Delft University of Technology

A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges

Mehrotra, S.; Degachi, C.; Vereschak, Oleksandra; Jonker, C.M.; Tielman, M.L.

DOI
10.1145/3696449

Publication date
2024

Document Version
Final published version

Published in
ACM Journal on Responsible Computing

Citation (APA)
Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M., & Tielman, M. L. (2024). A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM Journal on Responsible Computing*, *1*(4). https://doi.org/10.1145/3696449

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges

SIDDHARTH MEHROTRA, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, Netherlands
CHADHA DEGACHI, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, Netherlands
OLEKSANDRA VERESCHAK, Sorbonne Université, CNRS, ISIR, Sorbonne Université, Paris, France
CATHOLIJN M. JONKER, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, Netherlands and Leiden Institute of Advanced Computer Science, Leiden University, Leiden, Netherlands
MYRTHE L. TIELMAN, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, Netherlands

Appropriate trust in Artificial Intelligence (AI) systems has rapidly become an important area of focus for both researchers and practitioners. Various approaches have been used to achieve it, such as confidence scores, explanations, trustworthiness cues, and uncertainty communication. However, a comprehensive understanding of the field is lacking due to the diversity of perspectives arising from various backgrounds that influence it and the lack of a single definition for appropriate trust. To investigate this topic, this article presents a systematic review to identify current practices in building appropriate trust, different ways to measure it, types of tasks used, and potential challenges associated with it. We also propose a Belief, Intentions, and Actions mapping to study commonalities and differences in the concepts related to appropriate trust by (a) describing the existing disagreements on defining appropriate trust, and (b) providing an overview of the concepts and definitions related to appropriate trust in AI from the existing literature. Finally, the challenges identified in studying appropriate trust are discussed, and observations are summarized as current trends, potential gaps, and research opportunities for future work. Overall, the article provides insights into the complex concept of appropriate trust in human-AI interaction and presents research opportunities to advance our understanding on this topic.

This research was (partly) funded by the Hybrid Intelligence Center, a 10-year program funded the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, grant number 024.004.022 and by the EU H2020 ICT48 project "Humane AI Net" under contract# 952026.

Authors' Contact Information: Siddharth Mehrotra (Corresponding author), Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, Netherlands; e-mail: s.mehrotra@tudelft.nl; Chadha Degachi, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, Netherlands; e-mail: c.degachi@tudelft.nl; Oleksandra Vereschak, Sorbonne Université, CNRS, ISIR, Sorbonne Université, Paris, France; e-mail: vereschak@isir.upmc.fr; Catholijn M. Jonker, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, Netherlands and Leiden Institute of Advanced Computer Science, Leiden University, Leiden, Netherlands; e-mail: c.m.jonker@tudelft.nl; Myrthe L. Tielman, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, Netherlands; e-mail: m.l.tielman@tudelft.nl.

This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2024 Copyright held by the owner/author(s).

ACM 2832-0565/2024/11-ART26
https://doi.org/10.1145/3696449

CCS Concepts: • **General and reference → Surveys and overviews**; • **Human-centered computing → HCI theory, concepts and models**; • **Computing methodologies** → *Artificial intelligence;*

Additional Key Words and Phrases: Appropriate trust, systematic review, trust calibration, warranted trust, appropriate reliance, justified trust, trustworthiness, artificial intelligence

ACM Reference Format:
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 (November 2024), 45 pages. https://doi.org/10.1145/3696449

1 Introduction

Artificial Intelligence (AI) has become an increasingly ubiquitous technology in recent years, with applications in a wide range of industries and areas of life. The ability of AI to process and analyze large amounts of data quickly and accurately makes it particularly valuable for domains with high-stake decision making such as finance, healthcare, and transportation [129, 166]. While AI-embedded systems are powerful, they can still fail or behave unpredictably, leading to inappropriate trust, and introducing the corresponding risk of *misuse* and *disuse* [150].

Both disuse [115] and misuse [118] of AI-embedded systems by humans have led to severe issues, such as Amazon's AI recruiting tool being biased against women [40], a railroad accident in which the crew neglected speed constraints [171], and the use of facial recognition technology in law enforcement to target Black and Latino communities [89]. One of the major reasons of disuse and misuse of AI is people's over- or under-trust in it, or in other words, lack of appropriate trust in AI [142]. Appropriate trust is often linked to the alignment between the perceived and actual performance of the system [198] *or* a match between the human's and system's trustworthiness [122]. We argue that human trust in the AI system must be appropriate because, with appropriate trust in AI, people may be simultaneously aware of the potential and the limitations of AI. This should lead to reducing the harms and negative consequences of misuse and disuse of AI [148].

People have long been aware of the importance of appropriate trust in interpersonal relationships [49]. Taking an example from the Indian scripture *Bhagavad Gita*, dated 400 BCE [57], the deity *Krishna* advises that humans should be careful in trusting others and develop trust in degrees so that their trust is often appropriate [24]. Furthermore, he suggests that by cultivating appropriate trust, humans gradually move forward in interpersonal relationships. This highlights for how long this concept has played a role and is vital and helpful in understanding how people can develop appropriate trust in interpersonal relationships and AI systems.

To achieve appropriate trust in AI systems, different approaches have been taken, such as use of confidence scores [16, 93, 114, 151, 205], explanations [100, 101, 169, 193], and cues (alarms [30, 200], warning signals [139] or uncertainty communication [183]). Many studies aim to adjust the trust bestowed in a system to reflect the trustworthiness of said system [110, 164, 199, 205]. Despite these efforts, a comprehensive understanding of the field is currently lacking, and consensus on the definition of appropriate trust remains elusive. Different perspectives and varying definitions of trust, trustworthiness, and reliance contribute to this lack of clarity, as pointed out by Gille et al. [63].

According to an overview by Jacovi et al. [83], there are numerous types of trust that need to be more precisely defined and differentiated, such as the confusion between two similar yet different concepts, appropriate trust and appropriate reliance, which often stems from a lack of clear understanding of these terms' definitions. Various strategies have been employed to establish an appropriate level of trust in human-AI interaction. Researchers from diverse scientific fields have conducted empirical studies and developed theoretical models to explore different methodologies for building such trust. However, despite the crucial role of appropriate trust in ensuring the successful use of AI systems, there is currently a fragmented overview of its understanding [122].

To highlight and better understand appropriate trust in human-AI interaction, our article aims to present a comprehensive overview of the current state of research on human-AI trust by emphasizing definitions, measures, and methods of fostering *appropriate trust* in AI systems. Furthermore, we make an attempt to map different terms associated with appropriate trust and provide a comprehensive summary of current trends, challenges, and recommendations.

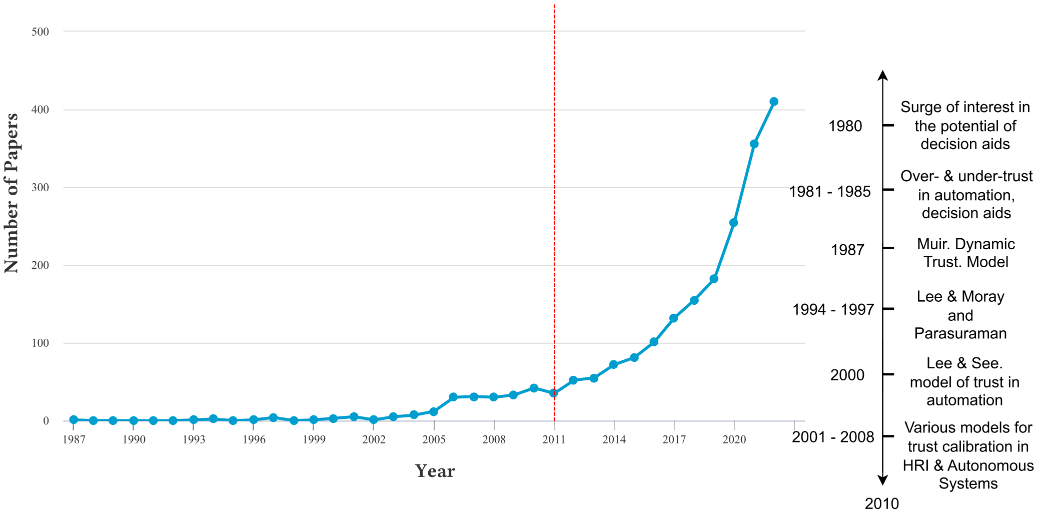
In this work, we study the state of the art in building appropriate trust by examining its evolution, definitions, related concepts, measures, and methods. Our research questions are the following:

(1) What is the history of appropriate trust in automation before AI systems?
(2) How does current research define appropriate trust, and what related concepts exist?
(3) How can we structurally make sense of these concepts related to appropriate trust?
(4) What is the state of the art in fostering appropriate trust in AI systems? This includes the following questions:
   (a) How do studies measure whether the trust is appropriate or not?
   (b) What kind of tasks do researchers employ in their studies to understand appropriate trust?
   (c) What different types of methods for building appropriate trust exist?
   (d) What are the results of the methods aimed at building appropriate trust?

To investigate the preceding questions, we provide a history of appropriate trust development and present a systematic review to identify current practices in the theoretical and experimental approaches. Furthermore, we identify potential challenges and open questions, allowing us to draw research opportunities to understand appropriate trust. First, we provide an overview of the history of understanding appropriate trust in automated systems. Next, we describe our systematic review methodology and the corpus, summarize the current understanding of appropriate trust, and propose a **Belief, Intentions, and Actions (BIA)** mapping to highlight commonalities and differences between concepts. Following this mapping, we present the results of the systematic review, discussing different ways to measure appropriate trust, types of tasks used, approaches to building it, and results of the appropriate trust interventions. Finally, we discuss the challenges identified in studying appropriate trust and summarize our observations as current trends, potential gaps, and research opportunities for future work.

Our novel and main contributions to the field are as follows:

(1) A novel BIA-based mapping of appropriate trust and related concepts.
Our mapping is the result of analyzing how authors define and quantify the abstract notion of appropriate trust and related concepts such as warranted trust, justified trust, and meaningful trust.
(2) An exhaustive presentation of different definitions used, measures of appropriate trust, tasks adopted by authors, and various methods for building appropriate trust through a systematic literature review.
Our presentation is based on similarities and differences in the approaches that authors have used to define, measure, and build appropriate trust in a variety of tasks.
(3) A set of future research opportunities highlighting current trends, challenges, and recommendations for future work.
Our set of future research directions results from a structured summary of our analysis based on the implications of the approaches (definitions, methods, tasks, and measures) adopted by the authors to foster appropriate trust in human-AI interaction.



(a) (b)

Fig. 1. (a) A timeline for the development of appropriate trust as a topic of research from 1987 to 2022 based on the hits from the SCOPUS database. The red dotted line indicates a rapid rise of research interest in appropriate trust research. Our search string for this chart was as follows: ("appropriate trust" OR "calibrated trust" OR "trust calibration" OR "over trust" OR "under trust" OR "over-trust" OR "under-trust") AND PUBYEAR > 1979 AND (LIMIT-TO (LANGUAGE, "English")). (b) Some key papers during the early-stage development of the topic. It is important to note that these are just some key developments and trends in the study of appropriate trust during this time period, and the field has continued to evolve and expand in the years since 2010. Copyright © 2023 Elsevier B.V. All rights reserved. Scopus© is a registered trademark of Elsevier B.V.

2 Background and History of Appropriate Trust

The topic of appropriate trust has been maturing for years. As technology evolved from automated machines to decision aids, virtual avatars, robots, and finally, AI teammates, appropriate trust has been studied in both depth and breadth across a variety of domains. As discussions of the failures of under- and over-trust in automation began to appear, researchers started to study how they could calibrate human trust in automation. One of these early studies defined trust calibration as the relation between user reliance and system reliability [18]. Trust calibration was studied by looking at how usage of a system over time changed trust levels, *calibrating* it to the demonstrated reliability of the system. The coining of this term marked the beginning of appropriate trust research within computer science communities, influenced by, but distinct from, previous trust research in, for example, psychology and philosophy.

Understanding the historical context and evolution of appropriate trust allows us to position this work within the broader context of the field. Therefore, in this section, we chronologically describe past efforts to study appropriate trust until the starting point of our systematic search. The background and history of appropriate trust can provide insights about its conceptualization and how technological and social factors have influenced the research field. Moreover, historical analysis can highlight the various theoretical frameworks that have been used to study trust calibration and their limitations. By examining the historical development of this topic, we can better understand its current conceptualization and identify gaps in the literature. In Figure 1, we illustrate the timeline for these developments.

2.1 1980–1990s: Over- and Under-Trust in Automation

The question of how and when to trust automation easily pre-dates the modern computer era. In the early 1980s, there was a surge of interest in the potential of computer-based decision aids to support decision making in various fields [97, 174, 179]. As automation gained further computing power and was able to solve tasks with high complexity, people started relying on the advice provided by these systems. However, early studies found that users tended to over-trust this advice, even in cases where it was clearly incorrect or irrelevant [79, 167]. This phenomenon was referred to as "automation bias" or "automation-induced complacency" [194]. This concern populated further in the late 1980s, where researchers were concerned about the reliability and safety of nuclear power plants.

Over-trust in automation is only one side of the coin, and under-trust is the other. One of the factors identified as contributing to various accidents such as the Baltimore train incident [170] or misuse of anti-ballistic missiles [55] was the tendency of operators to under-trust the information provided by the control systems and not to rely on them. This problem led to the development of various training and simulation programs aimed at improving operator trust in the automation [14].

In this era, researchers were interested in understanding how humans interact with automated systems and errors that can occur when trust is misplaced. Studying the operator role and human-system integration, Knee and Schryver [1996] found that over- and under-trust stem partially from consistent, reliable performance by the intelligent machines within tasks, problems, and so forth that the human operator may not fully understand (due to the lack of training, experience, or even the ability to be actively involved in system operation). According to them, such cases may support "blind reliance" on the part of the human operator—that is, acceptance of intelligent machine control actions without question of its intent or motives. In conclusion, the study of trust in automation from the 1980s to the 1990s sheds light on the pitfalls of misuse, disuse, and overuse of automated systems, highlighting the importance of understanding how humans trust automated systems.

2.2 1990s: Introduction of Human-Computer Interaction as a Field and Focus on Appropriate Trust

In 1987, Muir [130] presented a model based on dynamics of trust between humans and machines for calibrating a user's trust in decision aids. At this point in time, extensive research began in the **Human-Computer Interaction (HCI)** community to examine the factors that influence a human's trust in automation [75]. One of the themes of this research was calibrated trust.

