# A Review on Human–Machine Trust Evaluation: Human-Centric and Machine-Centric Perspectives

Biniam Gebru, Lydia Zeleke, Daniel Blankson, Mahmoud Nabil, Shamila Nateghi, Abdollah Homaifar, and Edward Tunstel

***Abstract*—As complex autonomous systems become increasingly ubiquitous, their deployment and integration into our daily lives will become a significant endeavor. Human–machine trust relationship is now acknowledged as one of the primary aspects that characterize a successful integration. In the context of human–machine interaction (HMI), proper use of machines and autonomous systems depends both on the human and machine counterparts. On one hand, it depends on how well the human relies on the machine regarding the situation or task at hand based on willingness and experience. On the other hand, it depends on how well the machine carries out the task and how well it conveys important information on how the job is done. Furthermore, proper calibration of trust for effective HMI requires the factors affecting trust to be properly accounted for and their relative importance to be rightly quantified. In this article, the functional understanding of human–machine trust is viewed from two perspectives—human-centric and machine- centric. The human aspect of the discussion outlines factors, scales, and approaches, which are available to measure and calibrate human trust. The discussion on the machine aspect spans trustworthy artificial intelligence, built-in machine assurances, and ethical frameworks of trustworthy machines.**

***Index Terms*—Human-machine trust, trust measurement, trust calibration, machine trustworthiness.**

## I. Introduction

As AUTONOMOUS systems become increasingly complex, the interaction between these systems and human users/operators relies heavily on how much and how well the users/operators trust them. Overtrust and lack of trust on the users' behalf lead to disuse and misuse of autonomous systems, respectively [1]. The following incidents demonstrate examples of the above-mentioned situations. In 2018, an Uber car operating in self-driving mode fatally struck a pedestrian on a bicycle in Arizona, USA [2]. The investigation by National Transportation Safety Board (NTSB) found that the vehicle's automatic system failed to identify the pedestrian and her bicycle as an imminent collision danger, as they were supposed to. NTSB has also found that during the accident, the human driver in the car had been streaming a television episode leaving the self-driving mode unattended. This suggests an overreliance on the system leading to misuse of autonomy. Another example is an incident with Southwest Airlines Flight 1455 that traveled from Las Vegas to Burbank, California, in March 2000 [3]. During the landing approach, cockpit warning signals alerted the captain and first officer that the flight speed and angle of descent were well outside the glide path. These warnings were ignored. As a result, the plane overran the runway, and crashed through a fence and wall. This incident fortunately had no fatalities, but illustrates how lack of trust may lead to disuse of autonomous systems. These examples also imply the need for appropriate and calibrated reliance on such systems. The seminal works [4] and [5] laid the motivational foundation by not only identifying trust as a major factor for the interaction between humans and autonomous systems involving uncertain situations, but also the necessity to calibrate trust correctly.

Trust has also been recognized to be one of the important factors for effective interaction and use of autonomous systems. According to [5] and [6], the critical outcome of trust is described to be reliance. While using these increasingly complex systems, we may not be concerned about trust itself, but the ultimate behaviors that trust is likely to produce—reliance or absence of it, i.e., nonreliance. Thus, the importance of trust in facilitating HMI and integration of autonomous systems into everyday use is noteworthy and necessary.

Trust in automated and autonomous systems has been explored in earlier research along with the notions of use, misuse, and disuse of automation [1]; the theoretical foundations of modeling human trust that can be used for empirical studies of human intervention in automated systems [4]; and trust consideration in automation design for appropriate reliance in [5]. Other topics ranging from trust dynamics in autonomous vehicles to trust etiquette and trust-distrust definitions were discussed in [7]–[12]. Furthermore, several survey papers have discussed human–machine trust with regard to multidisciplinary definitions of trust, trust frameworks, and factors affecting trust [13]. Other survey papers include studies of algorithmic assurances in human–autonomy trust relationships [14] and computational trust and reputation models [15] that focused on models applied to nonengineering fields. Notably, the review presented

Manuscript received 3 May 2021; revised 26 September 2021; accepted 17 December 2021. Date of publication 24 February 2022; date of current version 15 September 2022. This work was supported by the Air Force Research Laboratory and the OSD under Grant FA8750-15-2-0116, in part by NASA University Leadership Initiative (ULI) under Grant 80NSSC20M0161, and in part the National Institute of Aerospace's Langley Distinguished Professor Program under Grant C16-2B00-NCAT. This article was recommended by Associate Editor V. Fuccella. *(Corresponding author: Abdollah Homaifar.)*

Biniam Gebru, Lydia Zeleke, Daniel Blankson, Mahmoud Nabil, Shamila Nateghi, and Abdollah Homaifar are with the North Carolina A&T State University, Greensboro, NC 27411 USA (e-mail: btgebru@aggies.ncat.edu; lzeleke@aggies.ncat.edu; dblankson@aggies.ncat.edu; mnmahmoud@ncat.edu; snateghiboroujeni@ncat.edu; homaifar@ncat.edu).

Edward Tunstel is with the Motiv Space Systems, Inc., Pasadena, CA 91107 USA (e-mail: eddie.tunstel@rtx.com).

Digital Object Identifier 10.1109/THMS.2022.3144956

discussed how to conceptualize trust variability among various other research outputs on trust in automation.The model revealed three layers of human–automation trust variability. This layered and structured approach comprises dispositional, situational, and learned trust. Another review [17] summarized the research concerning autonomous systems by considering four categories of such systems, namely decision support systems, robots, self-driving cars, and autopilot systems. This article provided a quantitative definition of trust that appeared in the authors' prior work [18]. However, the quantitative definition provided has not been utilized as a criterion for evaluating any of the subsequent discussions presented there. Most recently, Kok and Soh [19] addressed how trust in robots can be gained, maintained, and calibrated. This article also outlined the challenges in terms of measurement of trust, and trust models in real-world scenarios.

Despite the fact that previous studies covered important themes in relation to trust in automation, there still remains much ongoing discussion regarding trust measurement and cal ibration. This includes standardized measurement techniques, standardized scales, and exploring trust calibration efforts that consider both the human and machine aspects of trust. In this article, a review of research outputs pertaining to human trust measurement approaches and trustworthiness of machines is provided. This article intends to provide a broader review of the topic focusing particularly on the following:

1) *Human and Machine Aspects of Trust:* This brings attention to a dual perspective of trust in human–machine teams. In the HMI context, an improved and effective use of machines/autonomous systems depends on both the human and machine counterparts. Proper calibration of the interaction requires a functional understanding of both aspects, and hence the relevance of this contribution.
2) *Human Trust Models, Measurements, and Calibration:* This includes a review of models used for studying trust evaluation with respect to autonomous systems, summary of various trust measurement, and experimental and empirical ways of trust calibration.
3) *Machine Trustworthiness:* It aims to cultivate machine trustworthiness and acceptance in society and a review of ethical frameworks, which lays the foundations of design, development, and deployment of trustworthy autonomous systems (TASs) to cultivate machine trustworthiness and acceptance in society. Also, a discussion on properties of trustworthy autonomy, verification of trustworthiness, and methods of trust calibration and assurance are also presented.

The remaining sections of this article are organized as follows. Section II starts with a definition of trust that subsumes the context of its use in the reviewed materials. Section III primarily focuses on the discussion of the human trust measurements and calibration along with validity and reliability testing requirements. In Section IV, a review of machine aspects of trust and trust influence based on the trustworthiness of autonomous systems are discussed. In Section V, the challenges of measuring and calibrating trust in light of the reviewed material are also discussed. Finally, Section VI presents the conclusions and the road map ahead for more studies.

## II. Background and Definitions

With the increased complexity of autonomous systems, understanding all the underlying operations of those systems becomes difficult. Yet, users will continue to delegate tasks, trusting their machine partners to various degrees in different settings. Therefore, it is important to discuss a definition of trust and autonomy that relates to the majority of the works reviewed in this article. We start by providing the definition of autonomy and follow up with a definition of trust.

