Grounding Trust in Human-Agent Interaction: A Systematic Literature Review of Data-driven Trust Assessment Methods

Chih-Wei (Charlotte) Ning
Intelligent Systems
Delft University of Techonology
Delft, Netherlands
C.W.Ning@tudelft.nl

Bernd Dudzik
Intelligent Systems/Pattern Recognition & Bioinformatics
Delft University of Technology
Delft, Netherlands
b.j.w.dudzik@tudelft.nl

Mark A. Neerincx
Intelligent Systems
Delft University of Technology
Delft, Netherlands
m.a.neerincx@tudelft.nl

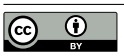
Myrthe L. Tielman
Interactive Intelligence
Delft University of Technology
Delft, Netherlands
m.l.tielman@tudelft.nl

Abstract

Trust plays an important role in human-agent interaction and has motivated researchers to develop various measurement approaches. Data-driven methods have been recognized as a promising direction by leveraging observable signals produced by the user during interaction. In this paper, we present an ongoing systematic literature review of data-driven trust measurement studies, focusing on how the abstract concept of trust is grounded throughout the assessment pipeline. We synthesize existing approaches based on input modalities and problem framing, and further discuss how two key characteristics of trust, temporal dynamics and socialness, are reflected in current operationalizations. Our initial analysis covers 44 more recent studies published between 2022 and 2026. The reviewed studies predominantly rely on psychophysiological signals, embedded indicators, and gaze-related metrics as input modalities. Although relational aspects of trust receive less attention, some studies demonstrate the potential of incorporating social signals to assess trust in more socially grounded interaction settings. Finally, most studies incorporate temporal information through windowing formulations, which rely on aggregated representations that may obscure fine-grained trust dynamics. Overall, this work provides an initial step toward systematically understanding how human-agent trust is operationalized in data-driven assessment methods. We will extend this ongoing review into a full paper by incorporating the complete corpus and further developing a systematic characterization of socialness in human-agent interaction.

CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models**; **User models**; • **Computing methodologies** → *Model development and analysis.*



This work is licensed under a Creative Commons Attribution 4.0 International License.
ICMI Companion '26, Napoli, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2319-3/26/10
https://doi.org/10.1145/3776591.3836637

Keywords

Human-Agent Trust, Trust, Affective Computing, Literature Review, Measurement

ACM Reference Format:
Chih-Wei (Charlotte) Ning, Bernd Dudzik, Mark A. Neerincx, and Myrthe L. Tielman. 2026. Grounding Trust in Human-Agent Interaction: A Systematic Literature Review of Data-driven Trust Assessment Methods. In *Companion of the INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI Companion '26), October 05–09, 2026, Napoli, Italy.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3776591.3836637

1 Introduction

Trust is a key component in human-agent interaction. When users maintain an appropriate level of trust, they can fully benefit from a system's assistance without either disusing it through under-trust or misusing it through over-trust [42, 68]. However, trust is a subjective attitude of the user and does not always match the system's actual, objective trustworthiness. When such misalignment occurs, trust becomes inappropriate and should be calibrated toward a more appropriate level [49, 68, 72], as illustrated in Figure 1 [68]. This need for trust calibration has motivated research on unobtrusive and real-time methods for trust measurement (e.g., [7, 24]).

The need for real-time trust assessment is closely tied to the view that trust evolves dynamically throughout an interaction [24, 49, 61]. This perspective is gaining increasing attention [49, 61], with researchers arguing that traditional questionnaire-based measures capture only a static snapshot of trust at a particular moment rather than its continuous development [24, 61]. Data-driven approaches address this limitation by leveraging behavioral signals that are naturally produced during interaction, offering the potential to capture temporal information.

Another challenge in measuring trust lies in its abstract and multifaceted nature. In human-human trust research, relational components are consistently emphasized alongside ability-based judgments [47, 48]. Similarly, in human-automation literature, trust is shaped not only by cognitive evaluations but also by affective responses [38, 50]. Taken together, human-agent trust can be seen as sitting in the intersection between human-human and human-automation trust. It is thus fundamentally a multidimensional construct [46, 53], especially as agents become increasingly socially

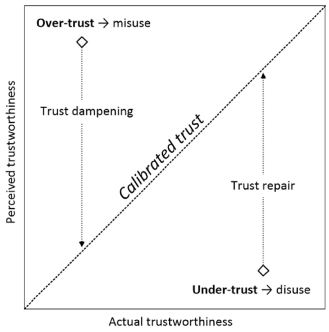


Figure 1: The concept of trust calibration. Figure from [68].

framed. This multidimensionality is particularly relevant for data-driven approaches, where rich multimodal behavioral signals hold the potential to capture different facets of trust [36].

While data-driven methods have been recognized as a promising direction in prior literature surveys [13, 36], existing reviews have not systematically examined how these approaches operationalize trust. Ajenaghughrure et al. [3] focus on psychophysiological signals but do not cover other behavioral modalities, such as speech and facial expressions, which have also been widely used in related areas like emotion recognition [33]. This work addresses this gap by reviewing data-driven trust assessment methods across a broader range of input modalities. In particular, we focus on how prior studies ground the abstract, multifaceted concept of trust in concrete, computational operationalization within different stages of data-driven pipelines.

In the remainder of this paper, we first provide background on broader human-agent trust research and outline the conceptual scope of this survey (Section 2). We then describe the methodology of our ongoing systematic literature review (Section 3), followed by a summary of how the reviewed studies operationalize trust in terms of input signals and problem framing (Section 4). In Section 5, we discuss how two key characteristics of human-agent trust, *temporality* and *socialness*, are reflected in current methodological choices. Finally, we conclude the paper and outline future research directions in Section 6.

2 Background

Trust has been defined in various ways. In this paper, we follow the definition of Lee and See [38] and conceptualizes trust as a positive *internal attitude* where the trustor feels positive about relying on the trustee, despite being in a situation characterized by uncertainty and vulnerability. The trustor is the one who trusts (in our context, the human user), while trustee is the one who is being trusted (in our context, the system or agent). In this section, we first review mainstream approaches of human-agent trust measurements. We then explain and define the scope of data-driven methods.

2.1 Measuring Human-Agent Trust

Researchers have proposed diverse approaches to assess users' trust in systems during interaction. Existing literature reviews provide comprehensive overviews of these methods [3, 13, 36, 58, 72]. Here, we briefly summarize three major categories.

The most common approach uses *questionnaires* to directly measure users' subjective attitudes [13, 36, 58]. Widely used examples include [31, 45, 63]. These instruments are theoretically grounded and psychometrically validated, enabling them to capture the multifaceted nature of trust. Nevertheless, administrating a survey can be time-consuming and intrusive to the ongoing interaction, thus is not ideal for capturing moment-to-moment trust dynamics.