In the CHI '94 conference, Bauhs and Cooke [18] showcased the effect of system information on trust calibration. The authors reported that the system information aided in calibrating users' confidence in system reliability, but it had little effect on users' willingness to take expert system advice. In the same year, Lee and Moray [105] showed how trust and self-confidence relate to the use of automation and refer trust calibration as correspondence between a person's trust in automation and the automation's capability. Following Lee and Moray's work, a seminal article by Parasuraman and Riley [150] in 1997 on the use, disuse, abuse, and misuse of automation indicated the issue of over- and under-reliance on machines due to lack of trust.

Many articles followed the research of Lee and Moray and Parasuraman and Riley. In 1988, Ostrom [143] showcased that effectively studying trust in automation can help alleviate the uncertainty in gauging the responses of others, thereby guiding appropriate reliance. Tangentially, Endsley and Kaber [51] introduced the concept of situational awareness to tackle the issue of mistrust in automated systems in the same year. Thus, the emergence of trust calibration studies signaled and ushered in a greater focus on user-centered design as a means of minimizing automation disuse and misuse.

2.3 2000s: Emergence of Appropriate Trust as a Key Topic of Research

A seminal article by Lee and See [106] in 2004 provided the first conceptual model of the relationship among calibration, resolution, and automation capability in defining appropriate trust in automation. This work by Lee and See was built on the key work by Cohen et al. [36] in 1998. The Lee & See model was based on purpose, process, and performance dimensions of information that describe the goal-oriented characteristics of the agent to maintain an appropriate level of trust.

In 2006, Duez et al. [48] followed Lee and See's model to study information requirements for appropriate trust in automation, whereas Dongen and van Maanen [187] investigated whether calibration improves after practice and whether calibration of own reliability differs from calibration of the aid's reliability. Thus, researchers were able to develop models of information communication [48] and asymmetrical reliability attribution [187] in automated systems, which improved understanding of how users calibrated their trust over time. Following the mentioned works and literature on calibrated trust, the **Human-Robot Interaction (HRI)** community developed an of understanding appropriate trust in robot capabilities, such as the measures of trust in HRI for detecting over- and under-trust by Freedy et al. [58] in 2007 or the meta-analysis of factors affecting trust in HRI by Hancock et al. [70] in 2011. Their results indicated that improper trust calibration could be mitigated by manipulating robot design, focusing on quantitative estimates of human, robot, and environmental factors. Similarly, Sanders et al. [160] provided a model of human-robot trust targeting performance, compliance, collaboration, and individual human differences to study how human trusts can be calibrated in situations of over- and under-reliance.

The topic of appropriate trust also started to pick up in industrial settings during the 2000s. For example, in 2008, Wang et al. [190] from a defense R&D studied the effectiveness of providing aid reliability information to support participants' appropriate trust in and reliance on a combat identification aid. Their results showed that participants who needed to be made aware of the aid's reliability trusted in and relied on the aid feedback less than those who were aware of its reliability, highlighting appropriate reliance on the aid.

Thus, the emergence of appropriate trust as a prominent topic in the 2000s was marked by the increasing prevalence of automation and innovative steps taken by researchers to study the role of this topic. Notably, the 2004 article by Lee and See [106] introduced a conceptual model that interconnected calibration, resolution, and automation capability to define appropriate trust in automation. This work, which was built on the work of Cohen et al. [36], was followed by many authors (e.g., [48, 70, 160, 190]), where fresh insights were seen considering purpose, process, and performance dimensions of information, offering a deeper understanding for trust calibration. Furthermore, the relevance of appropriate trust extended to industrial settings, as demonstrated by studies on combat identification aids and defense technology [191].

2.4 Parallel Developments: Influential Domains

While research in automation has made significant contributions to our understanding of how people develop and calibrate their trust in computer systems, appropriate trust is also studied in a variety of other fields, including psychology and philosophy. In many cases, our current understanding of appropriate trust has in fact stemmed from the paradigms established in these domains [58, 106, 149].

Different disciplines study appropriate trust differently; however, they all seek the capacity for accurate trust assessment, with the goal of establishing a robust basis for augmented decision making. Appropriate trust has been studied extensively in *psychology*. It is understood as trust that is based on a rational assessment of the risks and benefits of trusting another person or source of

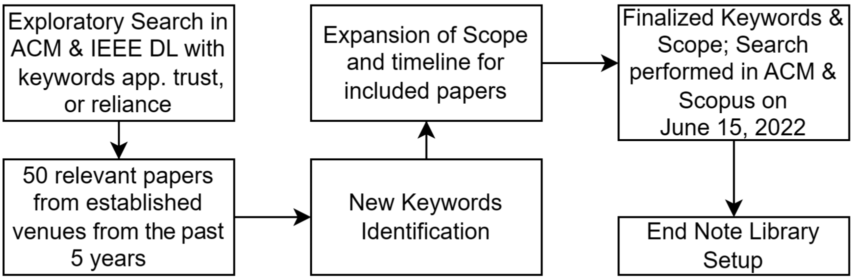

Exploratory Search in ACM & IEEE DL with keywords app. trust, or reliance
50 relevant papers from established venues from the past 5 years
New Keywords Identification
Expansion of Scope and timeline for included papers
Finalized Keywords & Scope; Search performed in ACM & Scopus on June 15, 2022
End Note Library Setup

Fig. 2. Search process for preparing the corpus of the systematic review.

information [107, 162]. For example, Evans et al. [53] showcased that older children (9–10 years of age) are more sensitive to changes in a trustee's characteristics, suggesting that they are not only more trusting but also more discerning in their decisions of when to trust. Similarly, Barnard [17] showcased that how medical professionals change their attitudes and behaviors to gain trust of their patients and proposed which conditions would win justified trust[1] of patients in them. Therefore, in *human-human interaction*, the ability to accurately calibrate trust is essential for building and maintaining strong relationships, as it helps individuals to avoid betrayals and to cultivate mutual respect and understanding. Overall, appropriate trust is an important aspect of social functioning and well-being.

In *philosophy*, the concept of appropriate trust is closely related to the idea of epistemic responsibility, which emphasizes the need for individuals to take responsibility for their beliefs and to use appropriate methods for evaluating evidence and making judgments [19, 152, 158]. In particular, according to Onora O'Neill [140], appropriate trust involves a "reasonable reliance on another's goodwill, competence, and reliability." Other philosophers, such as Karen Jones [90] and Katherine Hawley [72], have also explored the concept of appropriate trust and the importance of carefully calibrating one's trust based on various factors, such as past experiences, social norms, and situational factors.

The research interest in trust calibration evolved slowly compared to promoting trust in automation [33]. This can be partly due to a higher interest in understanding multi-dimensional aspects of trust, and partly due to the complex nature of automation systems. However, from 2012 to 2022, interest in appropriate trust research grew drastically (see Figure 1(a)). This trend was likely driven by the increasing cognitive complexity of AI and ubiquity of interpersonal interactions, as well as organizational interest. Therefore, it has become timely to provide an in-depth literature overview of the state of the art for building appropriate trust in AI. We follow the methodology outlined in this section to provide a comprehensive overview of studies from 2012 until June 2022 in the following section with our systematic review methodology.

3 Systematic Review Methodology

We conducted a systematic review to understand (a) current understanding about building appropriate trust in AI, (b) how appropriate trust and its related concepts have been defined and conceptualized, and (b) what measures and methods have been made to achieve appropriate trust in AI. We adapted the procedure by Calvaresi et al. [25] by developing the research protocol following inclusion and exclusion criteria. For search and identification of the relevant articles, we followed the PRISMA guidelines [146]. The specifications of these guidelines are illustrated in Figure 3.

[1] Different terms have been used in the literature that are related to appropriate trust, such as *justified trust* and *optimal trust*. Refer to Section 5.1 for details.

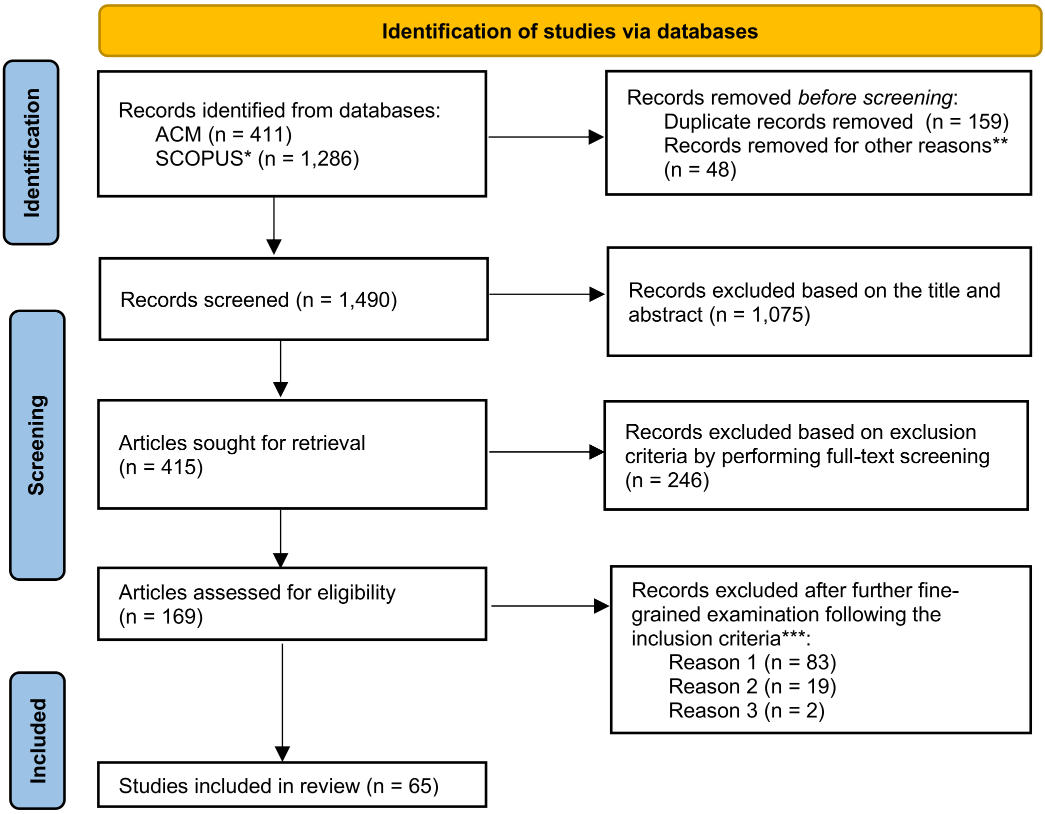

Identification of studies via databases
Identification
Screening
Included
Records identified from databases:
ACM (n = 411)
SCOPUS* (n = 1,286)
Records removed *before screening*:
Duplicate records removed (n = 159)
Records removed for other reasons** (n = 48)
Records screened (n = 1,490)
Records excluded based on the title and abstract (n = 1,075)
Articles sought for retrieval (n = 415)
Records excluded based on exclusion criteria by performing full-text screening (n = 246)
Articles assessed for eligibility (n = 169)
Records excluded after further fine-grained examination following the inclusion criteria***:
Reason 1 (n = 83)
Reason 2 (n = 19)
Reason 3 (n = 2)
Studies included in review (n = 65)

Fig. 3. *SCOPUS data includes all retrieved databases, such as IEEE, Springer, and SAGE. ACM data was excluded from the search, as it was taken from ACM Digital Library.
** Other reasons include records not retrieved, broken URLs, and blank pages in the published record.
*** Reason 1: The article's focus is *not* on appropriate trust derived from the primary or secondary research question. Reason 2: The article does not use a method or a measurement technique to measure or calculate trust. Reason 3: The article is published as a short version of a long paper (in this case, we included the longer version of the article).

3.1 Search String

Appropriate trust is a complex concept, and the term *appropriate* is often interchangeably used with terms for similar concepts (e.g., *appropriate reliance* and *justified trust*) [83, 182]. Therefore, we first conducted an exploratory search to determine which terms for similar concepts are used. In the ACM and IEEE Digital Libraries, we searched for articles with the keywords "appropriate trust" or "calibrated trust" from the previous 5 years.[2] This exploratory search produced 186 results. Among these 186 results, we focused on articles from four of the most most recurring and relevant computer science venues: FAccT, CHI, IUI, and HRI. We selected 50 articles (FAcct: 6, CHI: 20, IUI: 12, and HRI: 12) with the highest use of keywords and similar concepts throughout the articles.

We manually reviewed every title, keyword, and abstract to find new keywords to be included in our final search string (e.g., "optimal trust," "justified trust"). We iterated different combinations

[2] This phase was conducted in May 2022. We decided on the previous 5 years because it coincided with the recent rise of interest in appropriate trust research.

of the keywords until all papers deemed relevant in the exploratory step appeared among the ACM and IEEE Digital Libraries' search results. Analyzing the text of the relevant articles and their references, we noticed that scholars from the computer science community often cite scholars from other disciplines who also study appropriate trust. These disciplines include engineering, social sciences, psychology, mathematics, and decision sciences. Therefore, we decided to include these subjects in our search criteria. Furthermore, we decided to broaden our timeline to include articles published from 2012 to 2022[3] after examining the references of the articles. Figure 2 visualizes our search process and string finalization. The final search string used in ACM and SCOPUS search is as follows:

```
( ( "appropriate trust" ) OR ( "calibrated trust" ) OR ( "warranted trust" ) OR ( "justified trust") OR ( "optimal trust" ) OR ( "responsible trust" ) OR ( "trust calibration" ) OR ( "over trust" ) OR ( "under trust" ) OR ( "over-trust" ) OR ( "under-trust" ) OR ( "meaningful trust" ) ) AND PUBYEAR > 2011 AND PUBYEAR < JUL 2022 AND ( LIMIT-TO (SUBJAREA , "COMP" ) OR LIMIT-TO ( SUBJAREA , "ENGI" ) OR LIMIT-TO ( SUBJAREA , "SOCI" ) OR LIMIT-TO ( SUBJAREA , "PSYC" ) OR LIMIT-TO ( SUBJAREA , "MATH" ) OR LIMIT-TO ( SUBJAREA , "DECI" ) ) AND ( LIMIT-TO ( LANGUAGE , "English" ) )
```

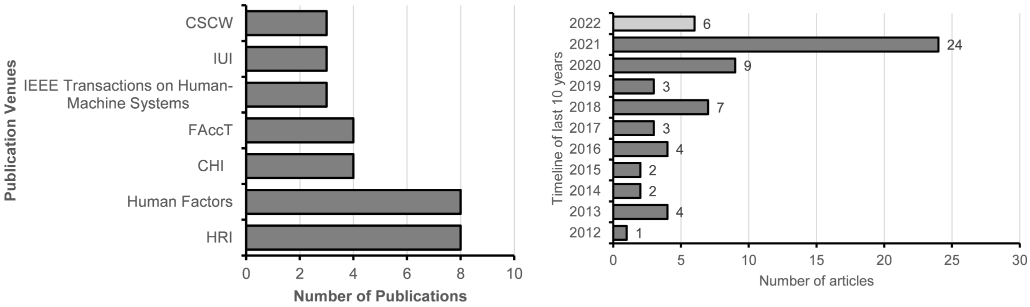
3.2 Selection Criteria

Our search string generated 1,697 articles from the ACM and SCOPUS databases. This phase of generating the final list of articles based on the search string was conducted on June 15, 2022. The screening of articles was carried out manually in three stages: (A) title and abstract screening based on the inclusion criteria, (B) full-text screening based on the exclusion criteria, and then (C) full-text screening with a fine-grained examination based on the inclusion criteria. Our inclusion criteria were as follows:

(1) *Language*: The article should be in English.
(2) *Peer reviewed*: The article should have been peer reviewed. For example, articles from arXiv, OSF (Open Science Foundation), magazine articles, and so forth were excluded.
(3) *Format*: Only full and short articles were included so that all reviewed articles could contain similar details about a study. Therefore, posters, dissertations, workshop papers, workshop calls, and so forth were excluded.
(4) *Publication singularity*: Only the complete version of the article is included.
(5) *Human centered*: Studies needed to have some form of human involvement to be included. For instance, full simulated multi-agent models were excluded.
(6) *Inclusion of a definition*: For a paper to be included, it should have a explicit definition or implicit definition through either referencing previous work or describing measures of appropriate trust or the similar concept (calibrated trust, warranted trust, etc.).
(7) *Conceptualization of appropriate trust*: The articles should conceptualize appropriate trust with measurable constructs or a similar concept. For example, the article uses measurable constructs for appropriate trust.

