly Kluck, Spencer C Kohn, James C Walliser, Ewart J de Visser, and Tyler H Shaw. 2018. Stereotypical of Us to Stereotype Them: The Effect of System-Wide Trust on Heterogeneous Populations of Unmanned Autonomous Vehicles. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 62. SAGE Publications Sage CA: Los Angeles, CA, 1103–1107.

[64] Spencer C Kohn, Daniel Quinn, Richard Pak, Ewart J De Visser, and Tyler H Shaw. 2018. Trust repair strategies with self-driving vehicles: An exploratory study. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 62. Sage Publications Sage CA: Los Angeles, CA, 1108–1112.

[65] Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*. Springer, 13–30.

[66] ES Kox, LB Siegling, and JH Kerstholt. 2022. Trust Development in Military and Civilian Human–Agent Teams: The Effect of Social-Cognitive Recovery Strategies. *International Journal of Social Robotics* (2022), 1–16.

[67] Esther S Kox, José H Kerstholt, Tom F Hueting, and Peter W de Vries. 2021. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous agents and multi-agent systems* 35, 2 (2021), 1–20.

[68] Johannes Kraus, David Scholz, Dina Stiegemeier, and Martin Baumann. 2020. The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human factors* 62, 5 (2020), 718–736.

[69] Johannes Maria Kraus, Yannick Forster, Sebastian Hergeth, and Martin Baumann. 2019. Two routes to trust calibration: effects of reliability and brand information on trust in automation. *International Journal of Mobile Human Computer Interaction (IJMHCI)* 11, 3 (2019), 1–17.

[70] Alexander Kunze, Stephen J Summerskill, Russell Marshall, and Ashleigh J Filtness. 2019. Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics* 62, 3 (2019), 345–360.

[71] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[72] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J König, and Nina Grgić-Hlača. 2022. "Look! It'sa Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?. In *CHI Conference on Human Factors in Computing Systems*. 1–28.

[73] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.

[74] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[75] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2023. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior* 139 (2023), 107539.

[76] Stephan Lewandowsky, Michael Mundy, and Gerard Tan. 2000. The dynamics of trust: comparing humans to automation. *Journal of Experimental Psychology: Applied* 6, 2 (2000), 104.

[77] Mengyao Li, Brittany E Holthausen, Rachel E Stuck, and Bruce N Walker. 2019. No risk no trust: Investigating perceived risk in highly automated driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 177–185.

[78] Cynthia Liem, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, and Cornelius J König. 2018. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and interpretable models in computer vision and machine learning*. Springer, 197–253.

[79] James Llinas, Ann Bisantz, Colin Drury, Younho Seong, and Jiun-Yin Jian. 1998. *Studies and analyses of aided adversarial decision making. phase 2: Research on human trust in automation.* Technical Report. STATE UNIV OF NEW YORK AT BUFFALO CENTER OF MULTISOURCE INFORMATION FUSION.

[80] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[81] Yidu Lu and Nadine Sarter. 2019. Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and system reliability.

IEEE Transactions on Human-Machine Systems 49, 6 (2019), 560–568.

[82] Yidu Lu and Nadine Sarter. 2019. Feedback on system or operator performance: Which is more useful for the timely detection of changes in reliability, trust calibration and appropriate automation usage?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 312–316.

[83] Joseph B Lyons, Izz aldin Hamdan, and Thy Q Vo. 2023. Explanations and trust: What happens to trust when a robot partner does something unexpected? *Computers in Human Behavior* 138 (2023), 107473.

[84] Stefanie M. Faas, Johannes Kraus, Alexander Schoenhals, and Martin Baumann. 2021. Calibrating Pedestrians' Trust in Automated Vehicles: Does an Intent Display in an External HMI Support Trust Calibration and Safe Crossing Behavior?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[85] Poornima Madhavan and Douglas A Wiegmann. 2007. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science* 8, 4 (2007), 277–301.

[86] M. Madsen and S. Gregor. 2000. Measuring Human-Computer Trust. In *Proceedings of 11th Australasian Conference on Information Systems*. 6e8.

[87] JB Manchon, Romane Beaufort, Mercedes Bueno, and Jordan Navarro. 2022. Why Does the Automation Say One Thing but Does Something Else? Effect of the Feedback Consistency and the Timing of Error on Trust in Automated Driving. *Information* 13, 10 (2022), 480.