Another line of work measures trust through *embedded behavioral indicators*, such as whether users delegate tasks to the agent or follow its suggestions [13, 36]. Despite being less intrusive, these measures are often tightly coupled to specific interaction contexts, limiting their generalizability. Moreover, demonstrated trust does not always align with perceived trust [1, 10, 15]. We therefore view such indicators as behavioral proxies that may reflect the underlying trust attitude to some degree, rather than a direct measurement of trust itself.

Finally, *data-driven* approaches have gained increasing attention [13, 36]. Rather than reducing trust to isolated questionnaire responses or discrete actions, these approaches aim to infer trust through heterogeneous observable evidence and are empowered by modern machine learning and deep learning techniques. Existing work, however, mainly focus on psychophysiological signals, such as heart rate, skin conduct, neural activities, etc. To complement prior review (e.g., [3]), we adopt a broader perspective that includes speech, facial expressions, and other socially expressive behaviors.

2.2 Behaviors as Manifestations of Attitudes

As illustrated in Figure 2, we conceptualize trust as a latent attitude of the trustor toward the trustee. This attitude is shaped by various factors and in turn manifests through observable signals. The arrows in the figure represent these assumed causal relationships.

Extensive human-agent trust research has focused on the left-hand side of this causal process: manipulating factors such as the trustor's personality, the trustee's characteristics (e.g., anthropomorphism), or the interaction context (e.g., risk), and examining their effects on trust [38, 72]. In contrast, this review focuses on the right-hand side: how an already-formed trust attitude manifests itself through observable multimodal signals during interaction. For examples, whether the attitude is shown on the users' faces or is even directly expressed in the conversation.

Following this conceptualization, the embedded behavioral indicators discussed in the previous subsection (e.g., accepting or rejecting an agent's recommendation) are interpreted as a context-specific manifestation of the underlying trust attitude, rather than as trust itself. They, therefore, represent one type of observable evidence from which trust may be inferred, alongside other behavioral and physiological signals.

This perspective aligns with the affective computing literature, where latent cognitive and affective states, such as emotions and stress, are commonly inferred from multimodal behavioral observations (e.g., [19, 33, 37]). It is also consistent with research in social computing, which seeks to understand social constructs from multimodal social behaviors (e.g., [28, 30]). As trust is inherently a relational construct involving both the trustor and the trustee, insights from these fields are particularly relevant.

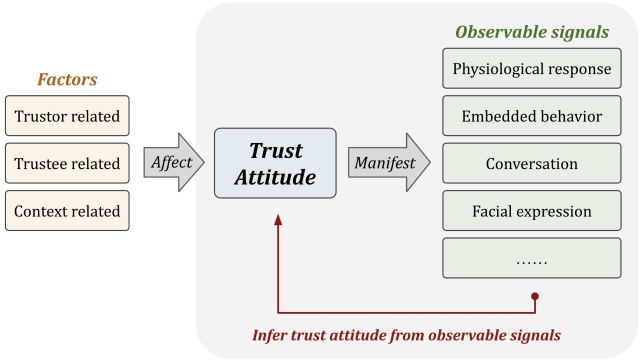


Figure 2: Conceptual scope of the present literature review. We see trust as an latent attitude, affected by various factors and manifest as various channels of observable signals. We review studies that infer this latent attitude from such observable signals. Arrows represent causal relationships.

To conclude, we view observable signals produced during interaction as potential *manifestations* of the trustor's underlying trust attitude. Accordingly, we scope this literature review to *data-driven* trust assessment studies that infer latent trust attitudes from such observable signals.

3 Review Method and Current Progress

We structured the systematic literature review according to the SALSA framework [22], consisting of search, appraisal, synthesis, and analysis.

3.1 Search

The literature search was conducted on Elsevier's Scopus database in April 2026. The initial search query was constructed based on four conceptual components of the review scope:

(1) **Trust**: trust
(2) **Human-agent interaction**: robot, agent, machine, automation, AI
(3) **Observable signals**: multimodal*, physiological*, behavior*, speech, acoustic, verbal, gaze, gesture, video, face, audio, linguistic
(4) **Assessment**: measure*, assess*, model*, predict*, infer*

Within each component, keywords were combined using *OR*, while different components were combined using *AND*. The term *trust* was restricted to the *TITLE* field, whereas all other components were searched within *TITLE–ABS–KEY*. Additional filtering criteria restricted the results to English peer-reviewed publications published between 2011 to 2026, and excluded irrelevant domains. The complete query string can be found in Appendix A.

3.2 Appraisal

The literature selection procedure is summarized in Figure 3. Using the search query described above, 1315 records were initially identified, followed by title, abstract, and full-text screening. The process was conducted by the first author. At the time of writing, full-text screening has been completed for 70 papers spanning the most recent five years (published 2022 – 2026), while the remaining 41 papers (published 2011 – 2021) are still under review. At each stage, papers were excluded if they did not satisfy any of the following criteria:

- The paper must investigate the trust of a human toward an agent/robot/machine during the one-on-one interaction of the two parties.
- The paper must intend to measure or infer such trust.
- The paper must utilize observable behavior data during the interaction to infer trust.

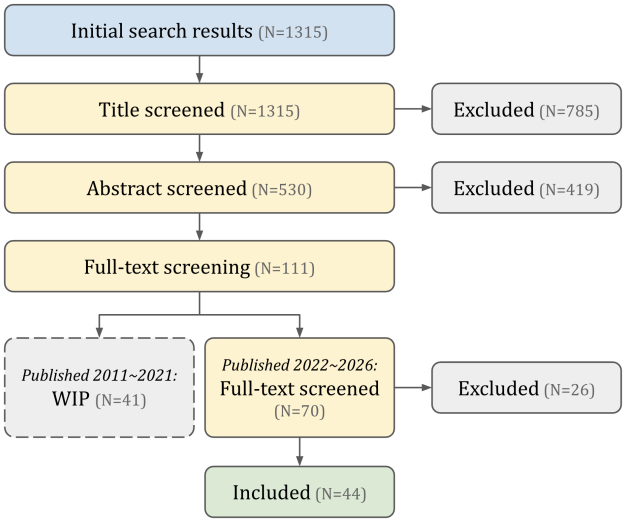


Figure 3: Systematic literature review procedure. After title and abstract screening, 111 documents remained for full-text review. The present paper reports an initial analysis of the 44 papers published within the most recent five years.

The preliminary observations presented in this workshop paper are based on the 44 currently reviewed studies and are intended to initiate discussion around data-driven trust assessment. Meanwhile, the full literature review is still ongoing, and the remaining corpus will be incorporated into an extended full paper.

3.3 Synthesis and Analysis

Following the screening process, the appraised studies were systematically reviewed. Data extraction and coding were also performed by the first author. How each study operationalizes trust were extracted from two perspectives within the data-driven pipeline:

(1) **Modalities**: We coded the types of observable signals used as inputs to data-driven trust assessment pipelines. We also considered how multiple modalities were combined, when applicable.
(2) **Problem framing**: On the other side of data-driven pipelines, we examined the intended outputs and how trust was defined as the prediction target or ground truth.