After applying the inclusion criteria in a two-step abstract and full-text screening process, 169 articles remained. On these 169 articles, we performed further fine-grained examination based on the following criteria:

[3] Since the aim is to identify the current trends and understand recent works in appropriate trust research, we chose to restrain this work to papers published from 2012 to 2022.



(a) Number of the selected papers per top six publishing venues

(b) Number of papers per year from 2012 to 2022

Fig. 4. Distribution of the selected articles from 2012 to 2022 and the top six of their publication venues. Please note that the data for 2022 is incomplete since the data collection for this literature review was conducted in mid-June 2022.

(1) *Contribution scope*: Articles whose primary contribution was unrelated to appropriate trust were excluded. Articles discussing the need for appropriate trust without any direct contribution to define, measure, or model it were also excluded.
(2) *Contribution type*: Surveys, Scoping Reviews, and Literature Reviews were excluded.

The research team registered the protocol of the review with OSF[4] to make the selection of reviews a transparent process. Once the registration was completed, the first and second author independently examined the full text of 169 articles. Both authors used the Rayyan web app [145] to organize their decisions. When there were discrepancies between their decisions, the two researchers involved the senior author in discussing it. This discussion process resulted in the final list of 65 articles for the systematic review.

3.3 Analysis and Corpus Overview

The final corpus consisted of 65 papers, on which we first performed a metadata analysis. Once we annotated each paper based on the preceding inclusion and exclusion criteria, we identified similarities in the measures, methods, and tasks used in each section and grouped them. This resulted in a categorized and complete overview of literature used for studying appropriate trust in human-AI interactions. This also allowed us to identify the common practices and differences in the community. To do so, the first two authors, with a background in HCI and psychology, relied on handbooks and reviews (e.g., [27, 64, 112]) to identify community discussions on methodology, raised issues, and proposed solutions for studying appropriate trust. We also propose research opportunities, challenges, and potential recommendations based on our analysis of the corpus in Section 6.

We were also interested in the publication venues, timeline of publications, and application scenarios. The top six publication venues and chronological distribution of articles from 2012 to 2022 are shown in Figure 4. In Figure 4(a), we can observe that the most popular venues, among others, are Human Factors and Ergonomics Society (Human Factors) and HRI (n = 8 each), and CHI and FAccT (n = 4 each, idem), which account for 47.3% of the final corpus. In addition, the last 5 years had experienced a growth in the number of publications related to the appropriate trust (see Figure 1). This trend reflects a growing interest in human-centered AI and the importance of studies focusing on appropriate trust, distinct from enhancing trust in AI. A spreadsheet containing

[4] https://osf.io/c78tw/?view_only=16c398038f474b9b8922277a3fd94c87

a list of all final papers is included in the supplementary material. Furthermore, we provide raw data files containing the list of papers for replicability and also to ensure that papers that appeared after our search string finalization can be added to perform an updated review.

4 Definitions and Related Concepts

Appropriate trust in AI systems is growing rapidly as a research field for both researchers and practitioners. To understand how to achieve appropriate human trust in AI (human-AI trust), it is important to understand how we define it and its related concepts. The increasing interest in trustworthy AI research [180] has brought to light a growing need for clarity among the community regarding the different concepts and definitions related to appropriate trust in AI.

Terms like *appropriate trust*, *calibrated trust*, and *appropriate reliance* are often used interchangeably in prior research [163]. There have been debates in the community about what appropriate trust is and how different concepts related to appropriate trust are different or similar—for instance, during the CHI TRAIT workshop in 2022 [15] and the CSCW '23 workshop on "Building Appropriate Trust in Human-AI Interactions" [7]. These debates are a result of the complex nature of trust in AI systems, which can be difficult to understand and evaluate. In this systematic review, we identified different terms related to appropriate trust in the literature, the most common ones being calibrated trust (n = 16), appropriate reliance (n = 8), and warranted trust (n = 6). The full list of terms is available in Table 1, with the corresponding definitions as given by the papers. We can see from Table 1 that there is often more than one definition of appropriate trust or its related concepts. This discord and diversity among different concepts motivates us to establish links between them and present a unified mapping.

4.1 BIA Mapping

Given the number of terms and slightly different definitions that exist, our first aim is to achieve a clearer understanding of the different concepts surrounding appropriate trust. To this end, we grouped all presented concepts at different levels of human perception in a conceptual mapping, following the theory of human-practical reasoning by Michael Bratman [22]. In this subsection, we will first discuss the relationships among appropriate trust and its related concepts following this mapping. Following, we relate the concepts to the definitions presented by the authors of the included papers.

We illustrate our categorization of the concepts associated with appropriate trust in Figure 5, which presents a BIA mapping of appropriate trust and related concepts. These levels allow us to separate the different perspectives on trust as a belief, intention, or action. More specifically, *Beliefs* describe a perception of the world and the other agents in it, including beliefs about other agents' intentions and actions. Beliefs may or may not be justified based on current information about the world and past agent behavior [56]. Second, *Intentions* represent the deliberative state of the human—what the human has chosen to do. Intentions are desires to which the human has to some extent committed [61] and provides a mechanism to translate beliefs into action plans. Finally, *Actions* describe events as they actually occur in the interaction [8], such as a doctor actually offering a patient an in-person consultation. In essence, the rationale behind the BIA mapping is to provide a structured framework for understanding appropriate trust and related concepts by delineating into three levels: Beliefs, Intentions, and Actions.

Our mapping identifies two actors: the human and the AI agent. The human actor is illustrated with a 'user' icon, and the AI agent is represented by a robot icon. To help distinguish between the different concepts, we formally define them. In our definitions, we use the following variables: $human \in H$ for a human, $agent \in A$ for an AI agent, and $T_{(x,y)}$ to denote the trust that trustor x has in trustee y. We then use the following notations:

Table 1. Definitions of Appropriate Trust and Its Related Concepts

Keyword	Definition
Appropriate Trust: Based on system performance or reliability	1. Appropriate trust is the alignment between the perceived and actual performance of the system. Appropriate trust is to [not] follow an [in]correct recommendation. Other cases lead to over-trust or under-trust [198]. 2. If the reliability of the agent matches with user's trust in the agent, then trust is appropriately calibrated [139]. 3. In human-robot teaming, appropriate trust is maintained when the human uses the robot for tasks or subtasks the robot performs better or safer while reserving those aspects of the task the robot performs poorly to the human operator [142].
Based on TW and beliefs	1. Appropriate trust in teams happens when one teammate's trust towards another teammate corresponds to the latter's actual trustworthiness [91]. 2. We can understand "appropriate trust" as obtaining when the trustor has justified beliefs that the trustee has suitable dispositions [39].
Based on the calculations	1. "Appropriate trust is the fraction of tasks where participants used the model's prediction when the model was correct and did not use the model's prediction when the model was wrong; this is effectively participants' final decision accuracy" [193]. 2. FORTNIoT (a smart home application) predictions lead to a more appropriate trust in the smart home behavior. Meaning, we expect participants to have reduced under-trust (i.e., they trust the system more when it is behaving correctly) and reduced over-trust (i.e., they trust the system less when it is behaving incorrectly) [38]. 3. Trust appropriateness was calculated by subtracting an ideal from a participant's allocation for a given round. Thus, a positive value indicates trust that is too high, a negative value indicates trust that is too low, and 0 indicates calibrated, appropriate trust [86]. 4. The level of trust a human has in an agent with respect to a contract is appropriate if the likelihood the human associates with the system satisfying the contract is equal to the likelihood of the agent satisfying that contract* [202]. 5. The term appropriate trust then is the sum of appropriate agreement and appropriate disagreement of humans with the AI prediction [110].
Warranted Trust	1. "Warranted trust describes a match between the actual system capabilities and those perceived by the user" [164]. 2. "Human's trust in an AI model (to Contract - C) is warranted if it is caused by trustworthiness in the AI model. This holds if it is theoretically possible to manipulate AI model's capability to maintain C, such that Human's trust in AI model will change. Otherwise, Human's trust in AI model is unwarranted" [83].
Justified Trust	1. "Justified trust is computed by evaluating the human's understanding of the model's decision-making process. In other words, given an image, justified trust means users could reliably predict the model's output decision" [2].
Contractual Trust	1. "Contractual trust is when a trustor has a belief that the trustee will stick to a specific contract" [83]. 2. "Contractual trust is a belief in the trustworthiness (with respect to a contract) of an AI" [56].
Calibrated Trust	1. "Trust calibration is the process by which a human adjusts their expectations of the automation's reliability and trustworthiness" [104]. 2. Calibrating trust is if explanations could help the annotator appropriately increase their trust and confidence as the model learns [62]. 3. Trust calibration refers to the correspondence between a person's trust in the automation and the automation's capabilities* (based on Lee and Moray [105] and Muir [130]) [106].
Well-Placed Trust*	"[T]he only trust that is well placed (intention) is given by those who understand what is proposed, and who are in a position to refuse or choose in the light of that understanding" [140].
Responsible Trust	"The area for responsible trust in AI is to explore means to empower end users to make more accurate trust judgments" [109].
Reasonable Trust*	"Reasonable trust requires good grounds for confidence in another's good will, or at least the absence of good grounds for expecting their ill will or indifference" [13].

*Articles before the year 2012 or after the end of the search date.

Note: The asterisked articles were not included in the review process.

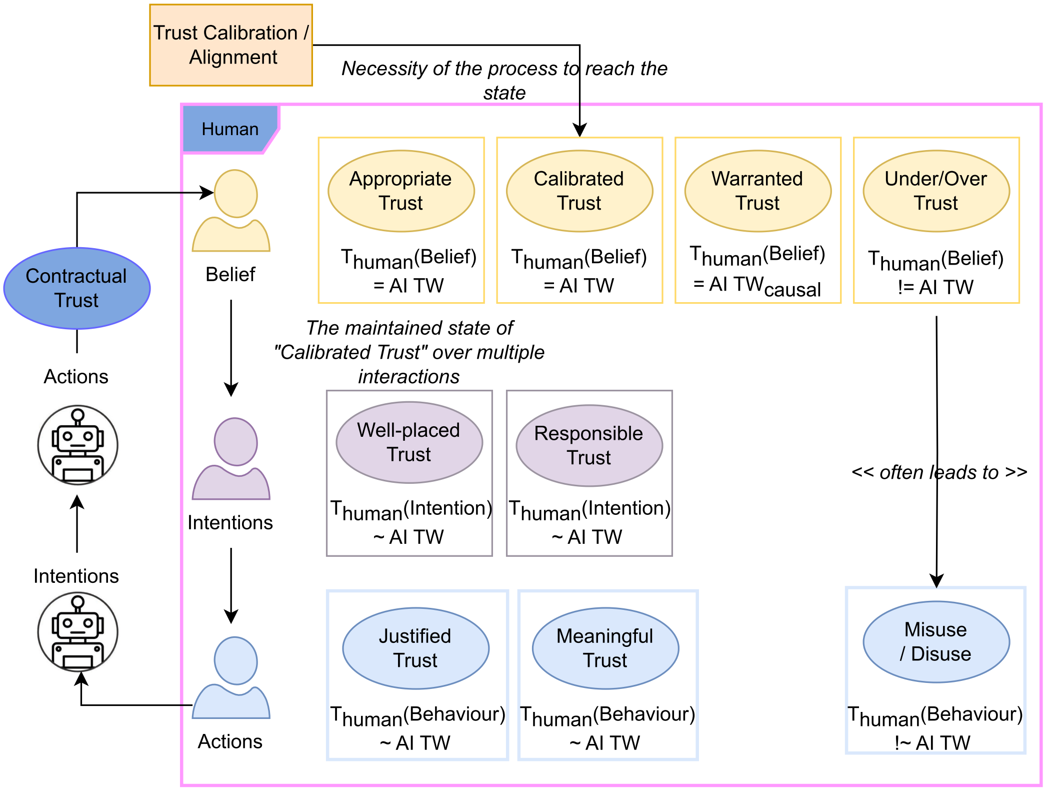

Trust Calibration / Alignment
Necessity of the process to reach the state
Human
Belief
Appropriate Trust
Thuman(Belief) = AI TW
Calibrated Trust
Thuman(Belief) = AI TW
Warranted Trust
Thuman(Belief) = AI TWcausal
Under/Over Trust
Thuman(Belief) != AI TW
Contractual Trust
Actions
The maintained state of "Calibrated Trust" over multiple interactions
Intentions
Well-placed Trust
Thuman(Intention) ~ AI TW
Responsible Trust
Thuman(Intention) ~ AI TW
<< often leads to >>
Intentions
Actions
Justified Trust
Thuman(Behaviour) ~ AI TW
Meaningful Trust
Thuman(Behaviour) ~ AI TW
Misuse / Disuse
Thuman(Behaviour) !~ AI TW

Fig. 5. In this figure, we present a BIA mapping of appropriate trust and related concepts. The pink outline represent the elements linked with the human (h) and the AI agent entity. The black colored circle with a robot icon represents the AI agent. For brevity, when writing T(Belief), we mean Trust(human,agent) (Belief), and for TW(agent), we use AI trustworthiness. In addition, Under/Over Trust and Contractual Trust are represented in different colors, as these types of trust are not (necessarily) appropriate.

$$
\begin{aligned}
T_{(human,agent)}(Belief) &= \text{Trust of the human in the AI agent is about the human's belief about the agent.}\\
TW_{agent} &= \text{Actual trustworthiness of the agent.}\\
T_{(human,agent)}(Intention) &= \text{Trust of the human in the agent is about the human's intentions toward the agent.}\\
T_{(human,agent)}(Behavior) &= \text{Trust of the human in the agent is about the human's behavior toward the agent.}
\end{aligned}
$$

We have divided the definitions of appropriate trust in the Table 1, see the keyword "Appropriate Trust" based on the similarities such as system performance or reliability, beliefs, and calculations. Based on this division, we can formulate our conceptualization of appropriate trust. We define trust to be appropriate when the human's trust formed by beliefs about the AI agent's trustworthiness (denoted as $TW_{(human,agent)}(Belief)$) is equal to the AI agent's actual trustworthiness $TW_{agent}(Actual)$ (refer to Equation (1)). In other words, the human's beliefs about the AI agent's trustworthiness match the agent's true level of trustworthiness. This definition aligns with the one proposed by Jorge et al. [91] and Mehrotra et al. [122].

$$\text{Appropriate Trust} \iff T_{(human,agent)}(Belief) = TW_{Agent} \tag{1}$$

Appropriate trust is based on the beliefs contrary to intentions or actions as it pertains to an individual's perception of the world and other agents. This form of trust reflects whether the individual's beliefs about other agents are appropriate based on actual trustworthiness. From this foundation, we go on to differentiate between the many related concepts for appropriate trust that we have encountered. As a note to our readers, the definitions and terms presented in the Table 1 do not always match one-to-one with our conceptualization in Figure 5, because sometimes different authors define the same term in different ways.