Although it is quite difficult to provide a governing definition of autonomy without the situational context of the application, Fisher *et al.* [20] defined autonomous systems as those systems that decide for themselves what to do and when to do it. In explaining this idea, Bradshaw *et al.* [21] emphasized that autonomy entails at least two dimensions: 1) self-directedness, which describes independence of an agent from its physical environment and social context; and 2) self-sufficiency, which describes self-generation of goals by the agent.

Trust has been studied in a broad range of fields of interest, including psychology, sociology, cognitive sciences, economics, computer science, human factors, and engineering. As a result, a unifying definition for trust is lacking. In particular, within the framework of HMI, the definitions span from the human-inspired notion of trust [22] to a computational model of trust accounting for cyber and network security [23], [24]. However, there exists a commonality across usage, which includes three components. These common components are: a trustor, a trustee, and a risk or uncertainty [16]. The definitions describe the relationship between the trustor (a user/operator) and the trustee (a machine) depending on the nature of the task, consequences, and conditions.

Trust is described in [5] as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." This definition of trust is by far the most widely used in subsequent studies of trust in automation [25]. In [19], this definition is elaborated to include the following:

1) consideration of trust arising in situations that entail risk;
2) the multidimensional and latent nature of trust;
3) trust will also account for the relationship between past experiences with the trustee and subsequent acts of reliance.

Combining these, [19] provided a definition of trust as "a multidimensional latent variable that mediates the relationship between events in the past and the trustor's subsequent choice of relying on the trustee in an uncertain environment."

A literature review of research covering trust in autonomous systems, mostly on materials published in the last decade, was conducted to build on the previous studies carried out in [16] and [17] using combinations of key terms, such as *trust, trust in autonomous systems, trust measurement, trust scales, trustworthiness, and trust calibration*.

In this article, the functional understanding of human–machine trust is viewed from two perspectives—the human aspect and the machine aspect. The human aspect of the discussion outlines trust models that have been developed and adopted through the years to characterize trust and facilitate trust measurement toward trust calibration. The discussions

TABLE I
Rust Models

| Authors | Description |
|---|---|
| Muir (1994) | Relationship between automation and operators' trust towards calibration and prediction of trust [4]. |
| Lee and See (2004) | A conceptual model of the dynamic process that accounts for contextual variations of individuals' trust evolution and its effect on reliance [5]. |
| Hancock et al (2011) | Human, robot and environmental - related factors affecting human-robot interaction [9]. |
| Schaefer et al (2014) | Extended factors affecting human-robot trust dynamics, including cognitive and emotive factors [29]. |
| Hoff and Bashir (2015) | Conceptual layered classification of trust types accounting for trust variability. Identified layers as - dispositional, situational, and learned trust types [16]. |
| Hoffmann and Söllner (2014) | Latent trust formation. Modeling trust in automated systems via dimensions of performance, process and purpose [30]. |
| Ekman et al (2018) | Modeled trust formation process based on specific driving scenarios for automated vehicle systems [31]. |

on the machine aspect span trustworthy artificial intelligence (AI), built-in machine assurances, and ethical frameworks of trustworthy machines.

## III. Human Trust Measurements

The U.S. Department of Defense's unmanned systems integrated roadmap [26] and the Institute of Defense Analysis's roadmap to trustworthy autonomy [6] emphasized the role of trust as an important determinant of reliance on autonomous systems. Establishing the appropriate levels of trust and trust calibration capabilities are part of the challenge areas identified in those roadmaps. Recalling whether someone trusts a system depends on the nature of the task, the consequences, and conditions [5]. Measurements of trust at those levels of task and conditions will be required to assess the situational calibration of operator trust. To facilitate our discussion of human trust measurement and calibration attempts, we subsequently outline trust models, validity considerations of trust measurement scales, and different types of measurement approaches.

### A. Trust Models

Researchers modeled human trust based on social interpersonal interactions [27], [28] to represent the psychological state or attitude toward delegation of tasks under implied risk. Various models for trust evaluation and explanation providing reproducible and methodical treatment of trust in HMI have been put forth by several research outputs. Table I summarizes some of these models that have been widely adopted. The early models of trust include those developed in [4] and [5]. These models adopted and modified the notion of interpersonal trust for use in automated systems. In [4], a qualitative model used to explain human–automation interaction to make behavioral predictions about a human operator toward automation and calibration of trust was provided. Furthermore, in [5], Lee and See devised a model that captures 1) the dynamic interaction among contextual awareness of individuals and the environment and 2) the design aspects of the interface between the automation and the operator.

Other widely used models of trust include those introduced in [9], [16], [30], and [31]. These models describe an increasingly detailed representation of trust and autonomy, factors affecting the relationship and trust formation under these interactions, and accounts of the multidimensionality of trust as a latent behavior.

### B. Trust Measurement—Scales and Approaches

Due to its latent and multifaceted nature, trust cannot be directly measured. As a result, measuring trust depends on capturing other factors or underlying constructs. As different types of trust are investigated: trust propensity [32], dispositional trust [16], history-based trust [33], and affective and cognitive trust [5], what is measured continues to vary widely within the existing literature. Trust measurements are mainly done via self-reports by the human subjects under study [34]–[38]. Another approach that has grown in recent years is the use of psycho-physiological sensors to measure neural and physical correlates of trust [10], [39]–[43].

In [44], Wei *et al.* pointed out that, despite the availability of a large number of research in trust measurement, it is not clear what psychometric level of measurement is most appropriate for trust in automation. They discussed the various levels of psychometric scales (nominal, ordinal, interval, and ratio) and investigated what could be a step toward having a standardized level of measurement for trust. They did so by carrying out a trust measurement using an experimental setup and measurement approach, which allows a permissible transformation between two levels of measurement. They found out that the interval level psychometric scale, adopted by the majority of trust measurement studies, might be a probable candidate for determining the most appropriate level of trust measurement. Also, Brzowski and Nathan-Roberts [45] iterated the need to validate trust scales. They indicated that most measurements are created with the exact application in mind and are validated in ways that are specifically relevant to the application (e.g., simulation of a decision aid) used for measurement purposes. Validity refers to the degree to which a construct is correctly measured in a quantitative study [46]. Another issue related with trust measurements is the reliability of those measurements. Reliability refers to the consistency of measurement [47]. It is the way in which a trust scale consistently produces the same results when used repeatedly in the same scenario. According to [19], validity, reliability, and related measurement variances need to be given the necessary attention to have reproducible trust measurement practices. Next, we discuss various subjective self-report measures and psychophysiological trust measurements.

*Self-Report Measurements:* These approaches measure trust by collecting responses in the form of surveys and questionnaires from participants. There is currently no standard for assessing human–machine trust [44], [48]. At present, it is measured by researchers using custom scales or validated measurements. This makes cross-study comparisons difficult to determine whether automation facilitates appropriate trust levels. Researchers have used a wide variety of approaches in measuring trust with self-reported accounts from human subjects under study [34]–[38]. In the following, the widely used self-report based scales are discussed.

In [37], Jian *et al.* developed one of the most frequently used scales, i.e., the Scale of Trust in Automated Systems, where they devised a 12-item scale to measure trust. They developed an experiment using a word elicitation study, a questionnaire study, and a paired comparison study. The 12 factors characterizing trust between people and machines were identified based on a cluster analysis carried out on the experimental data. In their conclusion, they particularly noted that the trust scale developed should be validated in experiments designed for a specific study of trust in automation.

In [49] and [50], a validated scale for assessing changes in a person's trust in a robot was developed. The 40-item trust instrument, developed over the course of six experiments, was designed to measure trust perceptions in the context of human–robot interaction. For this item pool, two experiments identified the robot and perceived functional features. The scale was reduced to those items from an original pool of 172 items, using item pool reduction techniques and content validation by subject matter experts. The scale was then validated through two final experiments.

The multidimensional nature of trust was employed in [51] to incorporate capacity trust (reliable, capable) and moral trust (sincere, ethical) aspects as factors. They developed a Multi-Dimensional Measure of Trust (MDMT) instrument, which captures the two trust factors. Repeated cycles of principal components analysis and item analysis resulted in study items distributed across four factors: reliable, capable, sincere, and ethical. The MDMT provided a tool that brings the interpersonal trust construct (moral) and a human–machine construct (capacity) into an amalgamated scale. This instrument, however, is yet to be validated.