Together, these two dimensions allowed us to characterize how existing studies ground the abstract concept of trust in concrete

multimodal interactions. They form the basis of our descriptive synthesis, which is presented in Section 4. Furthermore, recurring patterns identified during the synthesis motivated a conceptual analysis of how two fundamental characteristics of trust, temporality and socialness, are incorporated into current operationalizations throughout the pipelines. These initial observations are discussed in Section 5.

4 Preliminary Results

4.1 Modalities

First of all, we look at the data modalities being used to infer or measure trust. As our interest lies within social signals, and as complementary to previous review on psychophysiological and gaze data [3], this paper will primarily focus on the other modalities.

4.1.1 Type of modalities. As shown in Figure 4(a), **psychophysiological** signals remain the most commonly used modality for trust assessment, consistent with prior reviews [3, 13]. **Gaze**-related measures and **embedded** behavioral indicators are the second and third popular ones. In contrast, only 13 studies explored social signals beyond these modalities, as summarized in Table 1.

Speech has been investigated from both linguistic and acoustic perspectives. **Linguistic** features (e.g., *what* the user says) provide direct access to users' verbalized attitudes and reasoning processes, while **acoustic** features (e.g., *how* the user says it) capture nonverbal vocal cues associated with affective and cognitive states. Some recent studies leveraged large language models to infer trust from natural language responses [35, 69], while some others utilized acoustic features to estimate trust in domains such as child-robot interaction[67] and automated driving [16]. Li et al. further combined linguistic and acoustic information in the interactions with conversational agents, reporting improved trust prediction performance compared to either modality alone [40, 41].

Facial expressions constitute another naturally occurring social signal that reveal users' affective reactions dynamically during interactions and have thus been widely examined in emotion recognition literature [11, 60]. Following this line, some studies have incorporated facial features alongside other modalities, demonstrating their potential for real-time trust estimation [6, 9, 23].

Finally, a small number of studies have explored **body posture** and movement patterns as indicators of trust, capturing behavioral manifestations such as hesitation or avoidance under distrust. These studies typically consider embodied interaction contexts, including human–robot collaboration in shared spaces [12, 59], automated driving [43], and evacuation scenarios in virtual reality [64]. However, the resulting findings are often highly context-specific and may not generalize beyond the studied setting. For instance, the associations between lower-body movements (e.g., foot movements toward the brake pedal) and trust in automated driving [43].

4.1.2 Combination of modalities. In this subsection, we examine how different modalities are combined. As illustrated in Figure 4(b), about half of the reviewed papers ($N = 22$ out of 44) investigate more than one modality as input. Several studies explicitly compare unimodal and multimodal settings and report superior performance when multiple modalities are combined [6, 29, 32, 40, 41, 75]. Interestingly, some studies further discuss how ensemble-based machine learning models, such as Random Forest and XGBoost, are particularly well suited for multimodal fusion, as they can effectively integrate heterogeneous features extracted from different signal sources [2, 9, 14, 32]. Finally, several works contribute to this research direction by releasing multimodal datasets for trust assessment [6, 9, 55].

Paper	Lin.	Aco.	Fac.	Body	Others
Wang et al. (2025) [69]	✓				Embedded
Koehl et al. (2023) [35]	✓				
Li et al. (2024) [41]	✓	✓			
Li et al. (2022) [40]	✓	✓			
Velner et al. (2024) [67]		✓			
Deng et al. (2024) [16]		✓			
Alzahrani et al. (2025) [6]		✓	✓		
Green et al. (2025) [23]			✓		Psycho.
Bhalla et al. (2025) [9]			✓		Psycho. & Gaze
Campagna et al. (2024) [12]				✓	
Rehm et al. (2024) [59]				✓	
Lim et al. (2024) [43]				✓	
Tong et al. (2024) [64]				✓	
Count	4	5	3	4	

Table 1: Studies incorporating social signals in trust assessment. The abbreviations denote linguistic, acoustic, facial expression, body posture or movement, embedded behavioral indicators, psychophysiological signals, and gaze-related metrics.

4.2 Problem Framing

Reviewed studies have framed trust measurement in several different ways, as shown in Figure 4(c). The most common formulation treats trust as a continuous variable and models it as a **regression** problem ($N = 18$). In these studies, self-report trust ratings are typically used as the ground truth, while trust-relevant factors (e.g., system performance) are often manipulated to induce variation in trust levels. A wide range of regression methods have been applied, spanning from statistical models (e.g., [8, 62]), machine learning approaches (e.g., [14, 41]), and deep learning architectures (e.g., [32, 43]).

Classification is another common problem framing ($N = 14$). Similar to the trust measurement literature (see Section 2.1), the target classes are typically derived from either embedded behavioral indicators or self-report questionnaires. The former includes decisions such as taking over [5, 21] or following agent's suggestions [2, 6, 65]. Self-report trust ratings have been converted to target classes in several ways. Some directly treat rating levels as class labels [9, 39], whereas others divide participants into high- and low-trust groups using either median splits, balanced sampling, or other threshold-based procedures [7, 16, 55, 59]. While these discretization strategies facilitate classification, the resulting class boundaries are often somewhat arbitrary and may not correspond to theoretically meaningful distinctions in trust. Interestingly, Yi et al. [75, 76] derived trust categories by clustering both subjective trust ratings and objective behavioral measures, providing a more integrated operationalization of trust by jointly considering perceived and demonstrated trust.

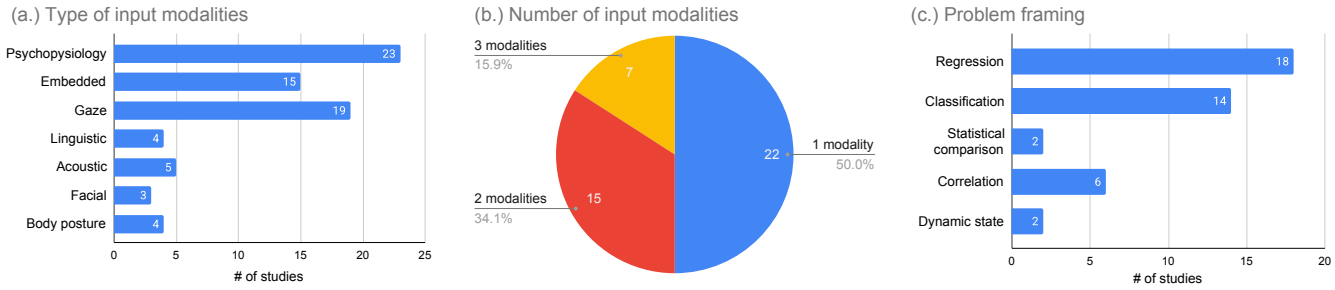


Figure 4: The descriptive statistics of the reviewed studies, including (a.) types and (b.) number of input modalities, as well as (c.) problem framing. Note that a single study may involve multiple input modality types and problem formulations.