[88] Daniel J McAllister. 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal* 38, 1 (1995), 24–59.

[89] John M McGuirl and Nadine B Sarter. 2003. How are we doing?: Presenting System Confidence Information to Support Trust Calibration and Adaptive Function Allocation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 47. SAGE Publications Sage CA: Los Angeles, CA, 538–542.

[90] John M McGuirl and Nadine B Sarter. 2006. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors* 48, 4 (2006), 656–665.

[91] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research* 13, 3 (2002), 334–359.

[92] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors* 58, 3 (2016), 401–415.

[93] Stephanie M Merritt. 2011. Affective processes in human–automation interactions. *Human Factors* 53, 4 (2011), 356–370.

[94] Stephanie M Merritt, Deborah Lee, Jennifer L Unnerstall, and Kelli Huber. 2015. Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors* 57, 1 (2015), 34–47.

[95] RC Meyer, JH Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[96] David Miller, Mishel Johns, Brian Mok, Nikhil Gowda, David Sirkin, Key Lee, and Wendy Ju. 2016. Behavioral measurement of trust in automation: the trust fall. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 1849–1853.

[97] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[98] Scott Mishler and Jing Chen. 2023. Effect of automation failure type on trust development in driving automation systems. *Applied Ergonomics* 106 (2023), 103913.

[99] Bonnie Marlene Muir. 2002. Operators' trust in and use of automatic controllers in a supervisory process control task. (2002).

[100] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2023. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies* 169 (2023), 102941.

[101] Mollik Nayyar and Alan R Wagner. 2018. When should a robot apologize? understanding how timing affects human-robot trust repair. In *International conference on social robotics*. Springer, 265–274.

[102] Birthe Nesset, David A Robb, José Lopes, and Helen Hastie. 2021. Transparency in hri: Trust and decision making in the face of robot errors. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 313–317.

[103] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.

[104] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *Plos one* 15, 2 (2020), e0229132.

[105] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic reviews* 10, 1 (2021), 1–11.

[106] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[107] Vlad L Pop, Alex Shrewsbury, and Francis T Durso. 2015. Individual differences in the calibration of trust in automation. *Human factors* 57, 4 (2015), 545–556.

[108] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[109] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[110] Dalai Dos Santos Ribeiro, Gabriel Diniz Junqueira Barbosa, Marisa Do Carmo Silva, Hélio Lopes, and Simone Diniz Junqueira Barbosa. 2021. Exploring the impact of classification probabilities on users' trust in ambiguous instances. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 1–9.

[111] Jorge Ribeiro, Rui Lima, Tiago Eckhardt, and Sara Paiva. 2021. Robotic process automation and artificial intelligence in industry 4.0–a literature review. *Procedia Computer Science* 181 (2021), 51–58.

[112] Paul Robinette, Ayanna M Howard, and Alan R Wagner. 2015. Timing is key for robot trust repair. In *International conference on social robotics*. Springer, 574–583.

[113] Julian Sanchez, Arthur D Fisk, and Wendy A Rogers. 2006. What determines appropriate trust of and reliance on an automated collaborative system? Effects of error type and domain knowledge. In *2006 9th International Conference on Control, Automation, Robotics and Vision*. IEEE, 1–6.

[114] Kristin Schaefer. 2013. The perception and measurement of human-robot trust. (2013).

[115] Markus Schedl. 2019. Deep learning in music recommendation systems. *Frontiers in Applied Mathematics and Statistics* (2019), 44.

[116] Beau G Schelble, Jeremy Lopez, Claire Textor, Rui Zhang, Nathan J McNeese, Richard Pak, and Guo Freeman. 2022. Towards Ethical AI: Empirically Investigating Dimensions of AI Ethics, Trust Repair, and Performance in Human-AI Teaming. *Human Factors* (2022), 00187208221116952.

[117] Younho Seong, Ann M Bisantz, and Ann M Bisantz. 2002. Judgment and trust in conjunction with automated decision aids: A theoretical model and empirical investigation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 46. SAGE Publications Sage CA: Los Angeles, CA, 423–427.