The Trust of Automated Systems Test (TOAST) measuring trust based on the theory of trust formation was proposed in [48]. Scale structure reliability and criteria validity were evaluated with civilian and military-affiliated samples. There, TOAST demonstrated a reliable, two-factor model representing system performance and understanding. The scores obtained through confirmatory factor analysis on the two factors demonstrated strong positive correlations within those factors.

Other trust scales include the perfect automation schema scale in [52], the psychometric instrument designed to measure cognitive and affective components of human–computer trust in [53], and the early scales include operators' trust and prediction of trust scale in [4] and the trust evolution and reliance scale in [5].

*Psychophysiological Measurements:* Although self- report trust measurements provide valuable insight into understanding users' trust in automation, they are unable to objectively evaluate user trust or trust correlates and are, therefore, not appropriate for real-time trust assessment [54]. Progress with sensing technology has resulted in the development of inexpensive and efficient psychophysiological sensors and a shift away from self-reporting scales toward more objective methods for assessing trust correlates using psychophysiological signals. Measurements using physiological traits, such as employing heart-rate variability (HRV) measurement [55], evaluation of brain activity using fMRI [56], [57], and fNIRS [58], and electroencephalogram (EEG) [59]–[63], have been used to study and evaluate trust via its psychophysiological correlates.

In [64], Gupta *et al.* investigated participants trust toward an auditory assistance system for a search task in a virtual reality (VR) environment. They found that the trust inferred this way is subject to different cognitive loads and agent accuracy levels. Participants data were gathered using a variety of sensors, including EEG, HRV, and Galvanic skin response (GSR) devices. The study identified that physiological and behavioral measurements can be used to assess human trust in virtual agents. It also showed that researchers can use a VR environment to simulate realistic environmental scenarios. Furthermore, they studied the effect of cognitive load on trust. Human-like cues were also demonstrated to play an important part in the neural response to an agent's technical capability [65]. Subjects played matrix games with a computerized agent while event-related potentials from EEG were averaged for each subject and trial. By analyzing brain signals, this study creates an environment that capability-based trust levels can be computed from those measurements.

In [59], an empirical trust sensor model was proposed using data from GSR and EEG and showed that psychophysiological signals could be used for real-time human trust measurement inferences in agents. In particular, Akash *et al.* [59] developed a binary classification model for trust and distrust based on the data they collected in this manner. They further expanded their experiments [66] to dynamically vary and calibrate automation transparency to optimize HMIs.

The application of psychophysiological signals to assess trust via its correlates is complex [67]. It is usually impossible to draw a one-to-one correspondence between the signals from these sensors and trust states [54]. Subsequently, the interpretation of the relevant signals to draw a causal relationship with trust is still an elusive task [68], [69]. In [54], it was suggested that researchers adjust the trust measurement scale, the number and types of physiological signals, trust relationship type, and the analysis technique used to analyze the data to infer the trust state when assessing trust using these approaches.

## IV. Machine Trustworthiness

### A. TASs and Trust

Trustworthiness is a property of trusted agents or organizations which engenders trust in other agents or organizations [70].

Any AI-driven autonomous system that embodies such properties is categorized under the term “Trustworthy Autonomous System (TAS).” TAS cements human trust in AI-based systems so that societies can assuredly design, develop, and reap the benefits of these technologies [71].

The integration of AI and machine learning is pervasive in many fields and affects various aspects of our daily lives. The lack of trust in AI systems is often a typical yet justifiable hindrance in achieving the next level of autonomous system ubiquity. Trust in intelligent systems is often compromised due to consistent failure in a broad range of systems and the black box behavior inherent in some data-driven models that obscure internal decision-making processes.

Although intelligent system benefits are far-reaching and undeniable, there are many safety-critical scenarios where the failure of AI systems could have detrimental consequences. The general category of reasons that lead to failure includes maliciously compromising system functionalities executed by unethical people (“intentionally”), engineering shortcomings, and even environmental factors [72]. Autonomous road vehicles are, for instance, exposed to adversarial attacks, which involve adding visual perturbations to stop signs to mislead the interpretations of their trained classifiers and ultimately compromise the vehicles’ safety [73].

A timeline consisting of first occurrence of intelligent systems failure, excluding deliberate attacks, is presented in [72]. It is critical to deliver on the advantages of intelligent systems and remit the drawbacks by building trustworthy systems. Additionally, system trustworthiness should be ensured throughout each stage of a product’s life cycle, including the technical design, development, implementation, testing, and deployment stages.

### B. *Ethical Frameworks for TASs*

Amidst the ever-evolving dynamics of HMI, it is essential to continuously monitor and ensure the development of intelligent technology aimed at benefiting both individuals and society at large. This has inspired the development of several frameworks and guidelines that inform key principles underlining the ethical operation of AI-based systems.

These frameworks are also relevant to trust, accentuated by references to trustworthy AI design and development principles for entailing customers’ trust [74]. Ethical frameworks and guidelines for building trustworthy intelligence list a set of qualities that should be adopted by AI-based systems to be deemed trustworthy.

Stakeholders of these technologies, including industries, governmental agencies, and academia, have released guidelines and frameworks around ethical AI. Some international organizations are instituting AI expert committees to draft guidelines. For instance, the High-Level Expert Group on Artificial Intelligence (AI HLEG) mandated by the European Commission [75] defined seven requirements that systems must meet to realize trustworthiness under its proposed trustworthy AI framework. These include technical robustness and safety, transparency, accountability, privacy and data governance, human agency and oversight, diversity, nondiscrimination, and fairness, as well as societal and environmental well-being. The guideline presented by AI HLEG also provides technical and nontechnical methods to meet these requirements. Industries too are establishing guiding principles and frameworks on ethical AI. Deloitte’s Trustworthy AI Framework [76] is an effort that proposes six pillars (transparent/explainable, robust/reliable, fair/impartial, privacy, secure/safe, and accountable/responsible) to consider when designing, developing, and deploying AI-based systems. The similarity in the essence of these principles suggested in both guidelines is reasonably vivid.

It is relevant to capture the global convergence among diverse principles and reconcile their existing differences to reach a consensus on trustworthy innovations. A comprehensive survey of guidelines is provided in [77] by investigating the overlap and divergence of principles and interpretations. This study identified 47 principles across frameworks and later grouped those related principles culminating in the following eight overarching themes:

1) safety and security;
2) transparency and explainability;
3) human control of technology;
4) professional responsibility;
5) promotion of human values;
6) fairness and nondiscrimination;
7) privacy;
8) accountability.

### C. *Elements of Autonomous Systems Affecting Trust*

Autonomous systems garner human trust in their capabilities in more than one way. Several of these methods are discussed in the following section.

*1) Trustworthiness Properties: Robustness and Safety.* Technical robustness in autonomous systems refers to the ability to counteract adverse conditions. Some autonomous systems are built to make sophisticated decisions in safety-critical scenarios. These systems must ensure their users’ safety and exhibit acceptable performance, especially in dynamic environments. There are various angles for addressing robustness in autonomous systems, such as security, reliability, safety, and resilience. Security is a sensitive issue in autonomous systems, given that the systems are more susceptible to attacks.

A survey of security in autonomous systems in [78] highlights major security threats in autonomous systems and provides insights on available cyber-security solutions along with their constraints. Maintaining the security of autonomous vehicles by verifying the integrity of sensor data is proposed in terms of LIDAR [79] and Radar [80] sensors, which use the quantization index modulation based data hiding technique. The research in [81] ensures the safety of autonomous vehicles in unstructured environments and robustness against adversarial attacks by developing controller-focused anomaly detection and system-focused anomaly detection techniques to enhance anomaly monitoring in sensor data and the overall system. Seemingly, minor adversarial perturbations to an AI model’s inputs can severely undermine the model’s reliability when used in safety-critical domains. To this effect, robust visual adversarial perturbations under changing environmental conditions are applied to road sign images causing intentional misclassification in autonomous driving deep neural networks [73], demonstrating how vulnerable these systems can be.