Two studies do not formulate trust assessment as a prediction task. Instead, they use **statistical comparisons** such as ANOVA to examine whether behavioral signals differ across trust conditions or self-reported trust levels [23, 77]. Although these approaches do not fully exploit the predictive capabilities of data-driven methods, they provide useful preliminary evidence regarding the sensitivity of such signals to variations in trust.

Another line focuses on exploratory **correlation analysis** (N = 6). These works investigate the association between trust ratings and behavioral signals, particularly gaze-related [4, 70] and neurophysiological signals such as EEG [65, 74] and fNIRS [26, 27]. Such studies often aimed to identify potential trust-related biomarkers before developing predictive models.

Finally, both [71] and [44] explicitly model trust as a **dynamic state** that evolves over time, using a partially observable Markov decision process without reward functions (POMDP/R) and reinforcement learning, respectively. This line of work builds upon earlier agent-based computational trust models that infer trust from system performance (e.g., [24, 73]), while additionally incorporating users' behavioral signals as observations. As such, these approaches fall within the scope of the present review.

5 Discussions

In this section, we present emerging patterns identified in the reviewed literature and discuss the implicit assumptions underlying the operationalization of data-driven trust assessment methods. Specifically, we focus on two concepts that have become increasingly salient in recent trust research: (1) temporal dynamics (i.e., how trust develops and changes throughout interactions) and (2) social grounding (i.e., the inherently relational nature of trust between the trustor and the trustee). These two perspectives cut across the entire trust measurement pipeline, from input signals to the formulation of trust labels and predictive objectives.

5.1 Incorporating temporal information

Trust is widely characterized as a dynamic construct that evolves throughout an interaction rather than a static snapshot at a single point in time [24, 42, 49]. This perspective underlies much of the trust calibration literature and motivates the development of unobtrusive trust assessment methods. Given that behavioral signals are continuously generated during interaction, data-driven approaches are, in principle, well positioned to capture temporal information and estimate trust dynamics. For example, continuous acoustic signals may reveal subtle changes in users' attitudes throughout an interaction that would not be reflected in a single post-interaction trust rating. However, it remains unclear to what extent existing studies exploit the temporal richness inherent in behavioral signals. This subsection presents preliminary observations in this regard.

Temporal information can be incorporated at both the input and output stages. On the input side, several studies explicitly model temporal dependencies using sequence-based architectures such as LSTMs and Transformers (e.g., [43, 75]). Temporal information can also be encoded through feature engineering. For example, studies using neurophysiological or acoustic signals often extract frequency-domain features [41, 56, 65]. While such features capture recurring temporal patterns, they are less suited for representing fine-grained changes in trust over time.

On the output side, dynamic trust labels are most commonly annotated through repeated trust questionnaires presented during interaction, typically at fixed intervals (e.g., every 25–30 seconds [7, 32]). Some studies further increase the temporal granularity by allowing participants to update their ratings between prompts [32]. Nevertheless, many studies still rely on a single trust rating collected at the end of each trial (e.g., [34, 62, 64, 70]). Even when paired with multiple behavioral segments, such ratings primarily reflect the final trust outcome rather than the trajectory through which trust evolved.

Taken together, most reviewed studies ultimately adopt a *windowing* formulation. Since trust labels are typically derived from self-reports or discrete task decisions, the temporal granularity of the output annotations is constrained by the frequency at which these measurements are collected. In contrast, behavioral signals are often recorded continuously at a much finer temporal resolution. As a result, even the most "dynamic" trust annotations remain aggregated representations of trust over a period of time and is ultimately more coarse-grained than the input signals.

A small number of studies have explicitly examined the effect of temporal granularity by comparing different window sizes for trust prediction [21, 54]. Cedric et al. found that longer temporal windows better predict subjective trust ratings collected at the end of an interaction, whereas shorter windows better predict individual takeover events during the interaction [21]. Niu et al. further

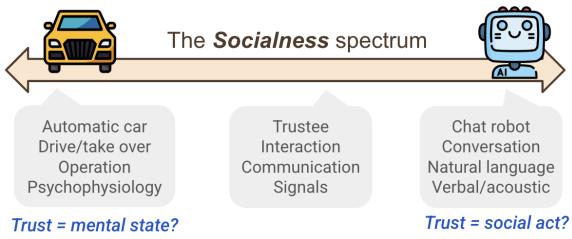


Figure 5: The "socialness" spectrum in trust research settings.

proposed the Selective Windowing Attention Network (SWAN), which automatically optimizes the window size for trust prediction [54]. Future research would benefit from more systematic investigations into the relationship between signal resolution, aggregation windows, and annotation frequency, thereby improving our understanding of how trust dynamics manifest across different temporal scales.

Finally, some studies move beyond window-based approaches and focus directly on what motivates the dynamic perspective of trust: trust change itself. As discussed in Section 4.2, state-transition models such as POMDPs provide a plausible framework in this regard [44, 71]. Other studies incorporate Bayesian assumptions to capture temporal dependencies [21, 57], while some focus on salient interaction events that may trigger trust changes [25, 76]. These directions may better align with the motivation behind data-driven trust assessment: trust calibration. The goal is not merely to estimate trust unobtrusively and automatically, but to enable systems to adapt or intervene when users' trust becomes inappropriate (revisit Figure 1). From this perspective, it may be more meaningful to investigate what triggers salient changes in trust and how trust evolves in response to such events, rather than maintaining precise estimates of trust at every temporal window.

5.2 The *Socialness* in human-agent trust

Socialness is a recurring theme in trust research and has been discussed in multiple layers. On the theory level, several models grounding in interpersonal trust inherently emphasize relational component in addition to performance or reliance based judgment. For instance, benevolence from the ABI (ability, benevolence, integrality) framework [47] and morality trust from the MDMT (multidimensional measurement of trust) questionnaire [45] both consider the feelings and relationships between trustor and trustee. Beyond theories, lots of studies have examined how socialness on the agent level affect user's trust, in terms of degree of anthropomorphism [52] communication channel, or warmth in communication style [20, 53]. The interaction context further shapes the social nature of trust. For example, a caregiving robot designed to provide emotional support inherently affords richer social interaction than an autonomous driving system. Taken together, we view the *socialness* of a research setting as a cross-cutting characteristic that spans these different levels. It is reflected in the choice of trustee, interaction context, communication channel, and the behavioral signals available for trust assessment. We conceptualize socialness as a spectrum and illustrate this perspective in Figure 5.

The settings of the reviewed data-driven research spread widely on this *socialness spectrum*. On the less social side, there are a big portion of, for example, automatic driving cars (e.g., [7, 16, 74]). These works often use psychophysiological or gaze metrics to assess driver's trust under various traffic conditions or system configurations [7, 32, 39]. These line of works often see trust as an internal state of trustor, and often jointly discussed with other mental state constructs like attention, stress, and fatigue [25, 32, 70].