First, we consider *calibrated trust*, the most common term in the reviewed corpus, which introduces notions of dynamic trust and trust variations to the human-AI interaction [118, 161]. We define calibrated trust as similar to appropriate trust in that a human's trust belief about the agent corresponds to their actual trustworthiness. However, calibrated trust necessarily involves a process of *trust calibration* or *trust alignment* that corrects for over- and under-trust over the course of time and repeated interactions. We postulate that appropriate trust is the maintained state of "calibrated trust" over multiple interactions. Nevertheless, human trust in an AI system may be appropriate even without calibration.

Distinct from appropriate or calibrated trust is *over-trust*—that is, the human's trust beliefs in an AI agent's is greater than the AI agent's actual trustworthiness TW_{Agent}. Similarly, when the human's trust belief in an AI agent is less than the AI agent's actual trustworthiness TW_{Agent}, then a state of *under-trust* is reached.

Next, *warranted trust* is defined as trust caused by the trustworthiness of the AI agent. More specifically, we talk about warranted trust if there is a causal relationship between trustworthiness of the trustee and the trust of the trustor [56]. Although we expect warranted trust to mostly be appropriate, not all appropriate trust needs to be warranted. In other words, while trust that is well supported by evidence and reasoning is probably appropriate, there may be situations where trust is appropriate even if there is no clear evidence or justification to support it. For instance, if an e-commerce website has a polished and visually appealing design, it may create an initial positive impression in the user's mind. This positive impression, in turn, may lead the user to trust the website's content to some degree, even though they lack in-depth evidence about the product's quality. Finally, *contractual trust* in the AI agent is based on the belief that the AI agent will uphold an explicit contract (upholds (AI, C)) which specifies what the AI agent is expected to do [71, 176]. Here, the contract may refer to any functionality of the AI agent that is deemed useful, even if it is not concrete performance at the end task for which the AI agent was trained [83]. It is important to highlight that contractual trust differs from many other definitions in that it does not directly imply appropriateness, as the human's beliefs about the agent might not be related to its actual trustworthiness.

Unlike the three preceding concepts, *well-placed trust* and *responsible trust* are built around intentions.[5] Here, both well-placed trust and responsible trust are defined as intentions about how to act toward the agent (denoted as TW_{human}*(Intent)*). This means that if a human has well-placed trust, its intentions are correct given the trustee's trustworthiness. Here, intentions are being referred to as a human's intentions to rely on, cooperate with, or be vulnerable to the agent (trustee) in some way. These intentions reflect the human's willingness to trust the agent.

We separate beliefs and intentions because people's intentionsdo not always directly follow from trust beliefs. For example, consider Mary, who has recently acquired a GPS system but does not fully trust its accuracy. Despite her reservations, she intends to use the GPS while driving to an unfamiliar location, as it offers some guidance compared to navigating without assistance. In this scenario, if the GPS successfully provides accurate directions, Mary's decision to use it would be considered well-placed trust. However, it would not be appropriate trust, as her level of confidence in the system does not match its actual performance. Furthermore, this example also

[5] Following the BDI (Belief-Desire-Intention) software model, intentions represent the deliberative state of the agent—that is, what the agent has chosen to do [153].

does not represent misplaced trust as it occurs when trust is placed in a system or person that is not deserving of that trust, leading to potential negative outcomes. Finally, warranted trust, which is built around intentions, has been an important factor for AI evaluation in hiring as shown by Agata Mirowska [127]. Mirowska showed how intentions can be measured using the two-item measure of Taylor and Bergmann [178] that helps in evaluating warranted trust.

Trust is justified when a human's behavior is appropriate given the agent's trustworthiness. In this case, the human actor is acting trustingly toward an agent. The ability of a user to evaluate trust does not make the AI agent more accurate, robust, and reliable in itself; it can only, at best, make the use of the AI by the human more accurate, robust, and reliable, leading to *justified trust*. In contrast, when a human's trust is not justified based on the AI's agent trustworthiness, it is plausible that the user can lean toward misuse or disuse of AI. Continuing Mary's example from before, imagine that in the end she remembers the route halfway through on seeing a familiar landmark. Because of this, she stops using the GPS system, as she believes that it is not very good, even though in reality it works well. This example shows a case of inappropriate distrust (as she wrongly believes the GPS to be bad), well-placed trust (as she intended to use the system), and unjustified mistrust (as she ends up not disusing the system where that is not necessary). Similarly, there might be an instance where Mary uses her GPS even though she does not believe that it is trustworthy, as she does not know the route herself. In this case, with a trustworthy GPS system, we would have inappropriate distrust but justified trust, as her behavior does match the system's trustworthiness.

So far, we have described our mapping related to the human actor. Now, we shift our focus to the AI agent. As mentioned earlier, we consider the AI agent attributed with intent. Here, the AI agent can form an intention based on the human behavior that can be associated with the action(s) it can take resulting in the human to form beliefs about the AI agent's trustworthiness, making it a closed-loop process. This process helps in arriving at the belief that a contract, as outlined by Jacovi et al. [83], has been established between the pair and will be adhered to in the future. In other words, an AI agent can form an intention based on human behavior or actions by observing the human behavior. Based on its intention, it can decide how to respond by taking an action. By doing so, the AI agent can ensure that it acts in accordance with the contractual agreement (actions the AI agent is authorized to take and the expectations for how the AI agent should behave) and maintains the trust of the human.

In addition to the aforesaid concepts, we found a few further, more minor, and less defined terms that are not completely covered in Figure 5, and these are the ones which we have not defined explicitly, namely meaningful trust (closely linked to *justified trust*), optimal trust (*justified trust*), moral trust (*responsible trust*), capacity trust (*perceived capability*), and well-deserved trust (*justified trust, warranted trust*). Our list of related concepts is not exhaustive, and there could be further concepts that appear outside the domain of our review that we have not included in our search criteria.

In summary, we have described the distinction between appropriate trust and related concepts stemming from human beliefs, intentions, and actions. The distinction among belief, intention, and action in defining concepts related to appropriate trust is crucial because it allows for a comprehensive understanding of dynamics of trust underlying human-AI interactions. By separating these components, we can pinpoint discrepancies between subjective perceptions, behavioral intentions, and actual actions, thus enabling the identification of factors influencing trust formation and maintenance. We believe that this is one of the first conceptualizations in human-AI interaction research to describe, associate, and categorize various concepts in a single framework, which could help reduce the discord among the community on approaching the concept of appropriate trust.

5 Results of the Systematic Review

In this section, we review how authors of the included papers define and measure appropriate trust,[6] what different domains, settings, and tasks they employ, methods for building appropriate trust, and the results achieved.

5.1 Measures (How to Measure Appropriate Trust?)

Human trust is studied differently based on whether it is conceptualized as a mental attitude [26, 54], a belief [91, 94, 204], or a behavior [27, 139, 197]. These approaches are typically linked to specific measures which either focus on subjectively measuring attitudes or beliefs (*linked to intentions and beliefs from the BIA mapping*), or which look at behavior (*linked to actions from the BIA mapping*) which demonstrates human trust. As measuring the appropriateness of trust naturally includes measuring trust, we draw the same distinction and divide this subsection into three parts in accordance with Wischnewski et al. [196]: (a) perceived trust, (b) demonstrated trust, and (c) mixed approach. Simply put, we say that *perceived trust* is about measuring a person's subjective beliefs, whereas *demonstrated trust* focuses on their behavior [124]. While measuring perceived trust is typically done via questionnaires, surveys, interviews, focus groups, and similar reporting tools, demonstrated trust is usually about measuring trust-related behaviors (i.e., in the form of reliance). In demonstrated trust, participants are given the option to use or rely on the system. The underlying assumption is that the more often people use or rely on a system, the more they trust it.

5.1.1 Perceived Trust. Among the papers in our corpus, nearly 40% measure appropriate trust or related concepts by examining a match between the system's capabilities and the user's trust as a belief. The most common strategy to measure appropriate trust was manipulating a system's trustworthiness and using self-report scales to compare how self-reported trust adapts to the trustworthiness' levels. For example, Chen et al. [31] presented participants with either 60%, 70%, 80%, or 90% reliable systems and measured trust through subjective self-report. Similarly, with a within-subjects experimental design, de Visser et al. [44] had participants interact with a system in which trustworthiness levels were manipulated through its reliability from 100% to 67%, 50%, and, finally, 0%. Then, the authors used a self-reported trust scale to measure trust, which then through comparison with the system's trustworthiness provided appropriateness of trust. In the prior examples, manipulating trustworthiness helped the authors do a before/after comparison. According to Miller [124], this comparison is crucial for measuring appropriate trust. The authors highlight that without manipulating the trustworthiness of the machine, we cannot establish whether the intervention has correctly calibrated trust.

We found that some authors measured perceived trust by performing a match between the trust ratings and the *static* reliability of the robot or the AI system [3, 21, 43, 85, 99, 111, 142]. However, the match between trust ratings and static reliability may not be perfect. There may be other factors such as appearance or behavior that influence how people rate the trustworthiness of a robot or AI system, even if the system is perfectly reliable. Furthermore, this method does not take into account the dynamic nature of trust. Therefore, we argue that it is difficult to match performance levels with subjective scale ratings.

5.1.2 Demonstrated Trust. We found that only about 26% of studies used behavioral measures for appropriate trust, and we identified three approaches to do so. The most common is *agreement percentage*, which is the percentage of trials in which the participant's final prediction agreed with the AI's correct prediction and cases where participants did not agree with the AI's wrong

[6] We follow the same terminology (appropriate trust/calibrated trust/warranted trust, etc.) as the authors of the reviewed papers to maintain consistency.

prediction [16, 23, 110, 133, 193, 205]. Usually, appropriate trust is seen as a sum of appropriate agreement ratio (human agreement with correct AI predictions) and appropriate disagreement ratio (disagreement with incorrect AI predictions) [38, 110, 139]. Another measure of appropriate trust is related to *switch percentage* [205]—that is, the percentage of trials in which the participant decided to use the AI's prediction as their final prediction instead of their own initial prediction. However, it is usually not a stand-alone measure of appropriate trust and is coupled with other measures. For example, Zhang et al. [205] used a statistically significant interaction between switch and agreement percentages and the AI's confidence level. When the AI's confidence level was high and the switch and agreement percentage was high (and vice versa), then trust was deemed appropriate.

A final method is to measure *ideal trusting behavior* during the task beforehand and compare to which extent the actual users' behavior matches it [74, 86]. For example, for an experiment where users have to delegate a number of tasks to AI, it is possible to calculate the most optimal number of tasks to delegate to AI to achieve the best speed and performance at a given AI's reliability [74]. The closer users are to this number, the more appropriate their trust in AI is.

5.1.3 Mixed Approach. Nearly 20%[7] of studies from our corpus used a combination of both self-report measures and behavioral measures to understand appropriate trust. These measures can be categorized into two different subgroups.

The first subgroup includes measures that focus on participants' decisions and compliance with the system's recommendations, along with self-reported scales. For example, Wang et al. [192] measured appropriate trust by letting users decide when and when not to trust a low-reliability robot. They measured self-reported trust by modifying the Mayer scale [117] and used behavioral measure of compliance as dividing the number of participant decisions that matched the robot's recommendation by the total number of participant decisions. Accordingly, when both measures matched the reliability of the robot, the trust was considered appropriate. Similarly, Kaniarasu et al. [93] conducted a study where participants rated trust at trust prompts and used buttons to indicate trust changes. Appropriate trust was measured by examining the degree of alignment of user's trust with the robot's current reliability (high or low). Finally, Zhang et al. [203] measured participants' reliance on AI using two behavioral indicators, agreement frequency and switch-to-agree frequency, as well as via subjective trust ratings. Their diverging reliance and subjective trust ratings results highlight the difference between these two types of measures.

The second subgroup includes measures that examine how participants calibrate their trust over time as they become more familiar with the system's capabilities and policies. For example, Albayram et al. [3] measured how participants calibrated their trust as they grew familiar with the system's capabilities by using subjective responses and number of images allocated to the automation for pothole inspection by varying automation reliability. Similarly, de Visser et al. [44] varied the anthropomorphism of the automation to understand trust calibration and appropriate compliance. By using both subjective ratings and a compliance measure, they measured appropriate trust as the match of a user's trust with the actual reliability of the aid. In both of the previous examples, the researchers manipulated the trustworthiness of the system to measure an appropriate level of trust. This approach is in line with Miller [123], who states that "there must be some known or estimated 'level' of trustworthiness that is manipulated as part of the evaluation."