*2) Explainability, Transparency, and Interpretability:* These concepts relate to the degree to which the operation and decision-making processes of AI-based systems are relayed in ways comprehensible to a human user. Humans are reluctant to espouse techniques that are not interpretable and trustworthy [82]. Any explanation of the decisions made by an AI-based system to engender trust in the system is enrolled in the concept of explainable AI (XAI). The nature of interpretability, whether inherent or aided, establishes the two facets of XAI. AI models with inherent interpretability are classified under transparent models, and those with aided interpretability are known to exhibit post-hoc explainability [83]. Transparent models have a level of intrinsic interpretability, whereas post-hoc explainability is due to deliberate incorporation of qualitative or quantitative explainable solutions. Visual [84] and natural language [85] explanations can be used to rationalize and elaborate the decisions of a system to external users who are unfamiliar with the system's inner workings. Black-box models in a data-driven domain such as neural networks adopt model- simplification techniques, like the DeepRED algorithm [86] and block-chain solutions [87], to ascribe transparency to the model. A computational explanation approach is used in [88] where feature coefficients are evaluated to understand their effect on the explainability of interpretable models using decision lists or decision trees. Researchers are often faced with tradeoff scenarios between the increasing complexity of algorithms to accommodate improved performance and the system's interpretability quotient [89].

*3) Verification of Trustworthy Autonomy:* Technological characteristics of trustworthy autonomy, such as reliability and performance efficiency, do not necessarily implicate positive trust responses nor ensure proper trust calibration. People's perceptions of trustworthiness can be affected by the environment, inclination toward technology, complexity of system (uncorroborated with explanations) and different interactions [70]. Trustworthy systems alone do not necessarily impose trust unless provided with the means to verify their capabilities. The importance of providing verifiable claims regarding the robustness, safety, fairness, and privacy protection of AI systems is emphasized as a prerequisite to building trust for those systems [90]. Thus, we draw attention to the notion of "Verified AI," whose objective is to provide evidential assurances for satisfying correctness levels in an AI-based system as a more substantial attempt to warranty trust.

An appeal to represent correctness justifications using mathematical specifications has led to a growing interest in extending formal verification to AI systems. However, traditional formal verification techniques are not always equipped to handle and generalize effectively in advanced intelligent systems. There are five key challenge areas derived from traditionally adopting formal verification methods in autonomous systems [91]. The first challenge is the presence of unknown and stochastic variables in modeling an environment addressed using probabilistic [92] and data-driven [93], [94] approaches to formally model uncertainties in human behavior and physical environments. High-dimensional input and state spaces in machine learning components are another challenge whose modeling has been facilitated using abstractions [95]. Third, formal specification concerns in learning systems are addressed by accommodating a wider mode of task specifications. Quantitative formulations are designed to specify quantitative properties [91]. Algorithmic improvisations [96] and counter example-guided training data generation [95] are some formal method-based research efforts used to optimize data specifications. The fourth challenge, which is designing scalable and efficient computational engines, was maneuvered using modular reasoning [97]. Finally, the theory of formal inductive synthesis [98] is an emerging solution used to address the challenge of synthesizing "correct-by-construction" design of models for a learning system.

A machine learning-based empirical verification of other intelligent systems is suggested as an alternative method to justify claims. Although not fully guaranteed, machine learning is better equipped to evaluate other machine learning- based systems' probabilistic nature. An interesting review of the use of a machine learning approach to verify another machine learning model's capability is provided in [99]. In this article, adaptive stress testing is deployed to validate the performance of the Next Generation Aircraft Collision Avoidance Software (ACAS X). Reinforcement learning is used to estimate and simulate the likeliest near mid-air collision events with aircraft, which opens doors to validate whether the software responds as expected.

An open challenge presented with verification of intelligent systems is the basic difficulty for specification of some abstract trustworthy AI properties, such as transparency. Conversely, trustworthy properties that are routinely verified in autonomous vehicles include safety, robustness, fairness, and privacy. Safety of autonomous vehicles is formally verified online [100]–[102] by continually monitoring vehicle maneuvers using reachability analysis of all possible behaviors given some knowledge of initial states and bounded uncertainty model. A safety analysis of whether a neural network-based controller prevents an unmanned underwater vehicle from colliding with a static object is provided in [103] using an efficient overapproximate reachability scheme. The fairness of a machine learning systems is probabilistically verified using a scalable algorithm and tool (FairSquare) in [104] and [105], respectively. The robustness of machine learning models is formally specified in [106], which can be used as the basis for verification. An analyzer, i.e., AI$^2$, was developed in [105] to certify large neural networks' robustness. Furthermore, the work in [107] formulated ethical policies in unmanned aircraft and executed formal verification of whether an autonomous agent does make ethical decisions.

Formal verification techniques have also been applied to evaluate human–machine interactive systems [108], [109]. They can exhaustively evaluate various categories of usability properties (e.g., reachability, visibility, reliability, and task-related) and mode-confusion related specifications (e.g., unintended side effects, operator authority limits, and inconsistent behavior) associated with formally modeled human–machine interfaces [108]. Therefore, formal verification helps find failure points overlooked by other traditional analysis techniques (e.g., simulations and human subject trials). The analysis of safety operations leads to design patterns that promote successful HMI characterized by calibrated trust. The limitations of formal verification discussed earlier still hold for HMIs involving autonomous systems. State-space explosion, expressiveness power of modeling formalisms, and model validity all determine the effectiveness of a formal verification method [108]. A new element of limitation is introduced in [110], which alludes to the difficulty of formalizing a model that evolves and adapts to change in human behavior. The Symbolic Analysis Laboratory model checker is used in [111] to formally verify the interaction between a human operator and a single unmanned aerial vehicle (UAV). The extended operator function model [109] was also used to model human operator behavior.

*4) Methods of Trust Calibration and Assurance:* User's trust must be consistent with an autonomous system's capabilities to ensure that the system is used within the bounds of its intended purpose. Concepts of calibration, resolution, and specificity are provided in [5] to describe the mismatch between user trust level and automation capabilities. Calibration is defined as "the correspondence between a person's trust in automation and the automation's capabilities" [112], [113]. Resolution refers to how well the range of capabilities in automation maps to the range of trust levels [114]. Moreover, specificity describes the interdependence level between trust and an element of a trustee. Following these definitions, the perspective in [5] emphasizes the importance of sound calibration, high specificity, and high resolution of trust in mitigating mismatch and guiding the design and evaluation of HMI.

The gradual process by which users build their faith in a particular system and become conditioned to its behavior suggests a characterization of low temporal specificity of trust with the system. Users have difficulty instantly building a mental model of a complicated system, thus requiring several interactions with the system. A faster approach in informing users was proposed for a human–machine setting that exposes users to the critical states of the system's policy, yielding an improved understanding of the system's performance and, as a result, efficiently calibrating users' trust. [115]. Israelsen [116] proposes an algorithmic approach, known as algorithmic assurances, to influence calibrated human trust in autonomous systems. An example of such an approach is when an unmanned ground vehicle communicates its competency boundaries, pertaining to a given task, to a human counterpart through the self-confidence score generated from the Factorized Machine Self-Confidence (FaMSeC) [117].

A few other studies on trust calibration include the following. In [118], Robinette *et al.* indicated how the provision of a robot's operational information and an indication of better performance on upcoming tasks on the robot's part is used as a trust repair mechanism. In [119], the investigation was carried out on how user interfaces that communicate both internal and external system awareness may increase the driver's perceived and measured awareness, as well as their trust in the system. Also in [60], Shahrdar *et al.* combined the use of immersive virtual reality experimentation with self-reporting software to gauge and measure the subjects' trust level in a self-driving autonomous vehicle. In their findings, they illustrated how trust could be reduced, escalated, mutated, and rebuilt by adjusting system performance, like the success rate of carrying out a task. Notably, [66] managed to dynamically vary and calibrate automation transparency, and the amount and utility information provided to the human, to optimize HMIs using feedback control policies.