On the other side of the spectrum, there are studies on conversational agents which inherently communicate with the user via natural language [40, 41] – the most nature channel with which people socially interact with each other. The use of humanoid robots further strengthen the social framing in some works [5, 67]. Given these cues, even if the task itself is not be relationship relevant, users still tend to perceive agents as social actor according to the CASA paradigm [51]. In this case, behaviour signals such as speech content and parody features may communicate trustor's trust toward trustee, more than just a state internally to the trustor. These signals are thus naturally seen as a manifestation of trust attitude and being used for measurement.

Despite the substantial differences in the social nature of the reviewed settings, the operationalization of trust are surprisingly similar across studies, with the lack of attention on the relational component of trust. Almost none of the reviewed works employ multidimensional trust measures that explicitly capture the social aspects. Although some questionnaires contain relevant items, they are ultimately aggregated with performance-based items, making the relational component difficult to interpret in isolation.

As a result, we observe a potential conceptual mismatch. Across the reviewed studies, the trustee, interaction context, communication channel, and behavioral signals often move together along the socialness spectrum as a package. However, the corresponding trust labels rarely reflect this variation and remain predominantly performance-oriented. The broader trust research community has been devoting attention on socially meaningful interaction as AI agents become increasingly social and ubiquitous. Data-driven approaches may align better with this trend by making better use of social signals and adopting multidimensional annotation ground truth.

6 Conclusion and Future Work

In this paper, we present an ongoing systematic literature review on data-driven assessment methods in human-agent trust. We reviewed 44 papers from the past five years and summarized their approaches in terms of input signals and problem framing.

We further examined how two key characteristics of trust research, temporality and socialness, are reflected in the operationalization of current data-driven approaches. For temporality, we find that while windowing formulations are commonly used to incorporate temporal information, they often rely on aggregated representations that obscure fine-grained dynamics of trust evolution. As an alternative, event-based and state-based formulations offer a more explicit modeling of trust as an evolving process. For socialness, current data-driven approaches tend to underemphasize the relational dimensions of trust. However, as social agents become

increasingly ubiquitous, there is growing potential for leveraging richer social signals in trust assessment.

While several interesting patterns have emerged from the current review, its work-in-progress nature inevitably introduces several limitations. First, the current analysis is based only on studies published in the past five years, which may introduce temporal sampling bias and limit the generalizability of the findings. As outlined in Figure 3, we plan to extend this work into a full paper after completing the review of the remaining 41 papers (2011–2021) and examining whether the current observations still hold.

In the extended review, we also aim to ground our preliminary observations on temporality and socialness more rigorously through systematic coding and quantitative synthesis of the complete (i.e., the 44 reviewed and 41 unreviewed ones) corpus. We plan to refine the currently intuitive notion of the *socialness* spectrum and systematically catalog and situate the reviewed interaction context within this space. Furthermore, we intend to distinguish finer-grained aspects of *temporal* representation, such as window size and annotation frequency, and report them quantitatively to better support our observations regarding windowing formulations. Finally, we will investigate event-based and state-based approaches, and establish stronger connections with literature on trust violation and repair (e.g., [17, 18]) as well as agent-based computational models of trust (e.g., [24, 66]), respectively.

Overall, this work takes an initial step toward systematically mapping how the abstract concept of trust is grounded and operationalized in multimodal human-agent interaction. By identifying common methodological choices and their limitations, we aim to support more faithful representation and measurement of trust in future research.

Safe and Responsible Innovation Statement

The motivation for developing unobtrusive trust assessment methods originates from trust calibration, which aims to support safer system use by detecting over-trusting and under-trusting. However, data privacy is an inevitable concern when the methods is data-driven, especially when modalities contain personally identifiable information. These risks should be mitigated through careful ethical review processes and by processing and storing only anonymized or pseudonymized data whenever possible. A further consideration concerns the ethics of inferring human internal attitudes. By utilizing observable signals, trust assessment systems do not necessarily form more intrusive judgments than an equivalent human interaction partner. However, as automated inference can be persistent and scalable, such technologies should only be used with explicit and retractable user consent and transparency of model uncertainty. Moreover, research on trust assessment should clearly communicate that inferred internal states are only for improving interaction quality and safety, rather than labeling or profiling individuals.

Acknowledgments

This publication is part of the project 'Hybrid Intelligence: augmenting human intellect' (https://hybrid-intelligence-centre.nl,) with project number 024.004.022 of the research programme 'Gravitation' which is (partly) financed by the Dutch Research Council (NWO).

References

[1] Mark Adamik, Karolina Dudzinska, Adrian J. Herskind, and Matthias Rehm. 2021. The Difference Between Trust Measurement and Behavior: Investigating the Effect of Personalizing a Robot's Appearance on Trust in HRI. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. 880–885. doi:10.1109/RO-MAN50785.2021.9515487

[2] Muneeb Ahmad and Abdullah Alzahrani. 2023. Crucial Clues: Investigating Psychophysiological Behaviors for Measuring Trust in Human-Robot Interaction. In *Proceedings of the 25th International Conference on Multimodal Interaction* (Paris, France) *(ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 135–143. doi:10.1145/3577190.3614148

[3] Ighoyota Ajenaghughrure, Sonia Sousa, and David Lamas. 2020. Measuring Trust with Psychophysiological Signals: A Systematic Mapping Study of Approaches Used. *Multimodal Technologies and Interaction* 4 (Sept. 2020), 63. doi:10.3390/mti4030063

[4] Anastasia Akkuzu, Natalia Calvo-Barajas, and Ginevra Castellano. 2023. Behavioural Observations as Objective Measures of Trust in Child-Robot Interaction: Mutual Gaze. In *Proceedings of the 11th International Conference on Human-Agent Interaction* (Gothenburg, Sweden) *(HAI '23)*. Association for Computing Machinery, New York, NY, USA, 452–454. doi:10.1145/3623809.3623960

[5] Abdullah Alzahrani and Muneeb Ahmad. 2024. Real-Time Trust Measurement in Human-Robot Interaction: Insights from Physiological Behaviours. In *Proceedings of the 26th International Conference on Multimodal Interaction* (San Jose, Costa Rica) *(ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 627–631. doi:10.1145/3678957.3688620

[6] Abdullah Alzahrani, Jauwairia Nasir, Ahmad J. Tayeb, Elisabeth André, and Muneeb I. Ahmad. 2025. What do the Face and Voice Reveal? Investigating Trust Dynamics During Human-Robot Interaction. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 400–409. doi:10.1109/HRI61500.2025.10974183

[7] Jackie Ayoub, Lilit Avetisian, X. Jessie Yang, and Feng Zhou. 2023. Real-Time Trust Prediction in Conditionally Automated Driving Using Physiological Measures. *IEEE Transactions on Intelligent Transportation Systems* 24, 12 (2023), 14642–14650. doi:10.1109/TITS.2023.3295783

[8] Gregory Bales, Allison P. A. Hayman, Torin K. Clark, Jason Dekarske, Sanjay Joshi, and Zhaodan Kong. 2025. An EEG-network-metric based approach to real-time trust inference in human-autonomy teaming. *Frontiers in Neuroergonomics* Volume 6 - 2025 (2025). doi:10.3389/fnrgo.2025.1627483