5.1.4 Synopsis. In summary, measures of appropriate trust typically involve either a comparison of two different measures—trust of the human and trustworthiness of the system—or some

[7] The remaining 14% of the reviewed papers presented frameworks or theoretical models where no user study was conducted in which a measure of appropriate trust was used.

form of agreement metric. The first type naturally involves knowing the trustworthiness of the system. Trustworthiness can be defined as absolute (e.g., the system is correct or not) or relative (the system gets better/worse over time). Although the first might give more insight into how good the system is, it does mean that the AI needs to be either wrong or right, which needs to be known. The relative measure allows for an easier comparison, as appropriateness is just about whether trust moves up or down in the same direction as trustworthiness. However, if trust is low for a nearly perfect system and slightly higher but still low for a perfect system, it is still inappropriate despite moving in the correct direction.

Comparing trustworthiness with trust naturally also involves measuring trust. In this as well, two methods can be distinguished. The first is subjective and behavioral measures based on questionnaires, and the second is on actions. The main disadvantage of questionnaires is that outcomes can be difficult to directly compare with trustworthiness, whereas it is easier to establish if reliability is correct. However, questionnaires better capture the concept of trust as a nuanced belief, as reliance behavior could be caused by more than just high trust. This is also reflected in the differences between behavioral and subjective scales that can occur when both are used [203]. This highlights the disadvantage of seeing appropriate trust in terms of an agreement metric; this is, by definition, about reliance behavior and often imposes constraints on the type of human-AI collaboration. Given the limitations of most current measures, the option to use different methods simultaneously has the opportunity to offer a more nuanced result. Which mix is the best might depend highly on the collaboration between the human and AI.

An example of simultaneous use of different methods is (a) the use of validated questionnaires to measure perceived trust combined with (b) behavioral measures to measure demonstrated trust that could offer a more insightful measurement than the use of one alone [196]. The underlying assumption is that these measures provide an accurate understanding of a human's trust. However, as human trust is a multi-dimensional concept, its measurement based on scales or behavior might not provide its complete understanding [173]. For example, behavioral measures are context specific and may not generalize well across different situations, and subjective measures may involve participants' individual biases or the willingness to disclose their true feelings. Therefore, we propose that the next steps in determining how to measure appropriate trust should be to examine combination of measures other than perceived or demonstrated trust. These measures can include personality traits [59], past experiences [66], social norms [181], and cultural values [206], and how these measures can differ across different contexts and populations.

Individual differences such as cultural background, prior experience with technology, and personality traits significantly influence trust in AI, offering a nuanced understanding of trust dynamics and its measurement. Cultural background shapes perceptions and attitudes toward AI, as evidenced by varying levels of trust across different countries. For instance, people in emerging economies like Brazil, India, China, and South Africa exhibit higher trust in AI compared to those in developed nations like Finland and Japan, where trust levels are notably lower [65]. Similarly, in a study by Rau et al. [154], Chinese participants preferred an implicit communication style of an intelligent robot than German participants and evaluated it as being more trustworthy than German participants. Prior experience with technology also plays a crucial role in trust formation. For example, individuals familiar with AI tend to trust it more due to their better understanding and positive past interactions [108]. Personality traits further impact trust in AI, with studies indicating that people with certain personality characteristics, such as openness to experience and lower levels of neuroticism, are more likely to trust AI systems [156]. Overall, these individual differences highlight the importance of considering diverse user profiles when measuring user trust in an AI system.

Table 2. Domains and Associated Tasks across Our Corpus

Domain	Tasks
Military	Object recognition [34, 85, 111], Prediction [21], Reconnaissance [41, 69, 84, 86, 137, 139, 192, 200], Remote operation [42, 88, 182], Search and rescue [93], Non-experimental [142, 183]
Transport	Automated driving [1, 11, 12, 50, 67, 73, 95, 98, 99, 113, 126, 186, 189, 195], Non-experimental [159]
Domain agnostic	Classification [135, 198], Multi-arm trust game [37], Object recognition [203], Non-experimental [32, 45, 76, 82, 91, 164, 165, 168]
Healthcare	Classification [74, 132, 133, 135], Meal design [23], Non-experimental [56, 83, 109, 119]
IT	Prediction [62], Classification [16], Question answering [16], Non-experimental [87]
Justice	Prediction [110, 193], Classification [16], Question answering [16]
Sustainability	Prediction [193], Disassembly [6]
Gaming	Prediction [78], Classification [43]
Robotics	Non-experimental [28], Classification [135]
Consumer products	Prediction [38]
Finance	Prediction [205]

5.2 Tasks

In this section, we describe the tasks and domains observed in the corpus of this review. We cluster these tasks around distinguishing characteristics that emerged.

We grouped all studies into different application domains to get an overview of the tasks. In enumerating the domains seen within our corpus of papers (Table 2), we observe that military operations, transport, and domain-agnostic applications are the most common in appropriate trust research. On a more granular level, we see that tasks such as automated driving (n = 14), prediction and classification (n = 14), and reconnaissance (n = 8) are most commonly given to users. Human-AI collaborative tasks such as working in a military environment with humans (e.g., [193]) and teaming for military missions [139, 182] are the particularly preferred cases of the reviewed articles. The popularity of military and transport application fields within the study of appropriate trust could follow from the more severe risks associated with the incorrect use of technology in those settings. Overall, the understanding and operationalization of appropriate trust varies by domain and task, driven by practicalities of the experiment design and the measures and definitions it affords. For example, Yang et al. [198] define appropriate trust as [not] following a [in]correct recommendation in a machine learning classification task, whereas Holzinger et al. [77] define it as the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency, and satisfaction in a medical context.

When analyzing the breadth of user studies included in this review (n = 46, 45.6% between, 32.6% within subject, 15.2% mixed design), we see a number of patterns emerge in the characteristics of the tasks assigned to participants. We group those characteristics along the dimensions of risk, dynamism, and users' expertise. Interestingly, only three studies [133, 137, 186] perform a non-controlled experiment, relying on think-aloud sessions, co-design, and interview sessions. They targeted medical, mobility, and military experts for interaction design. To some extent, this does suggest a lack of space within appropriate trust research for the voices of users and stakeholders, and little input on its design processes on their part.

5.2.1 Risk. We highlight risk as an integral part of experimental setups, as vulnerability is a key element of trust [106, 117]. Yet, it can be overlooked in studies of human-computer trust. We differentiate between explicit and implicit risk using the criteria proposed by Miller [124]. In these criteria, trust is characterized by the presence of vulnerability and stakes, which introduce a downside to inappropriate trust. The user must be aware of these stakes throughout the experiment so that actions can be adjusted to accommodate risk. We see that 78.3% of studies include an element of risk in their design. This element is largely implemented in one of two ways; simulated through points gained and lost (45.7%), or incentivized through performance-based pay bonuses (21.7%). Only one study [6] used a task that was risky in the experimental setting itself, namely disassembling traction batteries in a recycling context.

The remaining papers rely on the understood risk of a given task (automated driving and remote operation) in the real world to assume that users would engage realistically with their experiment [42, 88, 110, 182, 195] or do not discuss risk in their methodology [38, 198]. Given the importance of risk to trust, it is difficult to argue that users in such studies demonstrated trust at all, and with no consequences attached to over- and under-trust, users may rely on and positively perceive a system regardless of its actual trustworthiness.

5.2.2 Dynamism. The next element of task design we analyzed is dynamism—that is, changes in human-AI trust over time informed by the history of interaction [75]. Specifically, we investigate whether studies measure trust levels at multiple points, thus accounting for this dynamic aspect of trust. Across all studies, we find that 63% measure trust more than once throughout the task. In cases of automated driving tasks, this can sometimes even be a continuous measure of trust derived from driver behavior [1, 95, 195]. Meanwhile, a third of studies measured trust only once throughout the experiment, reducing the complexity of the trust relationship to one snapshot.

Moreover, most of the studies reviewed were either laboratory based, using simple tasks or theoretical models, which further fails to reflect real-world scenarios. Thus, generalizability of these findings to more complex and dynamic real-world situations is uncertain.

5.2.3 Participant Expertise. Overall, 67.4% of studies recruited non-expert participants, because often researchers design tasks so that the participant pool felt equally qualified to complete them without any specific training [12, 16, 23, 37, 193, 205]. Recruitment of non-experts also occurred for the tasks that could require more specialized knowledge, such as military-related tasks [93]. The main reason could be that candidates with required expertise are not available and/or are not easily found. This claim can be supported by the fact that all automated driving studies recruited licensed drivers to their experiments, whereas only three non-automated-driving user experiment studies recruited expert participants [62, 132, 182]. Given that a users' perception of their expertise can affect the extent to which they trust and rely on the automated system [203], participant expertise should more closely align with the expected expertise of the end user for more realistic results.

5.3 Methods for Building Appropriate Trust (How to Achieve It?)

In this section, we describe what different approaches were taken toward achieving appropriate trust in the reviewed corpus. A categorization of the methods revealed four broad categories: (1) improving system transparency; (2) cognition and perception; (3) models, guidelines, theories, and frameworks; and (4) relational framing and continuum of trust. These are further shown in Figure 6.

5.3.1 Improving System Transparency. The first category of methods aims to achieve appropriate trust by adding transparency to systems. About 52% of articles in our corpus target improving transparency of the system to build appropriate trust—that is, informing users about the specific capabilities and limitations of AI. This indicates that there is a common assumption that improving system transparency can help the human user better decide when to trust or distrust the AI system.

One way transparency is improved is through providing *explanations*. Explanations focus on the inner workings of the AI systems (n = 16), appearing either for every AI recommendation

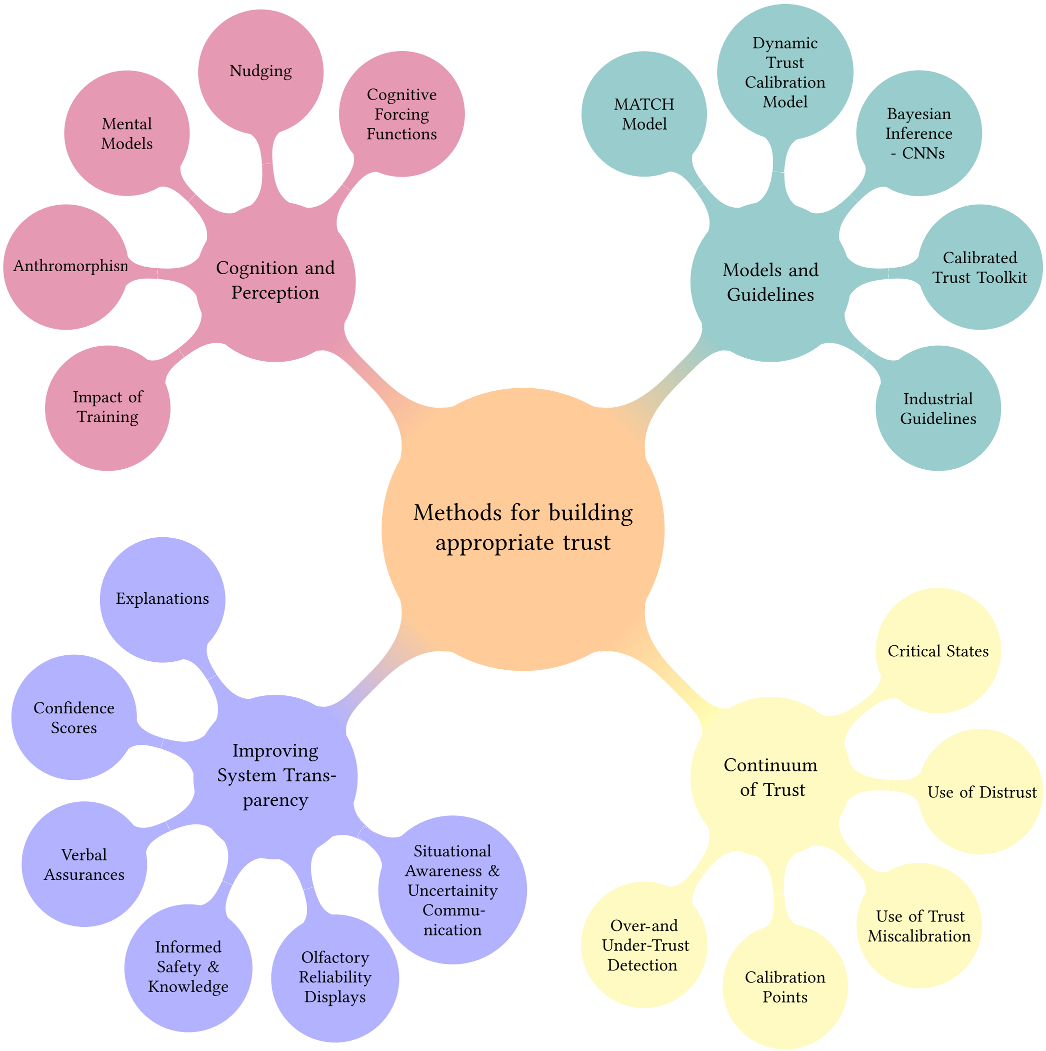

Methods for building appropriate trust
Cognition and Perception
Mental Models
Nudging
Cognitive Forcing Functions
Anthromorphisn
Impact of Training
Models and Guidelines
MATCH Model
Dynamic Trust Calibration Model
Bayesian Inference - CNNs
Calibrated Trust Toolkit
Industrial Guidelines
Improving System Trans-parency
Explanations
Confidence Scores
Verbal Assurances
Informed Safety & Knowledge
Olfactory Reliability Displays
Situational Awareness & Uncertainity Commu-nication
Continuum of Trust
Critical States
Use of Distrust
Use of Trust Miscalibration
Calibration Points
Over-and Under-Trust Detection

Fig. 6. Overview of the different appropriate trust building methods adopted in the articles from our corpus.

[62, 182, 203, 205] or under specific circumstances. For example, adaptive explanations by Bansal et al. [16] appear only for the predictions where the AI is quite confident and are absent for the low confidence predictions as a way to avoid human over-trust in the latter case. This explanation method was found to be effective in trust calibration, as here the AI system adjusts to the user's attitude and behavior following the signs of over- and under-trust. To further mitigate over-trust, Lakkaraju and Bastani [102] call for designing explanations as an interactive dialogue where end users can query or explore different explanations for building appropriate trust.

Another way to instill transparency is through *confidence scores* of the AI models to align the user's trustworthiness perception of the system with the actual trustworthiness (n = 12). These scores reflect the chances that the AI is correct, thus relating to its competence and capability. According to Zhang et al. [205], confidence scores are a simple yet effective method for trust calibration. However, it does not necessarily improve AI-assisted decision making [16]. Furthermore, confidence scores are not always well calibrated in machine learning classifiers [136], which can lead to inappropriate trust.