## V. Discussion

In the following, we discuss the challenges concerning human trust measurement and calibration efforts. These ideas reflect both the human-centric aspect of measuring trust and the machine-centric aspect of trustworthiness and its effect on measurement and calibration.

### A. Trust Measurement

Due to the latent nature of trust, it has been challenging to provide a definitive solution toward trust measurement and calibration. As discussed earlier, there have been notable developments that provide both subjective measurements of trust in terms of self-reporting and objective measurements of trust correlates.

The scales based on self-report measurements in [37], [48], [49], and [51]–[53] provided the most widely used scales in trust measurement. The need to validate these scales for their suitability of use under different application settings other than those they were developed under still largely remains. Also, most of the measurements reported on these scales lack confirmatory factor analysis to justify the relationship among the identified trust factors and the latent behavior, trust. The Structural Equation Model method [120] provides one such rigorous factor analysis method in accounting for how a latent behavior like trust is affected by the identified predictors. This approach used multiple regression and factor analysis to identify relationships between latent variables measured by multiple items. However, this method extracts only linear relationships among the factors.

The shift away from self-reporting reliance for trust measurement has paved a way toward using objective measurement approaches of trust correlates based on psycho- physiological sensors such as EEG, fMRI, and GSR. These measurement approaches enable real-time sensing of trust variation in response to the interaction with the machine in use. These methods help us study what the underlying psycho-physiological measured this way reveal about the human's trust state while interacting the machine. These signals provide a rich set of frequency, time, and spatial domain features that can be used to classify trust and distrust responses. However, like the case with self-reporting measurements there is a lack of confirmatory factor analysis and validity studies to cement these methods as definitive measurement approaches of trust and subsequent calibration. Additionally, there is difficulty in establishing a one-to-one relationship between such psycho-physiological correlates and trust states [67]–[69], [121]. Furthermore, the signals from these sensors are typically non-stationary [122]. And the majority of machine learning classification algorithms are based on the assumption of stationarity, and independent data samples [123]. These algorithms do not perform as high as expected for data collected in this a manner [122]. Hence, on top of addressing the issues of validity, reliability, and measurement invariance, it is also worth investigating the signal processing aspects of psycho-physiological sensors as these sensors provide us with a way to look into trust formation, trust measurement and calibration from a different angle by sensing directly observable correlates to trust.

### B. Trustworthiness and Trust Calibration

Lack of standardized trust measurements has hindered research progress in developing effective trust calibration techniques. We have yet to completely characterize the extent and magnitude of how the trustworthiness properties of machines influence trust. To our knowledge, users' trust should be appropriately calibrated to match the system's capabilities in a given situation to ensure a safe and effective HMI. The effects of technical robustness, transparency, fairness, and other properties of a TAS are intuitively linked with trust realization and calibration. Empirical-based research has surfaced in recent years demonstrating that this cause-and-effect relationship between the notions of trustworthiness and trust is more intricate and less crisp than anticipated. For instance, studies [4], [124] uphold the significance of system transparency in instigating regulated trust levels. Users may fall into an overtrust or undertrust category with respect to a system. Okamura and Yamada [125] assert that transparency of these systems alone is not enough to recover from such states of trust. Alternatively, they developed an adaptive trust calibration technique aided with cognitive cues to signal users when improper trust calibration is detected, thus prompting users to adjust their level of trust. Seppelt [126] demonstrated that continuous system feedback promotes trust calibration in automation. In contrast, after investigating various feedback levels, Mackay *et al.* [127] contest that more information does not always positively correlate to trust, allowing for the possibility of negative cognitive load effects. The way a system communicates its decisions to its users should also be tailored to user variability for a competent effect [31]. The composition of feedback relayed to users is investigated and automatically generated using XAI principles to influence desired trust effects in automated vehicles [128].

Although Israelsen *et al.* [117] devised a quantitative approach for trust calibration where an autonomous vehicle reports its self-confidence level depending on its capabilities, the application is highly domain-specific, and the competency boundary of the system is probabilistic. There is a persistent gap in terms of understanding how a machine's trustworthiness maps to the human–trust variable. Quantitatively outlining the association in a generalized manner is a long stride but shows the greatest promise toward effectively designing and implementing autonomous systems that are truly trustable.

Another challenge we may draw from the discussions in previous sections is that although there are hundreds of guidelines from different stakeholders outlining ethical principles that AI-based systems should adopt, less effort has been put in translating those principles into practice. A call for a "Practical-AI ethics" field enforces the development of action-oriented frameworks and guidelines [129]. As such, recent guidelines [75], [130] proposed recommendations on performing the translation. Implementation based on these recommendations demand inputs from researchers from various disciplines and ultimately draw stakeholders (with conflicting priorities) toward a consensus on these implementations.

Much work was presented in terms of quantifying the impact that trustworthy properties of autonomous systems have on human trust. The difficulty of interpreting these properties as relevant to machine-specific functions is a key impediment. Additionally, prior attempt to define and model trustworthy properties has been context-specific and require generalization. A Bayesian network-based trust modeling used in [131] demonstrates where the trust impact of modifying system behavior is evaluated using a utility function. The model enforces system behaviors that yield the highest trust utility value depending on the status of context variables. Despite the effort to computationally relate system properties to trust, the system under investigation [131] is not a learning system, and the system properties reflect user perceptions of them instead of actual measurements.

## VI. Conclusion

Evaluation, measurement, and calibration of trust in relationship to HMI can facilitate and improve the effective integration of complex autonomous systems into everyday use. Without an appropriately calibrated level of trust, there is a tendency to misuse autonomous systems because of overtrusting, or they may fall into disuse because of a distrust in these systems. To make the best use of ever-increasing interactions between humans and complex systems, trust measurement and appropriate calibration have become a critical task to solve. This survey has reviewed a large pool of research output that provide varying solutions toward the measurement and calibration of trust. Most importantly, the survey presents a discussion of the issue from both a human-centric trust measurement and machine-centric trustworthiness perspective. The survey leads the authors to conclude that proper calibration of trust depends on the appropriate factors affecting human trust being properly accounted for; on the relative importance of those factors being correctly measured and validated; and also on the degree which how properly the information regarding the operations, why and how the machine carries out a task the way it does, is relayed to the human user in an understandable way.

On one hand, concerning the human-centric aspect of trust measurement and calibration, the current state-of-the-art trust measurements have shifted from self-reporting approaches to more real-time sensing of trust using psychophysiological sensors. Measurements using these sensors provides an understanding of how human cognitive and physiological responses correlate to the level of trust projected toward how an autonomous system carries out a task. On the other hand, the machine-centric assessment of the same, research on critical aspects of machine trustworthiness influencing trust has gained further momentum to include detailed aspects such as robustness, fairness, and transparency. The review of these influences implies that providing cognitive cues about the machine's state (i.e., a continuous but appropriate level of system feedback about the machine's decision-making processes) and reporting the machine's self-confidence leads to promising results toward adaptive trust calibration.

## REFERENCES

[1] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Hum. Factors*, vol. 39, no. 2, pp. 230–253, 1997.

[2] M. Harris, "NTSB investigation into deadly Uber self-driving car crash reveals lax attitude toward safety," *IEEE Spectr.*, Nov. 2019.

[3] R. K. Dismukes, B. A. Berman, and L. Loukopoulos, *The Limits of Expertise: Rethinking Pilot Error and the Causes of Airline Accidents*. London, U.K.: Routledge, 2017.

[4] B. M. Muir, "Trust in automation: Part I. theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.

[5] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Hum. Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[6] D. Porter, M. McAnally, and C. Bieber, "Trustworthy autonomy: A roadmap to assurance - Part 1: System effectiveness," Inst. Def. Anal., Alexandria, VA, USA, Tech. Rep. AD1131283, 2020.

[7] D. H. McKnight and N. L. Chervany, "Trust and distrust definitions: One bite at a time," in *Trust in Cyber-Societies*. Berlin, Germany: Springer, 2001.