[9] Aditi Bhalla, Christian Hellert, Enkelejda Kasneci, and Nastassja Becker. 2025. TRUCE-AV: A Multimodal dataset for Trust and Comfort Estimation in Autonomous Vehicles. arXiv:2508.17880 [cs.HC] https://arxiv.org/abs/2508.17880

[10] J. M. Bindewald, C. F. Rusnock, and M. E. Miller. 2017. Measuring Human Trust Behavior in Human-Machine Teams. *Advances in Intelligent Systems and Computing* (2017), 47–58. doi:10.1007/978-3-319-60591-3_5

[11] Federico Cabitza, Andrea Campagner, and Martina Mattioli. 2022. The unbearable (technical) unreliability of automated facial emotion recognition. *Big Data & Society* 9, 2 (2022), 20539517221129549. arXiv:https://doi.org/10.1177/20539517221129549 doi:10.1177/20539517221129549

[12] Giulio Campagna, Mahed Dadgostar, Dimitrios Chrysostomou, and Matthias Rehm. 2024. A Data-Driven Approach Utilizing Body Motion Data for Trust Evaluation in Industrial Human-Robot Collaboration*. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. 1984–1990. doi:10.1109/RO-MAN60168.2024.10731437

[13] Giulio Campagna and Matthias Rehm. 2025. A Systematic Review of Trust Assessments in Human–Robot Interaction. *J. Hum.-Robot Interact.* 14, 2 (Jan. 2025), 30:1–30:35. doi:10.1145/3706123

[14] Hardik Chauhan, Youjin Jang, and Inbae Jeong. 2024. Predicting human trust in human-robot collaborations using machine learning and psychophysiological responses. *Advanced Engineering Informatics* 62 (2024), 102720. doi:10.1016/j.aei.2024.102720

[15] Jenna E. Cotter, Emily H. O'Hear, R. Cooper Smitherman, Addison B. Bright, Nathan L. Tenhundfeld, Jason Forsyth, Nathan R. Sprague, and Samy El-Tawab. 2022. Convergence Across Behavioral and Self-report Measures Evaluating Individuals' Trust in an Autonomous Golf Cart. In *2022 Systems and Information Engineering Design Symposium (SIEDS)*. 252–257. doi:10.1109/SIEDS55548.2022.9799431

[16] Miaomiao Deng, Jiaqi Chen, Yue Wu, Shu Ma, Hongting Li, Zhen Yang, and Yi Shen. 2024. Using voice recognition to measure trust during interactions with automated vehicles. *Applied Ergonomics* 116 (2024), 104184. doi:10.1016/j.apergo.2023.104184

[17] Connor Esterwood, Arsha Ali, Zariq George, Samantha Dubrow, Jonathon Smereka, Kayla Riegner, Dawn Tilbury, and Lionel Robert. 2023. Promises and Trust Repair in UGVs. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 67. doi:10.1177/21695067231196235

[18] Connor Esterwood and Lionel P. Robert Jr. 2023. Three Strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness. *Computers in Human Behavior* 142 (2023), 107658. doi:10.1016/j.

chb.2023.107658

[19] Debasmita Ghose, Oz Gitelson, and Brian Scassellati. 2024. Integrating Multimodal Affective Signals for Stress Detection from Audio-Visual Data. In *Proceedings of the 26th International Conference on Multimodal Interaction* (San Jose, Costa Rica) *(ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 22–32. doi:10.1145/3678957.3685717

[20] Ioanna Giorgi, Francesca Ausilia Tirotto, Oksana Hagen, Farida Aider, Mario Gianni, Marco Palomino, and Giovanni L. Masala. 2022. Friendly But Faulty: A Pilot Study on the Perceived Trust of Older Adults in a Social Robot. *IEEE Access* 10 (2022), 92084–92096.

[21] Cedric Goubard and Yiannis Demiris. 2025. Cognitive Modelling of Visual Attention Captures Trust Dynamics in Human–Robot Collaboration. 14, 4, Article 67 (July 2025), 21 pages. doi:10.1145/3732795

[22] Maria J. Grant and Andrew Booth. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal* 26, 2 (2009), 91–108. doi:10.1111/j.1471-1842.2009.00848.x

[23] Haley N. Green and Tariq Iqbal. 2025. Examining Physiological Response and Facial Expression as Indicators of Trust in a Robot Partner. *J. Hum.-Robot Interact.* 15, 2, Article 32 (Dec. 2025), 24 pages. doi:10.1145/3773895

[24] Y. Guo and X. J. Yang. 2020. Modeling and predicting trust dynamics in human–robot teaming: a bayesian inference approach. *International Journal of Social Robotics* 13 (2020), 1899–1909. Issue 8. doi:10.1007/s12369-020-00703-3

[25] Xiaolin He, Jork Stapel, Meng Wang, and Riender Happee. 2022. Modelling perceived risk and trust in driving automation reacting to merging and braking vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour* 86 (2022), 178–195. doi:10.1016/j.trf.2022.02.016

[26] S. K. Hopko and R. K. Mehta. 2022. Trust in Shared-Space Collaborative Robots: Shedding Light on the Human Brain. *Human Factors the Journal of the Human Factors and Ergonomics Society* 66 (2022), 490–509. Issue 2. doi:10.1177/00187208221109039

[27] Sarah K. Hopko, Yinsu Zhang, Aakash Yadav, Prabhakar R. Pagilla, and Ranjana K. Mehta. 2024. Brain–Behavior Relationships of Trust in Shared Space Human–Robot Collaboration. *J. Hum.-Robot Interact.* 13, 1, Article 5 (Jan. 2024), 23 pages. doi:10.1145/3632149

[28] T. M. Hrkalović, B. Dudzik, H. Hung, and D. Balliet. 2025. Partner perceptions during brief online interactions shape partner selection and cooperation. *Plos One* 20 (2025), e0318137. Issue 4. doi:10.1371/journal.pone.0318137

[29] Nikhil Hulle, Stéphane Aroca-Ouellette, Anthony J. Ries, Jake Brawer, Katharina Von Der Wense, and Alessandro Roncone. 2024. Eyes on the Game: Deciphering Implicit Human Signals to Infer Human Proficiency, Trust, and Intent. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. 453–460. doi:10.1109/RO-MAN60168.2024.10731204

[30] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Transactions on Multimedia* 12, 6 (2010), 563–575. doi:10.1109/TMM.2010.2055233

[31] Jiun-Yin Jian, Ann Bisantz, and Colin Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4 (03 2000), 53–71. doi:10.1207/S15327566IJCE0401_04

[32] Changjie Jiang, Tong Zhou, Haoran Shen, and Yanjun Wang. 2026. Understanding Trust and Fatigue Under Transparency Regulation in Air Traffic Control: A Multimodal Approach. *IEEE Intelligent Transportation Systems Magazine* 18, 2 (2026), 21–35. doi:10.1109/MITS.2025.3636860