A combination of explanations and confidence scores has been used for appropriate trust as well under the term of *informed safety and knowledge* in relation to autonomous vehicles [95]. The confidence scores informed the drivers of the vehicle's safety. At the same time, the explanations were provided to demonstrate the vehicle's knowledge of any maneuver, enabling the drivers to adjust their level of trust in the system appropriately.

Similar to confidence scores, *uncertainty communication* (n = 3)—that is, emphasizing the instances when AI is "unsure" of a prediction or does not have a definite answer—can also calibrate trust. For example, an AI agent can yield back the full control to humans and explicitly indicate that it does not "know" the solution [183]. The results of this method demonstrate that it helps users spot flows in the reasoning behind the AI predictions and when AI is "unsure" about them, and consequently rapidly calibrate their trust.

While confidence scores and uncertainty communication come mostly in a form of a text message, their more anthropomorphized counterpart is *verbal assurances*. Within this method of transparency, the system verbally indicates to the users what it can and cannot do in a form of promises [3, 6] or intent [113]. For example, the results of Albayram et al. [3] show that participants calibrated their trust based on the system's observed reliability following the promise messages. Besides written or verbal indicators, odors, presented as *olfactory reliability displays* [195], can also serve to communicate a change in reliability levels of AI for users to calibrate their trust. The authors communicated a change in reliability levels of an automated vehicle simulator using two odors: lemon for a change to low reliability and lavender for a change to high reliability. Their results indicate that olfactory notifications are useful for trust calibration.

Providing more information about not only the AI capability but also about the task and the context, or in other words, *situational awareness communication*, can provide transparency to achieve appropriate trust [12, 88]. For example, Azevedo-Sa et al. [12] showed that with activation of different communication styles to encourage or warn the driver about situational awareness when deemed necessary helps in calibration of trust. Similarly, the results of Johnson et al. [88] show that warning drivers about situational awareness is effective at increasing (decreasing) trust of under-trusting (over-trusting) drivers, and reducing the average trust miscalibration time periods by approximately 40%.

Studying various methods of improving system transparency for building appropriate trust in AI systems can provide valuable insights. Overall, these works show the value of understanding system transparency and how it can be increased in multiple ways. Some of the most common examples are confidence scores and explanations. However, we also see some unique solutions, such as using olfactory displays or verbal assurances. All of these solutions seem promising for improving system transparency, but that communicating system uncertainty or providing real-time situational awareness helps is only sometimes a given. This shows that there is still much to gain, especially in understanding why an AI system is uncertain or what can help improve its situational awareness for improving system transparency.

5.3.2 Cognition and Perception. Another group of methods to achieve appropriate trust is related to human factors, accounting for 21% of the reviewed papers. Several of them focus on the *users' mental model* of AI [142]. The more correct the mental model is, the more likely it is that trust will be calibrated appropriately, which links back to our previous method of increasing transparency. One of the ways to achieve this is through training users how to perform the task and how to collaborate with an AI-embedded system [88, 133]. The results show that training that emphasized the shortcomings of the system appeared to calibrate expectations and trust [88]. Another way to build a more correct mental model of AI is to let users observe the system's performance over time [16]. By observing the system performance over time in a study by Bansal et al. [16], participants developed mental models of the AI's confidence score to determine when to trust the AI.

Other human factors are related to *nudging and cognitive forcing functions*. For example, adding friction in the decision-making process of AI to purposefully slow down its recommendation and providing users a nudge gives them an opportunity to better reflect on the final decision [132]. The results of Naiseh et al. [132] show that with a nudging-based explainable AI approach such as "You are spending less time than expected in reading the explanation," users can calibrate their trust in AI. Similarly, introducing cognitive forcing interventions—that is, not automatically showing AI recommendations but on-demand or with forced wait—can significantly reduce over-reliance compared to the simple explainable AI approaches [23].

Another potential method to calibrate trust through understanding human factors was proposed by Johnson et al. [88]. The authors gave participants trust calibration *training* about task work and teamwork before the task. Their results showed that training that emphasized the shortcomings of the autonomous agent appeared to calibrate expectations and trust. Last, the characteristics of an AI-embedded system, notably its degree of *anthropomorphism*, contribute to appropriate trust [43]. The results showed that increasing the humanness of the automation increased trust calibration—that is, compliance rates matched with the actual reliability of the aid on increasing humanness.

Thus, studying cognition and perception can help us better understand how people interact with AI systems and how they form impressions of AI systems. In addition, studying the mental processes involved in perception, learning, reasoning, and decision making can help us in designing for appropriate trust in AI systems.

5.3.3 Models and Guidelines. Theoretical foundations can provide insights into how to establish appropriate trust in human-AI interaction (n = 12) [42, 82, 88, 126, 164, 168]. One example is using models and frameworks to understand how the actual and perceived trustworthiness of AI systems relate to each other. Several papers use different models to explain this relationship and suggest ways to improve it. For instance, Schlicker and Langer [164] use two models from organizational psychology to identify factors that influence the match between how trustworthy the system is and how trustworthy the user thinks it is. Similarly, Israelsen et al. [82] propose a three-level model that compares the user's and AI's abilities, analyzes the user's past experiences with similar systems, and measures the user's willingness to depend on the system. Some similarities between the theoretical models we reviewed are that they often try to explain how the user's perception of the AI system's trustworthiness is influenced by various factors, such as the system's performance, reliability, transparency, feedback, and context. These factors can help us understand how people interact with AI systems. By understanding these factors that influence trustworthiness, we can design AI systems that can be appropriately trusted [122].

Another type of model focuses on the *communication of trustworthiness cues* in AI systems. For example, Liao and Sundar [109] proposed the MATCH model for responsible trust, which describes how trustworthiness should be communicated in AI systems through trustworthiness cues. With their model, they highlight transparency and interaction as AI systems' affordances for designing trustworthiness cues. Apart from communicating trustworthiness cues, some authors studied building appropriate trust by allowing for real-time trust calibration [1, 69, 165]. For example, Shafi [165] provided a parametric model of machine competence that allowed generating different machine competence behaviors based on task difficulty to study *trust dynamics* for real-time trust calibration. Furthermore, Guo and Yang [69] modeled trust dynamics using *Bayesian inference* when a human interacts with a robotic agent over time. Here, based on the real-time trust values, a human can calibrate its trust in the robot.

Unlike theoretical models, *guidelines* offer practical design solutions to achieve calibrated trust in AI. For instance, a *calibrated trust toolkit* [186] aids transparent design of autonomous vehicles, analogous to methods in Section 6.3.1. These guidelines address post-design implementation, offering a road map for human factors in industrial robots and trust calibration for robotic teammates.

Besides academic efforts [32, 42, 133, 142, 183, 188], industrial organizations offer guidelines for designers and developers of AI-embedded systems [9, 10, 80, 147]. These guidelines are often recommendations or best practices that are developed to help people make informed decisions or take specific actions. The majority of *industrial guidelines* in the field of human-centered AI focus on building users' trust rather achieving appropriate trust (or related terms discussed in Section 5), and only one, Google PAIR guidebook [147], provides key considerations for users' trust calibration. Examples of their key considerations are telling users what the system cannot do, tying explanations to user actions, and considering the risks of a user trusting a false positive or negative. Overall, the key considerations outlined in the Google PAIR guidebook emphasize the importance of effective communication and transparency of the AI models in building appropriate trust in AI systems linking back to the importance of improving the system's transparency.

In synopsis, various theoretical models and guidelines have been proposed to understand the mechanisms around achieving appropriate trust in AI. Theoretical foundations, such as the models and frameworks, provide valuable insights into the factors influencing trustworthiness perception. By examining factors like system performance, reliability, transparency, feedback, and context, we understand how users interact with AI systems, ultimately aiding in designing AI systems that can be appropriately trusted.

Furthermore, models like the MATCH model by Liao and Sundar [109] focus on communicating trustworthiness cues, emphasizing transparency and interaction as essential elements in designing trustworthiness cues in AI systems. Like the one proposed by Shafi [165], real-time trust calibration models offer insights into how trust dynamics can be managed during human-AI interactions, allowing for adjustments based on task difficulty and performance. In addition to theoretical models, practical guidelines play a vital role in achieving calibrated trust in AI. These guidelines offer actionable recommendations for designers and developers, ensuring that AI systems align with their original design intent. It is worth noting that industrial organizations also contribute to this field, offering guidelines that often focus on building users' trust but increasingly recognize the importance of achieving appropriate trust through effective communication and transparency, as emphasized by the Google PAIR guidebook [147].

A nuanced approach is crucial in designing trust models for AI systems, considering the intricate interplay of various factors influencing trustworthiness. Likewise, when confronted with many guidelines on trust in AI, tailored selection and adaptation are crucial to ensuring that the chosen guidelines align closely with the unique context, objectives, and stakeholders of the AI system under consideration. Therefore, designing a comprehensive model that addresses all aspects is a complex challenge. Similarly, navigating the many guidelines for building appropriate trust in AI systems can be overwhelming. Therefore, it is essential to consider the specific context, domain, and stakeholders involved. Different guidelines may have varying focuses, such as ethics, explainability, or fairness, so selecting the most relevant ones based on the specific requirements and goals of the AI system can help guide the implementation of appropriate trust measures.

5.3.4 Continuum of Trust. To achieve appropriate trust, one has to be able to recognize when it is not there to fix this. Therefore, studying the entire continuum of trust beyond its appropriate level (i.e., over-, under-, mis-, and dis-trust) is helpful in achieving it. For example, it can be possible to achieve calibrated trust through fostering both trust and distrust in AI at the same time [126]. Sensibly placed distrust makes users not agree with the opinion of others automatically but rather increases their cognitive flexibility to trust appropriately [144]. Yet, only 14% of the reviewed papers look into this. The literature proposes terms like *calibration points* [119] or *critical states* [78] to classify the situations when the intervention for calibrating trust is needed. The former term is characterized as a way to classify situations in which the automation excels or situations in which the automation is degraded [119]. The latter is characterized by the situations in which it is quite important to take a certain action, such as an autonomous vehicle detecting a pedestrian [78]. In both of these situations, a mismatch can occur between levels of performances and expectations, which would allow users to reflect whether their trust levels are appropriate or not.

Generally, we find that the reviewed papers mostly rely on analyzing human behaviors to determine whether trust needs to be calibrated. For example, states of over- and under-trust are inferred from monitoring the user's reliance behavior rather than subjective trust measures [139]. Collins and Juvina [37] propose to watch out for any behaviors that can be considered as an exception out of principle of trust calibration (appropriately calibrated trust) to understand better long-term trust calibration in dynamic environments [37]. In their study with a multi-arm trust game, during critical states, users unexpectedly changed their trust strategy, tending to ignore the advice of the previously trusted AI advisors and leaning more toward the previously non-trusted ones. One of the unique findings from this work was that (a) trust decays in the absence of evidence of trustworthiness or untrustworthiness, and (b) perceived trust necessity and cognitive ability are important antecedents on the trustor's side to detect cues of trustworthiness.

The previous example teaches us that trust calibration is a complex process that requires a nuanced understanding of the context and user behavior, and that the ability to adapt and change trust strategies in response to changing situations is an important aspect of successful trust calibration. Similar to Collins and Juvina, Tang et al. [177] explicitly used distrust behaviors by leveraging data mining and machine learning techniques to model distrust with social media data. Distrust was conceptualized such that it can be a predictor of trust and of the extent to which it is miscalibrated. Last, one paper relied on physiological markers such as gauge behavior from an eye tracker coupled with the rate of reliance on AI and compared it with the system's capability to identify if trust is miscalibrated [12].

In conclusion, there are various approaches adopted by the authors ranging from examining behavior and performance to studying distrust and trust miscalibration for building appropriate trust. Authors have proposed over- and under-trust detection, calibration points, and critical states to study appropriate trust through the continuum. Furthermore, studies on distrust have shown that it can play a critical role in trust calibration, and trust miscalibration can be used to understand long-term trust calibration in dynamic environments.

5.4 Results of Calibration Interventions

In this subsection, we provide a general overview of the findings of the reviewed papers. In particular, we focus on the results of applying the methods for building appropriate trust described in Section 6.3.

From the categories of methods described in this section, improving system transparency was the most common. Most papers supported the hypothesis that transparency facilitates appropriate trust in a system. For example, it was found that uncertainty ratings [183], confidence scores [205], providing explanations [16, 102, 135], and reliability and situational awareness updates [12] improved appropriate trust in a system. However, other papers add some nuance to this conclusion. For instance, Bansal et al. [16] found that explanations increased the human's acceptance of an AI's recommendation, regardless of its correctness. Furthermore, Wang and Yin [193] found that only some of their tested explanations improved trust calibration, indicating that not all explanations are equal. Last, although confidence scores can help calibrate people's trust in an AI model, Zhang et al. [205] found that this largely depends on whether the human can bring in enough unique knowledge to complement the AI's errors. These results highlight that further research is necessary to study exactly what methods of increasing transparency are useful to facilitate appropriate trust, given the context of the interaction. We believe that opportunities lie in exploring how diverse factors such as user expertise, task complexity, and the type of explanation influence trust calibration. This could involve controlled experiments that manipulate different transparency elements to pinpoint their individual and combined effects on trust.

Improving system transparency had mixed results for building appropriate trust, and leveraging human cognition and perception for trust calibration yielded the similar results. For example, Riegelsberger et al. [157] found that changes in how a system interacts with the user impacted users' perception of trustworthiness. Similarly changing the interaction with the system, Buçinca et al. [23] found that cognitive forcing functions[8] reduced over-reliance on AI. However, the performance of human+AI teams was worse than the AI alone with these functions. Other than the use of cognitive forcing functions to compel people to engage more thoughtfully with AI systems, Naiseh et al. [132] found that nudging can also help users become more receptive and reflective of their decision, possibly leading to appropriately trusting the AI system. As nudging and cognitive forcing functions target cognitive and perceptual mechanisms for building appropriate trust, the effectiveness of training is also intricately linked to the these mechanisms. For example, two studies showed that teams receiving the calibration training reported that their overall trust in the agent was more robust over time [88, 133]. Based on these findings, it is crucial to focus on developing interventions that promote analytical cognitive thinking to foster appropriate trust in AI systems.

The appearance of a system plays a significant role in shaping how humans perceive and mentally process its attributes, which in turn impacts their levels of trust in the system. For example, Jensen et al. [85] discovered that a system with a more human-like appearance was perceived as more benevolent, but this did not lead to differences in trust in behavior leading to unsupported trust calibration. Similarly, both Christoforakos et al. [35] and de Visser et al. [44] found that more human-like systems were considered more trustworthy but did not help in trust calibration. These results highlight that the human-likeness strategies for building appropriate trust have been challenging so far. Although it seems clear that there is some effect of appearance on trust, how to use this properly to ensure the appropriateness of trust remains an open question.