[118] Mona SharifHeravi, John R Taylor, Christopher J Stanton, Sandra Lambeth, and Christopher Shanahan. 2020. It'sa Disaster! Factors Affecting Trust Development and Repair Following Agent Task Failure. In *Proceedings of the 2020 Australasian Conference on Robotics and Automation (ACRA 2020), 8-10 December 2020, Brisbane, Queensland.*

[119] Harold Soh, Pan Shu, Min Chen, and David Hsu. 2018. The Transfer of Human Trust in Robot Capabilities across Tasks.. In *Robotics: Science and Systems*.

[120] S Shyam Sundar. 2020. Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication* 25, 1 (2020), 74–88.

[121] Jonathan Tallant and Donatella Donati. 2020. Trust: from the Philosophical to the Commercial. *Philosophy of Management* 19, 1 (2020), 3–19.

[122] Claire Textor, Rui Zhang, Jeremy Lopez, Beau G Schelble, Nathan J McNeese, Guo Freeman, Richard Pak, Chad Tossell, and Ewart J de Visser. 2022. Exploring the Relationship Between Ethics and Trust in Human–Artificial Intelligence Teaming: A Mixed Methods Approach. *Journal of Cognitive Engineering and Decision Making* 16, 4 (2022), 252–281.

[123] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In *CHI Conference on Human Factors in Computing Systems*. 1–17.

[124] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second chance for a first impression? Trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*. 77–87.

[125] Qingqing Tu and Le Dong. 2010. An intelligent personalized fashion recommendation system. In *2010 International Conference on Communications, Circuits and Systems (ICCCAS)*. IEEE, 479–485.

[126] Daniel Ullrich, Andreas Butz, and Sarah Diefenbach. 2021. The development of overtrust: An empirical simulation and psychological analysis in the context of human–robot interaction. *Frontiers in Robotics and AI* 8 (2021), 554578.

[127] James C Walliser, Ewart J de Visser, and Tyler H Shaw. 2016. Application of a system-wide trust strategy when supervising multiple autonomous agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 133–137.

[128] Fei-Yue Wang. 2008. Toward a revolution in transportation operations: AI for complex systems. *IEEE Intelligent Systems* 23, 6 (2008), 8–13.

[129] Lu Wang, Greg A Jamieson, and Justin G Hollands. 2008. Improving reliability awareness to support appropriate trust and reliance on individual combat identification systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 52. SAGE Publications Sage CA: Los Angeles, CA, 292–296.

[130] Ning Wang, David V Pynadath, and Susan G Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 109–116.

[131] Pei Wang. 2019. On defining artificial intelligence. *Journal of Artificial General Intelligence* 10, 2 (2019), 1–37.

[132] Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Transactions on Interactive Intelligent Systems (TiiS)* (2022).

[133] Douglas A Wiegmann, Aaron Rich, and Hui Zhang. 2001. Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science* 2, 4 (2001), 352–367.

[134] Yihan Wu and Ryan M Kelly. 2020. Online Dating Meets Artificial Intelligence: How the Perception of Algorithmically Generated Profile Text Impacts Attractiveness and Trust. In *32nd Australian Conference on Human-Computer Interaction*. 444–453.

[135] Yaqi Xie, Indu P Bodala, Desmond C Ong, David Hsu, and Harold Soh. 2019. Robot capability and intention in trust-based decisions across tasks. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 39–47.

[136] Jin Xu and Ayanna Howard. 2022. Evaluating the Impact of Emotional Apology on Human-Robot Trust. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1655–1661.

[137] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[138] X Jessie Yang, Christopher Schemanske, and Christine Searle. 2021. Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *arXiv preprint arXiv:2107.07374* (2021).

[139] X Jessie Yang, Vaibhav V Unhelkar, Kevin Li, and Julie A Shah. 2017. Evaluating effects of user experience and system transparency on trust in automation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 408–416.

[140] Michelle Yeh and Christopher D Wickens. 2001. Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors* 43, 3 (2001), 355–365.

[141] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[142] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd international conference on intelligent user interfaces*. 307–317.

[143] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do i trust my machine teammate? an investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 460–468.

[144] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *CHI Conference on Human Factors in Computing Systems*. 1–28.

[145] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.