[33] Sepideh Kalateh, Luis A. Estrada-Jimenez, Sanaz Nikghadam-Hojjati, and Jose Barata. 2024. A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges. *IEEE Access* 12 (2024), 103976–104019. doi:10.1109/ACCESS.2024.3430850

[34] Jacob R. Kintz, Neil T. Banerjee, Johnny Y. Zhang, Allison P. Anderson, and Torin K. Clark. 2023. Estimation of Subjectively Reported Trust, Mental Workload, and Situation Awareness Using Unobtrusive Measures. *Human Factors* 65, 6 (2023), 1142–1160. PMID: 36321727. doi:10.1177/00187208221129371

[35] Derek Koehl and Lisa Vangsness. 2023. Measuring Latent Trust Patterns in Large Language Models in the Context of Human-AI Teaming. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 67, 1 (2023), 504–511. doi:10.1177/21695067231192869

[36] Spencer C. Kohn, Ewart J. de Visser, Eva Wiese, Yi-Ching Lee, and Tyler H. Shaw. 2021. Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology* 12 (Oct. 2021). doi:10.3389/fpsyg.2021.604977

[37] Nikola Kovacevic, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. On Multimodal Emotion Recognition for Human-Chatbot Interaction in the Wild. In *Proceedings of the 26th International Conference on Multimodal Interaction* (San Jose, Costa Rica) *(ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 12–21. doi:10.1145/3678957.3685759

[38] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. PMID: 15151155. doi:10.1518/hfes.46.1.50_30392

[39] Huimin Li, Mengxuan Liang, Ke Niu, and Yaqiong Zhang. 2025. A Human-Machine Trust Evaluation Method for High-Speed Train Drivers Based on Multi-Modal Physiological Information. *International Journal of Human–Computer Interaction* 41, 4 (2025), 2659–2676. doi:10.1080/10447318.2024.2327188

[40] Mengyao Li, Isabel M Erickson, Ernest V. Cross, and John D. Lee. 2022. Estimating Trust in Conversational Agent with Lexical and Acoustic Features. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 66, 1 (2022), 544–548. doi:10.1177/1071181322661147

[41] Mengyao Li, Isabel M Erickson, Ernest V Cross, and John D Lee. 2024. It's Not Only What You Say, But Also How You Say It: Machine Learning Approach to Estimate Trust from Conversation. *Human Factors* 66, 6 (2024), 1724–1741. PMID: 37116009. doi:10.1177/00187208231166624

[42] Magnus Liebherr, Ellen Enkel, Effie L.-C. Law, Mohammad Reza Mousavi, Matteo Sammartino, and Philipp Sieberg. 2026. Dynamic calibration of trust and trustworthiness in AI-enabled systems. *International Journal on Software Tools for Technology Transfer* (Feb. 2026). doi:10.1007/s10009-026-00840-6

[43] Cherin Lim, Tianhao Xu, and Prashanth Rajivan. 2025. Measuring Trust Degradation in AVs: Opportunities and Challenges Using Body Posture and Deep Learning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 69, 1 (2025), 802–806. doi:10.1177/10711813251369369

[44] Chin-Teng Lin, Hsiu-Yu Fan, Yu-Cheng Chang, Liang Ou, Jia Liu, Yu-Kai Wang, and Tzyy-Ping Jung. 2022. Modelling the Trust Value for Human Agents Based on Real-Time Human States in Human-Autonomous Teaming Systems. *Technologies* 10, 6 (2022). doi:10.3390/technologies10060115

[45] Bertram F. Malle and Daniel Ullman. 2021. Chapter 1 - A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction*, Chang S. Nam and Joseph B. Lyons (Eds.). Academic Press, 3–25. doi:10.1016/B978-0-12-819472-0.00001-0

[46] Bertram F. Malle and Daniel Ullman. 2023. Measuring Human-Robot Trust with the MDMT (Multi-Dimensional Measure of Trust). arXiv:2311.14887 [cs.RO] https://arxiv.org/abs/2311.14887

[47] R. C. Mayer, J. H. Davis, and F. D. Schoorman. 1995. An integrative model of organizational trust. *The Academy of Management Review* 20 (1995), 709. Issue 3. doi:10.2307/258792

[48] D. J. McAllister. 1995. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal* 38 (1995), 24–59. Issue 1. doi:10.2307/256727

[49] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM J. Responsib. Comput.* 1, 4, Article 26 (Nov. 2024), 45 pages. doi:10.1145/3696449

[50] L. Miller, J. Kraus, F. Babel, and M. Baumann. 2021. More than a feeling—interrelation of trust layers in human-robot interaction and the role of user dispositions and state anxiety. *Frontiers in Psychology* 12 (2021). doi:10.3389/fpsyg.2021.592711

[51] C. Nass, J. Steuer, and E. R. Tauber. 1994. Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (1994), 72–78. doi:10.1145/191666.191703

[52] Manisha Natarajan and Matthew Gombolay. 2020. Effects of Anthropomorphism and Accountability on Trust in Human Robot Interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 33–42. doi:10.1145/3319502.3374839

[53] Chih-Wei (Charlotte) Ning, Carolina Centeio Jorge, Myrthe L. Tielman, and Mark A. Neerincx. 2026. "What's on your mind?": Understanding the Development of Multidimensional Trust in Social Robots. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction* (Edinburgh, Scotland, UK) *(HRI '26)*. Association for Computing Machinery, New York, NY, USA, 924–933. doi:10.1145/3757279.3785556

[54] Minxue Niu, Zhaobo K. Zheng, Kumar Akash, and Teruhisa Misu. 2024. Beyond Empirical Windowing: An Attention-Based Approach for Trust Prediction In Autonomous Vehicles. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5615–5619. doi:10.1109/ICASSP48485.2024.10446116

[55] Jesse Parron, Rui Li, Weitian Wang, and Mengchu Zhou. 2024. Characterization of Human Trust in Robot through Multimodal Physical and Physiological Biometrics in Human-Robot Partnerships. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*. 2901–2906. doi:10.1109/CASE59546.2024.10711764

[56] Aakriti Prasad, Rahul Agnihotri, and Gunjan Ansari. 2022. Virtual Reality based test batteries for evaluating Trust and Reliability in Human Machine teams. In *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. 691–696. doi:10.1109/ICCCIS56430.2022.10037599

[57] Rodrigo Chacón Quesada, Fernando Estévez Casado, and Yiannis Demiris. 2025. An Integrated 3D Eye-Gaze Tracking Framework for Assessing Trust in Human–Robot Interaction. *J. Hum.-Robot Interact.* 14, 3, Article 57 (June 2025), 28 pages. doi:10.1145/3725861

[58] Yosef S. Razin and Karen M. Feigh. 2024. Converging Measures and an Emergent Model: A Meta-Analysis of Human-Machine Trust Questionnaires. *J. Hum.-Robot Interact.* 13, 4 (Nov. 2024), 58:1–58:41. doi:10.1145/3677614