So far, we have looked at results of the trust calibration interventions related to improving system transparency and understanding human cognition and perception including human-likeness. Distinct from these methods, understanding the continuum of trust was also helpful in certain cases for building appropriate trust. For instance, calibration points and critical states prompted users to adjust their trust in the system by facilitating specific moments of engagements [78, 119]. Furthermore, detecting over- and under-trust was critical in providing trustworthiness cues to the user in calibrating their trust levels. However, the use of these cues was found to not necessarily improve the performance of the human-AI teams [139]. Finally, miscalibrated (i.e., over- or under-) trust [37] and distrust [98] were also promising to calibrate human trust in the system in certain situations, such as under conditions of increased trust necessity. Miscalibration affected interactions with new trustors, as a reputation for past trustors preceded the entity, causing potential new trustors to approach with caution [98]. Therefore, understanding the continuum of trust through user studies can help in building appropriate trust, which can improve the human-AI team performance and be helpful in trust repair. In particular, opportunities lie in conducting more empirical

[8] Interventions implemented during decision making to disrupt heuristic reasoning and prompt analytical thinking such as on-demand explanation or forced waiting for output [103].

studies investigating trust development over time with different contexts and how this impacts human decision making.

In summary, the methods applied in the selected papers yielded mixed results—on the one hand where improving system transparency and understanding human perception and cognition had an impact on appropriateness of trust, but on the other hand it did not improve the human+AI joint performance. Similarly, studying the continuum of trust helped in fostering appropriate trust but also failed to improve human-AI team performance as well as in repairing trust. Overall, it remains complicated to find a one-size-fits-all solution for building appropriate trust in AI systems. Therefore, we recommend that future researchers give careful consideration to (a) how they define appropriate trust, (b) specify what they mean by it, (c) how they conceptualize their measures, and (d) avoid using related concepts in particular.

6 Discussion

In this systematic review, we have discussed the (a) history of appropriate trust, (b) differences and similarities in concepts related to appropriate trust, (c) a BIA mapping to understand commonalities and differences of related concepts, (d) different methods of developing appropriate trust, and (e) results of those methods. In this section, we reflect on our findings by providing critical insights on elaborating limitations of the current research. Furthermore, we provide some novel perspectives on understanding appropriate trust and finally acknowledge the limitations of this work.

6.1 Limitations in Current Research

With appropriate trust constituting a central variable to the appropriate adoption of AI systems, different approaches have been taken to understand it. Our aim with this study was to provide an overview of the field's current state. In doing so, we reflected on our findings and found some challenges that exist in our way of understanding this research area. In this subsection, we elaborate on the aforementioned key challenges and how to overcome possible limitations, as well as summarize critical points with research opportunities for future work. Our identified key challenges are as follows:

(1) Discord and diversity in concepts related to appropriate trust such as calibrated trust, justified trust, and responsible trust.
(2) A strong focus on appropriate trust in capability, leaving out other aspects of trust such as benevolence and integrity [91].
(3) The issues involved in adequately measuring appropriate trust.
(4) Designing appropriate experimental tasks involving risk and vulnerability in the study design.

6.1.1 Discord and Diversity in Understanding Appropriate Trust. From the analysis of the reviewed definitions of appropriate trust, we identify three major challenges for the current theoretical discourse on the topic. First, as seen in Section 5, there is no uniform understanding on what appropriate trust is: some papers define appropriate trust based on system performance or reliability [137, 139, 142, 189, 198], some relate it to trustworthiness and beliefs [39, 91], and some base it on calculations [38, 86, 193]. Such a variety of the appropriate trust definitions stems from different understanding of what "the right amount of trust" implies. The common denominators of having various definitions of appropriate trust can be linked to the following: (a) the context in which it is studied often differs from one study to another; (b) the multi-dimensional nature of trust, often associated with attitude or subjective beliefs, adds complexity to understanding appropriate trust; and (c) different academic fields approach the study of trust in unique ways, leading to divergent interpretations of appropriate trust. For example, in the HRI domain, trust is often linked to the robot's performance [93], whereas in psychology, it is commonly linked to understanding social and interpersonal aspects [155].

In addition to the variety of definitions of appropriate trust, we also found that the literature proposes various related concepts[9] (see Table 1), sometimes used interchangeably in the discourse about appropriate trust [16, 139, 195]. For example, we would like to especially stress the difference between appropriate trust and another most used related construct—calibrated trust. Although the logical formulation of the two concepts is similar as shown in the BIA mapping in Figure 5, trust calibration requires a process. In contrast, appropriate trust is the maintained state of the calibrated trust over a series of interactions. This conceptual overlap raises questions about the precise boundaries and distinctions between these concepts and highlights the need for a more refined and standardized conceptual framework.

These challenges surrounding understanding appropriate trust emphasize the significance of shaping our research agenda in this domain. To address the need for consensus among researchers, in this work we proposed a framework that explicitly defines appropriate trust and its boundaries. Our framework considers multiple dimensions, such as system capability, trustworthiness, beliefs, and task requirements, while accounting for contextual variations. Moreover, we made an attempt to clarify the relationships between appropriate trust and related concepts, establishing clear definitions and boundaries to facilitate meaningful discussions and avoid conceptual confusion. By addressing these challenges and shaping a coherent research agenda, we can advance our understanding of appropriate trust and its implications for various domains.

6.1.2 Prominent Focus on the System's Capabilities in Definitions. The majority of appropriate trust definitions or its related concepts focus on the capability or ability of an agent. Here, appropriate trust is the alignment between the perceived and actual capabilities of the agent by the human [110, 139, 198]. Much of previous research has looked at 'ability' as the core factor of establishing trust [91, 121], which brings the focus upon the engineering aspect of trustworthiness. However, we view trustworthiness as more than just ability. Our interpretation of trustworthiness can be enhanced when we not only focus upon agent capabilities but also on understanding other factors such as integrity and benevolence [117, 149] or process and purpose [106].

Hoffman [76] states that "a thorough understanding of both the psychological and engineering aspects of trust is necessary to develop an appropriate trust model." Our examination of the psychological aspects of trust in human-AI interaction has revealed a need for improvement in the existing literature regarding modeling the integrity and benevolence of an AI agent toward a human as highlighted by Ulfert et al. [185], Mehrotra et al. [122], and Jorge et al. [91]. Mayer et al. [117] propose that the effect of integrity on human trust will be most salient early in the relationship, before the development of meaningful benevolence—that is, X has the disposition to do good for Y [47]. Therefore, we believe that it is important to first investigate how humans perceive the AI system's integrity and how to model this relationship for fostering appropriate trust in AI system. Then it becomes vital to study the effect of perceived benevolence on trust as it increases over time as the relationship between the parties develops [121]. Throughout, the perceived ability of the system remains important. However, we believe that it is crucial to not forget these other factors in research on appropriate trust.

6.1.3 Adequately Measuring Appropriate Trust. While analyzing our corpus, we encountered common issues with appropriate trust measurements identified by Miller [124]. These issues include the robustness of single/multiple-item questionnaires in capturing changes in trust levels

[9] From our understanding, a "concept" is a general idea representing a category, whereas a "definition" is a precise statement that clarifies the meaning of a term or concept.

over time, reliance on agreement/disagreement with model predictions without considering discrepancies in human goals, and the use of appropriate situational awareness as a proxy for trust.

First, it is difficult to establish whether single/multiple-item questionnaires are robust enough to capture changes in trust levels over time [6, 37, 88, 95, 98]. In addition, in almost 40% of studies, trust is measured before and after the user study, although it is not always appropriate to reflect on users' attitude at such a high level of granularity. A focus on trust dynamics over time as indicated by some studies [6, 11, 69, 98] could be a better approach.

Second, measures of trust related to whether humans agree or disagree with a model prediction are employed in some studies [23, 110, 203, 205]; however, what happens when the model targets differ from human goals? Third, reliance was often used as a proxy for trust or even treated as the same thing. As Tolmeijer et al. [182] highlighted *trust in an agent* as the belief that "an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" [106], whereas *reliance on AI* is defined as "a discrete process of engaging or disengaging" [106] with the AI system.

Finally, some authors [12, 88] acknowledge the ambiguity of using appropriate situational awareness as a proxy for measuring appropriate trust in their approach.

6.1.4 Designing Appropriate Experimental Tasks. We observe several methodological limitations recurring within the scope of this review, including an absence of risk and vulnerability elements in user studies, overlooking instances of under-trust, and uncertainty regarding the extent to which behavioral experiments can capture trust.

First, we found some papers in our corpus [6, 38, 93, 133, 164] that had little or no element of risk in the task design. We posit that in a questionnaire, survey, or field study, it is crucial that participants have experienced or currently experience vulnerability to the possibility of the AI system failing. Trust cannot exist without the element of risk, and participants must have a personal stake in the situation. Including risk and vulnerability factors allows researchers to evaluate the trustworthiness of systems or services accurately.

Second, we observed that some articles focused on capturing over-trust in AI [23, 82, 87, 192, 195]; however, under-trust was often overlooked. We posit that calibrated trust requires equal consideration of both scenarios. Appropriate trust necessitates equal consideration of both over-trust and under-trust scenarios because a skewed focus on one aspect can lead to suboptimal outcomes.

Third, it is not clear to what extent behavioral experiments, which account for 70% of experiment designs, especially physiological and empirical measures, can be used as a proxy to capture trust. While behavioral experiments can offer valuable insights into trust-related behaviors, their ability to fully capture the complexity of trust can be unclear due to simplified environments, artificial motivations, lack of context, limited generalizability, and the subjective nature of trust [52].

Thus, we have yet to completely characterize adequate experiment design for capturing appropriate trust. More is needed to fully incorporate the element of risk or vulnerability, to have a clear distinction between reliance and trust, and to address the uneven focus between over-trust and under-trust.

7 Addressing Challenges in Designing for Appropriate Trust

In this section, we first discuss practical challenges and recommendations in experimental design where building appropriate trust is the goal, followed by some novel perspectives that appeared in our corpus analysis. We further delve into some recent developments within the field of explainable AI as well as an emerging critique of the explainability framework, detailing what this could mean for building appropriate trust. Finally, we articulate on the practical implications of our findings for AI practitioners, which includes discussions on overcoming challenges to foster appropriate trust in AI systems.

7.1 Practical Challenges and Recommendations for Designing for Appropriate Trust

While analyzing the text from our corpus, we discovered some open questions and practical challenges of designing for appropriate trust in human-AI interaction. First, what to take into account when deciding whether human's trust in the AI system is over-trust or under-trust? From the reviewed articles, this distinction seems to be primarily based on the AI accuracy—that is, correct or incorrect AI recommendations [198, 203, 205]. We argue that this process of determining where the threshold lies in deciding over- or under-trust cannot be solely about making a right or wrong decision; instead, it should consider multiple aspects. For example, while accuracy indicates human reliance on the AI system's outputs, it does not capture the nuanced nature of trust. Trust involves more than mere reliance; it encompasses perceived reliability, multiple interactions, transparency, and the belief that the AI system has the user's best interests. For instance, a user may rely on an AI-based navigation system when using it for the first time to reach their destination, leading to 100% reliance. However, trusting the system 100% may require interacting with it multiple times in different contexts. Hence, we argue that a comprehensive evaluation of trust should consider a multi-dimensional approach that incorporates both accuracy and factors related to transparency, interpretability, adaptability, longitudinal interactions, user feedback, and the cognitive and emotional aspects of trust. This broader perspective will enable researchers to understand better when human trust in an AI system gears toward over-trust or under-trust [138].

Second, how to calculate appropriate trust for a task with non-binary decision making? In other words, when the decisions are non-binary (e.g., price estimation), it is relatively difficult to identify over- and under-trust at regular time intervals. This could be because it involves a continuous scale of possibilities, making it challenging to define clear boundaries for what constitutes over-trust or under-trust. However, when the decisions are binary, it is easier to assess trust since one can directly compare the outcomes to the binary decisions (e.g., correct or incorrect). In our analysis, we could not find any articles from the reviewed corpus that clarify how to calculate appropriate trust if the decisions are non-binary. We believe that in such cases it is essential to consider a more nuanced approach that takes into account the specific characteristics of the task and the decision-making process, such as by assigning probabilities to different outcomes or decision options.

Third, and relating to the previous point, as AI systems can change over time, how can we measure appropriate trust, or even reliance, as they become moving targets? Consider the automated vehicle that is highly reliable in dry, clean weather but whose performance degrades in rainy conditions, forcing the driver to dynamically adjust their trust. We only find mention of this limitation in five of the articles we reviewed. Further, we could not find reviewed articles addressing how periodicity in the trust gain and loss is affected by the task (i.e., frequency or regularity with which trust is gained or lost in a task), thus we have limited understanding of trust dynamics in real-world long-term interactions. We postulate that a common reason we could not find articles relating to periodicity of trust is because dynamics of trust development and erosion is itself a complex topic that can impact task performance and efficiency. Hence, we need further research on generating empirical evidence, insights, and theoretical frameworks to address the gap in knowledge regarding the influence of task frequency and regularity on the periodicity of trust gain and loss.

Fourth, how appropriate trust is conceptualized and measured across different fields (e.g., healthcare vs. autonomous vehicles) has direct implications for designing and evaluating AI systems. For instance, in healthcare, appropriate trust may entail patients' confidence in AI-assisted diagnosis and treatment recommendations, which necessitates considerations of accuracy, reliability, and ethical standards. However, in autonomous vehicles, appropriate trust may involve users' reliance on the vehicle's decision-making abilities and safety features, requiring assessments of performance, predictability, and risk mitigation strategies. Understanding these divergent conceptualizations and measurement approaches is crucial for tailoring AI systems to meet domain-specific appropriate trust requirements.

Finally, the intricate nature of AI systems and the requisite skills for designing them present significant challenges. Addressing this aspect involves exploring the multifaceted nature of AI design and the diverse skill sets required. For instance, AI designers must possess basic competencies in various domains, such as machine learning, natural language processing, and ethics, to effectively navigate the complexities inherent in developing trustworthy AI systems. Mapping these competencies required for fostering appropriate trust in AI systems can provide valuable insights into the interdisciplinary nature of efforts. This entails identifying and integrating expertise from fields such as computer science, psychology, ethics, and sociology to address the multifaceted challenges associated with trust in AI.