[8] T. B. Sheridan, *Humans and Automation: System Design and Research Issues*. Santa Monica, CA, USA: Human Factors and Ergonomics Society, 2002.

[9] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Hum. Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[10] C. Basu and M. Singhal, "Trust dynamics in human autonomous vehicle interaction: A review of trust models," in *Proc. AAAI Spring Symp. Ser.*, 2016, Art. no. 53307505.

[11] S. M. Merritt and D. R. Ilgen, "Not all trust is created equal: Dispositional and history-based trust in human-automation interactions," *Hum. Factors*, vol. 50, no. 2, pp. 194–210, 2008.

[12] R. Parasuraman and C. A. Miller, "Trust and etiquette in high-criticality automated systems," *Commun. ACM*, vol. 47, no. 4, pp. 51–55, 2004.

[13] J.-H. Cho, K. Chan, and S. Adali, "A survey on trust modeling," *ACM Comput. Surv.*, vol. 48, no. 2, pp. 1–40, 2015.

[14] B. W. Israelsen and N. R. Ahmed, ""Dave … i can assure you … that it's going to be all right …" a definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–37, 2019.

[15] D. D. S. Braga, M. Niemann, B. Hellingrath, and F. B. D. L. Neto, "Survey on computational trust and reputation models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–40, 2018.

[16] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Hum. Factors*, vol. 57, no. 3, pp. 407–434, 2015.

[17] S. Shahrdar, L. Menezes, and M. Nojoumian, "A survey on trust in autonomous systems," in *Proc. Sci. Inf. Conf.*, 2018, pp. 368–386.

[18] M. Nojoumian and D. R. Stinson, "Social secret sharing in cloud computing using a new trust function," in *Proc. 10th Annu. Int. Conf. Privacy, Secur. Trust*, 2012, pp. 161–167.

[19] B. C. Kok and H. Soh, "Trust in robots: Challenges and opportunities," *Curr. Robot. Rep.*, vol. 1, pp. 1–13, 2020.

[20] M. Fisher, L. Dennis, and M. Webster, "Verifying autonomous systems," *Commun. ACM*, vol. 56, no. 9, pp. 84–93, 2013.

[21] J. M. Bradshaw, R. R. Hoffman, D. D. Woods, and M. Johnson, "The seven deadly myths of" autonomous systems," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 54–61, May/Jun. 2013.

[22] M. Nojoumian, "Trust, influence and reputation management based on human reasoning," in *Proc. AAAI Workshop: Incentive and Trust in E-Communities*, Palo Alto, California, USA: AAAI Press, 2015.

[23] J.-H. Cho, A. Swami, and R. Chen, "A survey on trust management for mobile ad hoc networks," *IEEE Commun. Surv. Tut.*, vol. 13, no. 4, pp. 562–583, 2011.

[24] S. Adalı, *Modeling Trust Context in Networks*. Berlin, Germany: Springer, 2013.

[25] B. French, A. Duenser, and A. Heathcote, "Trust in automation," *Intell. Syst.*, vol. 28, no. 1, pp. 84–88, 2018.

[26] United States Department of Defense, *Unmanned Systems Integrated Roadmap FY2011–2036*. Morrisville, NC, USA: Lulu Com, 2015.

[27] C. Castelfranchi and R. Falcone, "Principles of trust for mas: Cognitive anatomy, social importance, and quantification," in *Proc. Int. Conf. Multi Agent Syst.*, 1998, pp. 72–79.

[28] J. Holmes and J. Rempel, "Trust in close relationships," *J. Pers. Social Psychol.*, vol. 49, 1985, Art. no. 07.

[29] K. E. Schaefer *et al.*, "A meta-analysis of factors influencing the development of trust in automation: Implications for human-robot interaction," Army Res. Lab. Aberdeen Proving Ground, Baltimore, MD, USA, Tech. Rep. ADA607926, 2014.

[30] H. Hoffmann and M. Söllner, "Incorporating behavioral trust theory into system development for ubiquitous applications," *Pers. Ubiquitous Comput.*, vol. 18, no. 1, pp. 117–128, 2014.

[31] F. Ekman, M. Johansson, and J. Sochor, "Creating appropriate trust in automated vehicle systems: A framework for HMI design," *IEEE Trans. Hum.-Mach. Syst.*, vol. 48, no. 1, pp. 95–101, Feb. 2017.

[32] S. A. Jessup, T. R. Schneider, G. M. Alarcon, T. J. Ryan, and A. Capiola, "The measurement of the propensity to trust automation," in *Proc. Int. Conf. Hum.-Comput. Interact.*, 2019, pp. 476–489.

[33] S. Merritt and D. Ilgen, "Not all trust is created equal: Dispositional and history-based trust in human-automation interactions," *Hum. Factors*, vol. 50, no. 2, pp. 194–210, May 2008.

[34] L. Buckley, S.-A. Kaye, and A. K. Pradhan, "Psychosocial factors associated with intended use of automated vehicles: A simulated driving study," *Accident Anal. Prevention*, vol. 115, pp. 202–208, 2018.

[35] S.-Y. Chien, M. Lewis, Z. Semnani-Azad, and K. Sycara, "An empirical model of cultural factors on trust in automation," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2014, pp. 859–863.

[36] D. Garcia, C. Kreutzer, K. Badillo-Urquiola, and M. Mouloua, "Measuring trust of autonomous vehicles: A development and validation study," in *Proc. Int. Conf. Hum.-Comput. Interact.*, 2015, pp. 610–615.

[37] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cogn. Ergonom.*, vol. 4, no. 1, pp. 53–71, 2000.

[38] R. E. Yagoda and D. J. Gillan, "You want me to trust a robot? The development of a human–robot interaction trust scale," *Int. J. Social Robot.*, vol. 4, no. 3, pp. 235–248, 2012.

[39] I. Ajenaghughrure, S. Sousa, and D. Lamas, "Measuring trust with psychophysiological signals: A systematic mapping study of approaches used," *Multimodal Technol. Interact.*, vol. 4, no. 3, Sep. 2020, Art. no. 63.

[40] W. Payre, J. Cestac, and P. Delhomme, "Fully automated driving: Impact of trust and practice on manual control recovery," *Hum. Factors*, vol. 58, no. 2, pp. 229–241, 2016.

[41] J. M. Bindewald, C. F. Rusnock, and M. E. Miller, "Measuring human trust behavior in human-machine teams," in *Proc. Int. Conf. Appl. Hum. Factors Ergonom.*, 2017, pp. 47–58.

[42] K. Yu, S. Berkovsky, R. Taib, D. Conway, J. Zhou, and F. Chen, "User trust dynamics: An investigation driven by differences in system performance," in *Proc. 22nd Int. Conf. Intell. User Interfaces*, 2017, pp. 307–317.

[43] E. J. De Visser *et al.*, "A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents," *Hum. Factors*, vol. 59, no. 1, pp. 116–133, 2017.

[44] J. Wei, M. L. Bolton, and L. Humphrey, "The level of measurement of trust in automation," *Theor. Issues Ergonom. Sci.*, vol. 22, no. 3, pp. 274–295, 2020.

[45] M. Brzowski and D. Nathan-Roberts, "Trust measurement in human–automation interaction: A systematic review," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2019, pp. 1595–1599.

[46] K. Fitzner, "Reliability and validity: A quick review," *Diabetes Educator*, vol. 33, no. 5, pp. 775–780, Sep. 2007.

[47] E. Elsayed, "Overview of reliability testing," *IEEE Trans. Rel.*, vol. 61, no. 2, pp. 282–291, Jun. 2012.

[48] H. Wojton, D. Porter, S. Lane, C. Bieber, and P. Madhavan, "Initial validation of the trust of automated systems test (TOAST)," *J. Social Psychol.*, vol. 160, no. 2, pp. 1–16, Apr. 2020.

[49] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the 'Trust perception Scale-HRI'," in *Proc. Robust Intell. Trust Auton. Syst.*, 2016, pp. 191–218.