[59] Matthias Rehm, Ioannis Pontikis, and Kasper Hald. 2024. Automatic Trust Estimation From Movement Data in Industrial Human-Robot Collaboration Based on Deep Learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. 11245–11251. doi:10.1109/ICRA57147.2024.10610822

[60] Amjad Rehman, Muhammad Mujahid, Alex Elyassih, Bayan AlGhofaily, and Saeed Ali Omer Bahaj. 2025. Comprehensive Review and Analysis on Facial Emotion Recognition: Performance Insights into Deep and Traditional Learning with Current Updates and Challenges. *Computers, Materials and Continua* 82, 1 (2025), 41–72. doi:10.32604/cmc.2024.058036

[61] Jimin Rhim, Sonya S. Kwak, Angelica Lim, and Jason Millar. 2023. The dynamic nature of trust: Trust in Human-Robot Interaction revisited. arXiv:2303.04841 [cs.CY] https://arxiv.org/abs/2303.04841

[62] Erin E. Richardson, Jacob R. Kintz, Savannah L. Buchner, Torin K. Clark, and Allison P. Hayman. 2025. Operator-agnostic and real-time usable psychophysiological models of trust, workload, and situation awareness. *Frontiers in Computer Science* Volume 7 - 2025 (2025). doi:10.3389/fcomp.2025.1549399

[63] Kristin Schaefer. 2016. *Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI"*. 191–218. doi:10.1007/978-1-4899-7668-0_10

[64] Feng Tong, Ruohan Wang, Depeng Xu, and Aidong Lu. 2025. Exploring People's Trust towards AI with Human Motions in a Virtual Evacuation Study. In *2025 International Conference on Metaverse Computing, Networking and Applications (MetaCom)*. 88–93. doi:10.1109/MetaCom65502.2025.00021

[65] Abdullah Tutul, Theodora Chaspari, Sarah Levitan, and Julia Hirschberg. 2025. *Detecting Trust Calibration Traits in AI with EEG Signals for Speech Deception*. doi:10.3233/FAIA250631

[66] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. 2013. The Impact of Benevolence in Computational Trust. In *Agreement Technologies*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Carlos Iván Chesñevar, Eva Onaindia, Sascha Ossowski, and George Vouros (Eds.). Vol. 8068. Springer Berlin Heidelberg, Berlin, Heidelberg, 210–224. Series Title: Lecture Notes in Computer Science. doi:10.1007/978-3-642-39860-5_16

[67] Ella Velner, Thomas Beelen, Bob Schadenberg, Roeland Ordelman, Theo Huibers, Khiet P. Truong, and Vanessa Evers. 2024. 'Uhm... Are you sure?' An Exploratory Study of Trust Indicators in Robot-Directed Child Speech. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents* (GLASGOW, United Kingdom) *(IVA '24)*. Association for Computing Machinery, New York, NY, USA, Article 35, 4 pages. doi:10.1145/3652988.3673933

[68] E. J. d. Visser, M. Peeters, M. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx. 2019. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics* 12 (2019), 459–478. Issue 2. doi:10.1007/s12369-019-00596-x

[69] Xin Wang, Stephanie Tulk Jesso, Sadamori Kojaku, David M Neyens, and Min Sun Kim. 2025. VizTrust: A Visual Analytics Tool for Capturing User Trust Dynamics in Human-AI Communication *(CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 581, 10 pages. doi:10.1145/3706599.3719798

[70] Zhenyuan Wang, Qingkun Li, Wenjun Wang, Weimin Liu, Zhaocong Sun, and Bo Cheng. 2025. Deconstructing Driver State in Smart Cockpits: Physiological Validation of Perceived Risk, Trust, and Fatigue during Automated Driving. doi:10.4271/2025-01-7345

[71] Katherine J. Williams, Madeleine S. Yuh, and Neera Jain. 2023. A Computational Model of Coupled Human Trust and Self-confidence Dynamics. *J. Hum.-Robot Interact.* 12, 3, Article 39 (June 2023), 29 pages. doi:10.1145/3594715

[72] 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)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3544548.3581197

[73] Anqi Xu and Gregory Dudek. 2015. OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 221–228. doi:10.1145/2696454.2696492

[74] Tao Xu, Andrei Dragomir, Xucheng Liu, Haojun Yin, Feng Wan, Anastasios Bezerianos, and Hongtao Wang. 2022. An EEG study of human trust in autonomous vehicles based on graphic theoretical analysis. *Frontiers in Neuroinformatics* Volume 16 - 2022 (2022). doi:10.3389/fninf.2022.907942

[75] Binlin Yi, Haotian Cao, Xiaolin Song, Jianqiang Wang, Wenfeng Guo, and Zhi Huang. 2023. Measurement and Real-Time Recognition of Driver Trust in Conditionally Automated Vehicles: Using Multimodal Feature Fusions Network. *Transportation Research Record* 2677, 8 (2023), 311–330. doi:10.1177/03611981231156576

[76] B. Yi, H. Cao, X. Song, J. Wang, S. Zhao, W. Guo, and D. Cao. 2023. How Can the Trust-Change Direction be Measured and Identified During Takeover Transitions in Conditionally Automated Driving? Using Physiological Responses and Takeover-Related Factors. *Human Factors the Journal of the Human Factors and Ergonomics Society* 66 (2023), 1276–1301. Issue 4. doi:10.1177/00187208221143855

[77] Yinsu Zhang, Aakash Yadav, Sarah K. Hopko, and Ranjana K Mehta. 2024. In Gaze We Trust: Comparing Eye Tracking, Self-report, and Physiological Indicators of Dynamic Trust during HRI. In *Companion Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI Companion '24)*. Association for Computing Machinery, New York, NY, USA, 1188–1193. doi:10.1145/3610978.3640649

A Complete Scopus Search Query

```
TITLE(trust)
AND TITLE-ABS-KEY(
    robot OR agent OR machine OR automation OR AI
)
AND TITLE-ABS-KEY(
    multimodal* OR physiological* OR behavior*
    OR speech OR acoustic OR verbal OR gaze OR gesture
    OR video OR face OR audio OR linguistic
)
AND TITLE-ABS-KEY(
    measure* OR assess* OR model* OR predict* OR infer*
)
AND PUBYEAR > 2010 AND PUBYEAR < 2027
AND LIMIT-TO(PUBSTAGE, "final")
AND (
    LIMIT-TO(DOCTYPE, "ar")
    OR LIMIT-TO(DOCTYPE, "cp")
    OR LIMIT-TO(DOCTYPE, "ch")
)
AND LIMIT-TO(LANGUAGE, "English")
AND NOT TITLE-ABS-KEY(
    "zero trust" OR "zero-trust"
    OR "trust aware" OR "trust-aware"
)
AND (
    EXCLUDE(SUBJAREA, "MATH")
    OR EXCLUDE(SUBJAREA, "ECON")
    OR EXCLUDE(SUBJAREA, "MATE")
    OR EXCLUDE(SUBJAREA, "ARTS")
)
```