[50] K. Schaefer, "The perception and measurement of human-robot trust," Ph.D. dissertation, Dept. Model. Simul. College Sci., Univ. Central Florida, Orlando, FL, USA, 2013.

[51] D. Ullman and B. F. Malle, "What does it mean to trust a robot? Steps toward a multidimensional measure of trust," in *Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2018, pp. 263–264.

[52] S. M. Merritt, J. L. Unnerstall, D. Lee, and K. Huber, "Measuring individual differences in the perfect automation schema," *Hum. Factors*, vol. 57, no. 5, pp. 740–753, 2015.

[53] M. Madsen and S. Gregor, "Measuring human-computer trust," in *Proc. 11th Australas. Conf. Inf. Syst.*, 2000, pp. 6–8.

[54] I. B. Ajenaghughrure, S. C. Sousa, I. J. Kosunen, and D. Lamas, "Predictive model to assess user trust: A psycho-physiological approach," in *Proc. 10th Indian Conf. Hum.-Comput. Interact.*, 2019, pp. 1–10.

[55] H. M. Khalid *et al.*, "Exploring psycho-physiological correlates to trust: Implications for human-robot-human interaction," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2016, pp. 697–701.

[56] K. Drnec, A. R. Marathe, J. R. Lukos, and J. S. Metcalfe, "From trust in automation to decision neuroscience: Applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction," *Front. Hum. Neurosci.*, vol. 10, 2016, Art. no. 290.

[57] K. Goodyear *et al.*, "An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents," *Social Neurosci.*, vol. 12, no. 5, pp. 570–581, 2017.

[58] S. Palmer, D. Richards, G. Shelton-Rayner, D. Inch, and K. Izzetoglu, "Human-agent teaming-an evolving interaction paradigm: An innovative measure of trust," in *Proc. 93rd Int. Symp. Aviation Psychol.*, 2019, Art. no. 438.

[59] K. Akash, W.-L. Hu, N. Jain, and T. Reid, "A classification model for sensing human trust in machines using EEG and GSR," *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 4, pp. 1–20, 2018.

[60] S. Shahrdar, C. Park, and M. Nojoumian, "Human trust measurement using an immersive virtual reality autonomous vehicle simulator," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2019, pp. 515–520.

[61] M. Wang, A. Hussein, R. F. Rojas, K. Shafi, and H. A. Abbass, "EEG-based neural correlates of trust in human-autonomy interaction," in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2018, pp. 350–357.

[62] C. Park, S. Shahrdar, and M. Nojoumian, "EEG-based classification of emotional state using an autonomous vehicle simulator," in *Proc. IEEE 10th Sensor Array Multichannel Signal Process. Workshop*, 2018, pp. 297–300.

[63] F. Chao, X. Yao, X. Yang, L. Zheng, J. Li, and Y.-W. Wang, "The trust game database: Behavioral and EEG data from two trust games," *Front. Psychol.*, vol. 10, 2019, Art. no. 2656.

[64] K. Gupta, R. Hajika, Y. S. Pai, A. Duenser, M. Lochner, and M. Billinghurst, "Measuring human trust in a virtual assistant using physiological sensing in virtual reality," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces*, 2020, pp. 756–765.

[65] S.-Y. Dong, B.-K. Kim, K. Lee, and S.-Y. Lee, "A preliminary study on human trust measurements by EEG for human-machine interactions," in *Proc. 3rd Int. Conf. Hum.-Agent Interact.*, 2015, pp. 265–268.

[66] K. Akash, G. McMahon, T. Reid, and N. Jain, "Human trust-based feedback control: Dynamically varying automation transparency to optimize human-machine interactions," *IEEE Control Syst. Mag.*, vol. 40, no. 6, pp. 98–116, Dec. 2020.

[67] D. Novak, "Engineering issues in physiological computing," in *Advances in Physiological Computing*, S. H. Fairclough and K. Gilleade, Eds. New York, NY, USA: Springer, 2014, pp. 17–38.

[68] J. Cacioppo and L. Tassinary, "Inferring psychological significance from physiological signals," *Amer. Psychol.*, vol. 45, no. 1, pp. 16–28, Jan. 1990.

[69] J. M. Kivikangas *et al.*, "A review of the use of psychophysiological methods in game research," *J. Gaming Virtual Worlds*, vol. 3, no. 3, pp. 181–199, Sep. 2011.

[70] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. van Moorsel, "The relationship between trust in AI and trustworthy machine learning technologies," in *Proc. Conf. Fairness, Accountability, Transparency*, 2020, pp. 272–283.

[71] C. S. Wickramasinghe, D. L. Marino, J. Grandio, and M. Manic, "Trustworthy AI development guidelines for human system interaction," in *Proc. 13th Int. Conf. Hum. Syst. Interact.*, 2020, pp. 130–136.

[72] R. V. Yampolskiy, "Taxonomy of pathways to dangerous artificial intelligence," in *Proc. Workshops 30th AAAI Conf. Artif. Intell.*, 2016.

[73] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1625–1634.

[74] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Mach. Intell.*, vol. 1, no. 9, pp. 389–399, 2019.

[75] N. A. Smuha, "The EU approach to ethics guidelines for trustworthy artificial intelligence," *Comput. Law Rev. Int.*, vol. 20, no. 4, pp. 97–106, 2019.

[76] "Deloitte," Deloitte introduces trustworthy AI framework to guide organisations in ethical application of technology in the age of with, Aug. 26, 2020. [Online]. Available: https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-introduces-trustworthy-ai-framework.html

[77] J. Fjeld, H. Hilligoss, N. Achten, M. L. Daniel, S. Kagay, and J. Feldman, "Principled artificial intelligence: Mapping consensus and divergence in ethical and rights-based approaches," Project of the Berkman Klein Center for Internet and Society at Harvard University, Boston, MA, USA, vol. 13, p. 2019, 2019, Data Visualization retrieved September.

[78] S. Katzenbeisser, I. Polian, F. Regazzoni, and M. Stöttinger, "Security in autonomous systems," in *Proc. IEEE Eur. Test Symp.*, 2019, pp. 1–8.

[79] R. Changalvala and H. Malik, "LIDAR data integrity verification for autonomous vehicle," *IEEE Access*, vol. 7, pp. 138018–138031, 2019.

[80] R. Changalvala, B. Fedoruk, and H. Malik, "Radar data integrity verification using 2D-QIM based data hiding," *Sensors*, vol. 20, no. 19, 2020, Art. no. 5530.

[81] N. Patel, A. N. Saridena, A. Choromanska, P. Krishnamurthy, and F. Khorrami, "Adversarial learning-based on-line anomaly monitoring for assured autonomy," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 6149–6154.

[82] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood, "Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation," in *Proc. IEEE Conf. Comput. Intell. Games*, 2018, pp. 1–8.

[83] A. B. Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[84] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[85] F. Costa, S. Ouyang, P. Dolog, and A. Lawlor, "Automatic generation of natural language explanations," in *Proc. 23rd Int. Conf. Intell. User Interfaces Companion*, 2018, pp. 1–2.

[86] J. R. Zilke, E. L. Mencía, and F. Janssen, "DeepRred Rule extraction from deep neural networks," in *Proc. Int. Conf. Discov. Sci.*, 2016, pp. 457–473.

[87] M. Nassar, K. Salah, M. H. ur Rehman, and D. Svetinovic, "Blockchain for explainable and trustworthy artificial intelligence," *Wiley Interdiscipl. Rev.: Data Mining Knowl. Discov.*, vol. 10, no. 1, 2020, Art. no. e1340.

[88] R. L. Rivest, "Learning decision lists," *Mach. Learn.*, vol. 2, no. 3, pp. 229–246, 1987.

[89] F. K. Došilovic´, M. Brčic´, and N. Hlupič, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron.*, 2018, pp. 0210–0215.

[90] J. M. Wing *et al.*, "Trustworthy AI," *Commun. ACM*, vol. 64, no. 10, pp. 64–71, 2021.

[91] S. A. Seshia, D. Sadigh, and S. S. Sastry, "Towards verified artificial intelligence," 2016, *arXiv:1606.08514*.

[92] D. J. Fremont *et al.*, "Scenic: A language for scenario specification and scene generation," in *Proc. 40th ACM SIGPLAN Conf. Program. Lang. Design Implementation*, 2016, pp. 63–78.

[93] D. Sadigh *et al.*, "Data-driven probabilistic modeling and verification of human driver behavior," in *Proc. AAAI Spring Symp.-Tech. Rep.*, 2014, pp. 56–61.

[94] D. Sadigh, S. S. Sastry, S. A. Seshia, and A. Dragan, "Information gathering actions over human internal state," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 66–73.

[95] T. Dreossi, A. Donze´, and S. A. Seshia, "Compositional falsification of cyber-physical systems with machine learning components," *J. Automated Reasoning*, vol. 63, no. 4, pp. 1031–1053, 2019.

[96] I. Akkaya, D. J. Fremont, R. Valle, A. Donze´, E. A. Lee, and S. A. Seshia, "Control improvisation with probabilistic temporal specifications," in *Proc. IEEE 1st Int. Conf. Internet-of- Things Des. Implementation*, 2016, pp. 187–198.

[97] S. Berezin, S. Campos, and E. M. Clarke, "Compositional reasoning in model checking," in *Proc. Int. Symp. Compositionality*, 1997, pp. 81–102.

[98] S. Jha and S. A. Seshia, "A theory of formal synthesis via inductive learning," *Acta Informatica*, vol. 54, no. 7, pp. 693–726, 2017.

[99] R. Lee, M. J. Kochenderfer, O. J. Mengshoel, G. P. Brat, and M. P. Owen, "Adaptive stress testing of airborne collision avoidance systems," in *Proc. IEEE/AIAA 34th Digit. Avionics Syst. Conf.*, 2015, pp. 6C2–6C1.

[100] M. Althoff and J. M. Dolan, "Online verification of automated road vehicles using reachability analysis," *IEEE Trans. Robot.*, vol. 30, no. 4, pp. 903–918, Aug. 2014.

[101] M. Althoff, D. Althoff, D. Wollherr, and M. Buss, "Safety verification of autonomous vehicles for coordinated evasive maneuvers," in *Proc. IEEE Intell. Veh. Symp.*, 2010, pp. 1078–1083.

[102] M. Althoff and J. M. Dolan, "Set-based computation of vehicle behaviors for the online verification of autonomous vehicles," in *Proc. 14th Int. IEEE Conf. Intell. Transp. Syst.*, 2011, pp. 1162–1167.

[103] D. M. Lopez, P. Musau, N. Hamilton, H.-D. Tran, and T. T. Jonhson, "Case study: Safety verification of an unmanned underwater vehicle," in *Proc. IEEE Secur. Privacy Workshops*, 2020, pp. 189–195.

[104] O. Bastani, X. Zhang, and A. Solar-Lezama, "Probabilistic verification of fairness properties via concentration," in *Proc. ACM Program. Lang.*, 2019, pp. 1–27.

[105] T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev, "AI2: Safety and robustness certification of neural networks with abstract interpretation," in *Proc. IEEE Symp. Secur. Privacy*, 2018, pp. 3–18.

[106] S. A. Seshia, S. Jha, and T. Dreossi, "Semantic adversarial deep learning," *IEEE Des. Test*, vol. 37, no. 2, pp. 8–18, 2020.

[107] A. Albarghouthi, L. D'Antoni, S. Drews, and A. V. Nori, "FairSquare: Probabilistic verification of program fairness," in *Proc. ACM Program. Lang.*, 2017, pp. 1–30.

[108] M. L. Bolton, E. J. Bass, and R. I. Siminiceanu, "Using formal verification to evaluate human-automation interaction: A review," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 3, pp. 488–503, May 2013.

[109] B. Weyers, J. Bowen, A. Dix, and P. Palanque, *The Handbook of Formal Methods in Human-Computer Interaction*. Berlin, Germany: Springer, 2017.

[110] H. Kress-Gazit *et al.*, "Formalizing and guaranteeing human-robot interaction," *Commun. ACM*, vol. 64, no. 9, pp. 78–84, 2021.

[111] M. van Paassen, M. L. Bolton, and N. Jime´nez, "Checking formal verification models for human-automation interaction," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2014, pp. 3709–3714.

[112] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *Int. J. Hum.-Comput. Stud.*, vol. 40, no. 1, pp. 153–184, 1994.

[113] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man-Mach. Stud.*, vol. 27, no. 5/6, pp. 527–539, 1987.

[114] M. S. Cohen, R. Parasuraman, and J. T. Freeman, "Trust in decision aids: A model and its training implications," in *Proc. Command Control Res. Technol. Symp.*, 1998.

[115] S. H. Huang, K. Bhatia, P. Abbeel, and A. D. Dragan, "Establishing appropriate trust via critical states," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 3929–3936.

[116] B. W. Israelsen, "Algorithmic assurances and self-assessment of competency boundaries in autonomous systems," Ph.D. dissertation, Faculty Grad. School, Dept. Comput. Sci., Univ. Colorado Boulder, Golden, CO, USA, 2019.

[117] B. W. Israelsen, N. R. Ahmed, E. Frew, D. Lawrence, and B. Argrow, "Algorithmic assurances and self-assessment of competency boundaries in autonomous systems," Ph.D. dissertation, Univ. Colorado Boulder, Boulder, CO, USA, 2019.

[118] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *Proc. 11th ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2016, pp. 101–108.

[119] T. Rezvani, K. Driggs-Campbell, D. Sadigh, S. S. Sastry, S. A. Seshia, and R. Bajcsy, "Towards trustworthy automation: User interfaces that convey internal and external awareness," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst.*, 2016, pp. 682–688.

[120] W. Kim, N. Kim, J. Lyons, and C. Nam, "Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach," *Appl. Ergonom.*, vol. 85, May 2020, Art. no. 103056.

[121] R. Grech *et al.*, "Review on solving the inverse problem in EEG source analysis," *J. Neuroeng. Rehabil.*, vol. 5, no. 1, 2008, Art. no. 25.

[122] K. Akash, T. Reid, and N. Jain, "Adaptive probabilistic classification of dynamic processes: A case study on human trust in automation," in *Proc. Annu. Amer. Control Conf.*, 2018, pp. 246–251.

[123] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerg. Artif. Intell. Appl. Comput. Eng.*, vol. 160, no. 1, pp. 3–24, 2007.

[124] J. E. Mercado, M. A. Rupp, J. Y. Chen, M. J. Barnes, D. Barber, and K. Procci, "Intelligent agent transparency in human–agent teaming for multi-UXV management," *Hum. Factors*, vol. 58, no. 3, pp. 401–415, 2016.

[125] K. Okamura and S. Yamada, "Adaptive trust calibration for human-AI collaboration," *PLoS One*, vol. 15, no. 2, 2020, Art. no. e0229132.

[126] B. Seppelt, "Supporting operator reliance on automation through continuous feedback," Ph.D. dissertation, Graduate College, Univ. Iowa, Iowa City, IA, USA, Dec. 2009.

[127] A. Mackay *et al.*, "The impact of autonomous vehicles' active feedback on trust," in *Proc. Int. Conf. Appl. Hum. Factors Ergonom.*, 2019, pp. 342–352.

[128] P. Wintersberger, H. Nicklas, T. Martlbauer, S. Hammer, and A. Riener, "Explainable automation: Personalized and adaptive UIS to foster trust and understanding of driving automation systems," in *Proc. 12th Int. Conf. Automot. User Interfaces Interact. Veh. Appl.*, 2020, pp. 252–261.

[129] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, "The role and limits of principles in AI ethics: Towards a focus on tensions," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2019, pp. 195–200.

[130] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 4, pp. 1–31, 2020.

[131] S. Hammer, M. Wißner, and E. Andre´, "Trust-based decision-making for smart and adaptive environments," *User Model. User-Adapted Inter.*, vol. 25, no. 3, pp. 267–293, 2015.