This discussion on the practicalities of implementing appropriate trust also raises questions of developer responsibility, autonomy, and ethical considerations within the AI design process. With the potential of appropriate trust to moderate algorithmic harms realized through misuse, such as biased mortgage approval [116] and hiring decisions [60], the question remains—are AI developers and designers interested in, and in a position to, spearhead the operationalization of appropriate trust? In their work investigating the ethical agency of AI developers, Griffin et al. [68] show that developers may not perceive themselves as agentic as they are in reality and further may not conceptualize the decisions that drive AI tool development as value laden. This work, along with others in the field of responsible AI [120, 131], suggests that research may need to operationalize appropriate trust beyond guidelines and toolkits, as well as account for AI industry culture, to enable real adoption and acceptance in the field and empower developer ethical decision making. Initial work to understand how developers and designers may understand and attempt to employ appropriate trust in their practice represents a starting step along that path [46].

7.2 Novel Perspectives in Designing for Appropriate Trust

We found some distinct perspectives on understanding appropriate trust in AI while analyzing our corpus. First, Chiou and Lee [32] argue that the current approach to studying trust calibration neglects relational aspects of increasingly capable automation and system-level outcomes, such as cooperation and resilience. They adopt a relational framing of trust to address these limitations based on the decision situation, semiotics, interaction sequence, and strategy. They stress that the goal is not to maximize or even calibrate trust but to support a process of trusting through automation responsivity. We resonate with the perspective put forward by the authors; however, to achieve a higher degree of automation responsivity, human values, societal norms, and conflicts are to be studied and implied in the AI systems.

Second, Toreini et al. [184] suggest that we need to study the locus of trust to understand appropriate trust in the systems. They raise the questions such as whether we trust the people who developed the system or the system itself. What purpose are the broader organizations serving? Furthermore, the authors acknowledge the limitations of individuals' capabilities concerning assessing ability and benevolence and propose that individuals accomplish this indirectly by assessing the ability and benevolence of the entity developing the AI.

Finally, among the enormous amount of methods and approaches presented in the review, the work by Collins and Juvina [37] highlights the importance of trust miscalibrations to study appropriate trust. According to the authors, when the need for trust becomes stronger, individuals may stop trusting their previous trusted partners and instead try to establish trust with those they previously distrusted. Studying these exceptions to the principle of trust calibration might be critical for understanding long-term trust calibration in dynamic environments. We believe that this change in trust tactics that is known in human-human interaction is missing in the human-AI interaction studies. Furthermore, we could not find any studies in which humans interact with several AI systems in real life, so this aspect of trust strategies needs to be studied if we wish to learn about how trust miscalibration can be a useful tool to understand appropriate trust in AI systems.

7.3 Is Explainable AI the Answer for Building Appropriate Trust?

The field of explainable AI has seen a few important developments in recent years. For one, visual explanations are increasingly used in human-AI communication [141] for supporting users in decision making. Similarly, counterfactual explanations—explanations that highlight how AI output would change in response to a change in input [175]—have been gaining attention among the research community with the hope that they would enable users a higher degree of control over algorithmic decision making by outlining actionable features of the input data [20]. Furthermore, methods of evaluating explainable AI have grown as well [134], enabling more rigorous validation, benchmarking, and comparison between explainable AI models.

As we have seen in this review, explanations have been heavily utilized in designing for appropriate trust. However, its effectiveness is still questionable in designing for appropriate trust. The challenge mainly centers on the usefulness, usability, and perception of explanations in end users interacting with AI systems. For examples, Miller [125] presented their framework for evaluative AI in response to the perceived failings of explainable AI. They argued that explanations cannot support human judgment in high-stakes situations, as they remove control over the decision making from the human and reduce their likelihood of engaging critically with recommendations. Thus, over- and under-reliance is better moderated by frameworks that focus on exploring the solution space over defending recommendations. The contestable AI [5] framework similarly highlights weaknesses of the explainable AI approach. By shifting focus from human-agent teams and end users to decision subjects, the contestable AI framework reveals contexts where explainability is superfluous at best and actively overwhelming at worst [4]. Under this framework, trust fostered by explanations is always inappropriate unless paired with means through which decision subjects can utilize explanations to contest algorithmic decisions [201].

Within the framework of explainable AI, Sivaraman et al. [169] reveal a 'negotiate' interaction pattern wherein experts selectively implement aspects of AI recommendations, complementing the system's performance with their own domain knowledge, and employing explanations to increase a user's confidence in their own decisions, rather than reflecting on the AI's recommendation. This expanding body of work shows that explainable AI is a complex domain where human-AI expertise overlap [128] and effectiveness of explanations for building appropriate trust is an ongoing research topic.

We recommend that in choosing to calibrate trust through explanations, designers and developers must also confront the preceding challenges of designing explanations and interaction workflows such that they are useful for their target audience. For example, Bertrand [20] presents several directions for future explanation design that better align with the complexity of human cognitive factors. Their recommendations include interactive explanations, explanations framed as questions, multimodal explanations, and adaptive explanation complexity.

7.4 Implications for AI Practitioners

We discussed in this review how the discord and diversity of understanding around appropriate trust in the literature may have negative downstream effects on both researchers and designers. In this section, we focus on how the takeaways of our review may guide practitioners through this field. Research has already begun to investigate how concepts like appropriate [46], responsible [109], or general [172] trust can be conceptualized for practical development work. This work highlighted the complexity and ethical considerations of integrating these concepts into design and development processes but also highlighted enabling trust judgments [109], reducing triggers for over/under-trust, and centering design process around human values and ethics as concrete actions for AI practitioners.

Based on the results of this review, we further recommend designers and developers to prioritize understanding which facet of appropriate trust is most relevant to their work (responsible, contractual, justified, etc.) early on in the design process to ensure that their measures of trust and their design choices are fit to purpose. Moreover, akin to Sousa et al. [172], we emphasize the importance of designers understanding both risks associated with their system and the perception of those risk in the end user population. As we stated previously, risk and vulnerability are a large part of trust judgment formation, as well as the need for trust, and thus it is important to characterize throughout the development process.

In addition, the introduction of measurable constructs for appropriate trust and considerations of AI system integrity and explanations offers practical elements for the evaluation and refinement stages of the design cycle. Designers can use these constructs as measurable metrics to assess the effectiveness of appropriate trust-building features and refine their designs based on empirical insights gathered from user interactions. A useful tool for practitioners would be using design patterns proposed by Google's PAIR guidebook [147] for trust calibration, which is based on user interactions. Some of these design patterns include setting right expectations, letting users supervise automation, and adding context from human sources that have been applied to a real case study in Google Flights.[10]

Finally, addressing ethical implications and biases throughout the design cycle at every stage prompts designers to incorporate ethical considerations from the outset and continually reassess the societal impact of their designs.

8 Limitations and Ideas for Future Research

Despite the systematic review's comprehensive analysis of the state of the art in fostering appropriate trust, there are several limitations to this study that need to be acknowledged.

First, while we included studies from a number of relevant disciplines (refer to our search string in Section 3.1), it is possible that some relevant studies were missed. Additionally, we only focused on studies published in English, which may have led to language bias. Future reviews should consider including studies in other languages to increase the generalizability of the findings.

Second, our mapping of concepts related to appropriate trust—based on beliefs, desires, and intentions—is just one of many possible frameworks. Future research should aim to develop a clear and concise mapping from a multidisciplinary perspective. Additionally, the absence of empirical validation for the proposed BIA framework is a limitation. Future work should include empirical studies to test this framework's applicability in real-world settings, providing valuable insights and refining the model.

Third, the search period for this review covered publications from 2012 until June 2022, potentially excluding recent studies in this rapidly evolving field. Moreover, while this review identified current trends and potential research gaps in fostering appropriate trust in AI systems, it did not explore the development of new approaches or design techniques. Further research is necessary to incorporate more recent findings and to delve deeper into innovative strategies for enhancing appropriate trust in AI systems.

[10] https://medium.com/people-ai-research/pair-guidebook-google-flights-case-study-1ba8c7352141

8.1 Ideas for Future Research

The first evident step of future work is to employ the findings of this work by the academic researchers, AI system designers, and UX researchers in real-world settings. AI system designers and UX researchers working within large organizations can test the findings on a larger scale. Going forward, future work can aim to refine the formalization and implement a method to evaluate the definition of appropriate trust. In particular, one can conduct user studies to evaluate our notions regarding beliefs of beliefs in an experimental setting. This can both help us understand how trust beliefs are formed in humans and how agents can appropriately use these beliefs to improve teamwork.

Future research can focus on the development of a clear and concise mapping of definitions of appropriate trust and related concepts from a multidisciplinary perspective. Based on the identified research gaps in our study, future work should aim for (a) a clear definition of appropriate trust; (b) defining, distinguishing, treating, and measuring concepts related to appropriate trust as independent concepts; (c) focus on integrity and benevolence of the AI systems; and (d) work more closely with AI developers and designers to understand the facilitators and barriers to appropriate trust. Finally, future work should incorporate recent developments in the field of building appropriate trust, such as large language models, and empirically validate our proposed BIA mapping.

9 Summary of Systematic Review

This subsection aims to summarize the current trends, challenges, and recommendations concerning the definitions, conceptualizations, measures, implications of measures, and results for establishing appropriate trust in AI systems (Table 3). By addressing the evolving trends, inherent challenges, and potential solutions, we aim to enrich the overall understanding of the topic, enabling readers to grasp the broader context and implications associated with building appropriate trust in AI systems.

Our aim with this summary is to provide a well-structured gateway for both experts and newcomers to understand the trends and challenges with an actionable set of recommendations. With

Table 3. Detailed Summary of Current Trends, Challenges, and Recommendations Based on the Results of the Systematic Review

Section	Current Trends	Challenges	Recommendations
Definitions	(1) 75.3% (n = 312) of articles from our corpus that were sought for retrieval did not provide a definition of appropriate trust or a related concept.[11]	(1) A lack of clear definition creates confusion among readers from different backgrounds.	(1) Provide a clear definition of appropriate trust or a related concept.
	(2) Of the articles that provided a definition in our final corpus, 25% (n = 16) of them provided new definitions that were often not related to prior works (see Table 1).	(2) A variety of definitions inherent to multidisciplinary fields without relating it to other fields can cause misunderstanding to the reader.	(2) We need to converge in the future to establish common ground to define what appropriate trust means in human-AI interaction.
Conceptualization	(1) Many types of appropriate trust concepts are only sometimes explicitly distinguished. For example, the differences between optimal trust, well-placed trust, meaningful trust, and justified trust, and so on, are often unclear and used interchangeably.	(1) A plethora of concepts related to appropriate trust is causing the HCI community to diverge in multiple ways. For example, this unclear connotation of similar concepts often creates confusion among researchers, especially new graduate student.	(1) Related concepts that are distinct from the goal of appropriate trust should be defined, distinguished, treated, and measured as independent concepts. For example, warranted trust and contractual trust have different goals than appropriate trust.

(Continued)

Table 3. Continued

Section	**Current Trends**	**Challenges**	**Recommendations**
Conceptualization	(2) Interchangeable use of *appropriate trust* with *appropriate reliance.*	(2) A core distinction in philosophy, which is often neglected in the empirical HCI literature, regards trust and reliance as distinct concepts.	(2) We propose the distinction of Hoff and Bashir [75], where trust is the belief that "an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability," and the reliance distinction of Lee and See [106], "a discrete process of engaging or disengaging."
	(3) A total of 38% of articles in our final corpus conceptualize appropriate trust or related concepts as the measure of alignment between the perceived and actual ability of the system.	(3) To explore the extent and magnitude of how the trustworthiness properties of machines, beyond their ability, impact trust. For example, what do we mean by the integrity of a machine, and how can we measure it?	(3) We must focus on measuring less studied dimensions of trustworthiness (i.e., integrity and benevolence) to understand human trust in AI systems.
Measures	(1) A total of 40% (n = 26) of articles in our final corpus study appropriate trust in binary decision-making tasks (i.e., to [not] follow an [in]correct AI recommendation).	(1) To develop strategies for building appropriate trust in AI systems that continuously make decisions, such as in price estimation. In addition, the potential issues that arise when the AI model targets diverge from human goals.	(1) We need to investigate new measures to assess dynamic trust in practice. For example, we can use situational reference points to keep aligning the goal [29].
Results	(1) Around 37% of reviewed articles report the effect of improving system transparency for establishing appropriate trust in human-AI interaction.	(1) A disadvantage of single focus on improving system transparency requires ground truth, which is often not available or there is really no "ground" at all.	(1) Include post-experiment surveys or interviews where the participants can give their impressions on the trustworthiness of the AI agents.
	(2) In 43% of the included articles, the objective of the designed task had direct influence on the results of appropriate trust in human-AI interaction.	(2) If the objective of the task to foster appropriate trust in the AI agent is built around improving the fairness of the AI agent, then the results will be different compared to the objective of improving the accuracy.	(2) Ensure control of the initial participants' expectations about the AI system, and report results with scientific rigor about how the design of the task may have influenced human trust.
Implication of the measures	(1) A total of 45% of articles involving a user study focused on detecting over-trust in AI, and under-trust in AI systems is often overlooked.	(1) Under-trust in AI systems is a common challenge.	(1) Investigate and adopt methodologies from social sciences and psychology to study under trust in AI [92].
	(2) Around 10% of articles in our corpus follow some already established guidelines to design for fostering appropriate trust.	(2) There are multiple guidelines from academia and industrial organizations outlining trust calibration principles that AI-based systems should adopt. However, there is less effort that has been put in translating those principles into practice.	(2) Adopt established guidelines while designing a user study, and report if those guidelines did not scale for the user study.
	(3) Locus of trust in the AI systems: are we trusting the people who developed the system is unexplored?	(3) Identify and explore the fundamental correlations between appropriate trust in AI systems and the manufacturers of AI.	(3) Adopt the recommendation by Toreini et al. [184] on analyzing factors such as the transparency of the AI development process, the track record of the manufacturer in delivering trustworthy AI, and the level of accountability and responsibility taken by the manufacturer for the AI's outcomes.

these recommendations, we make an attempt to connect all sections of this article to provide broader context and implications of building appropriate trust in AI.

10 Conclusion

Appropriate trust in AI systems is crucial for effective collaboration between humans and AI systems. Various approaches have been taken to build and assess appropriate trust in AI systems in the past. This article provided a comprehensive understanding of the field with a systematic review outlining different definitions of appropriate trust, methods to achieve it, results of those methods, and a detailed discussion on challenges and future considerations. Through this review of current practices in building appropriate trust, we identified the challenge for a single definition of appropriate trust and the ambiguity surrounding related concepts such as warranted trust, appropriate reliance, and justified trust.

Our review proposed a BIA mapping to study commonalities and differences among different concepts related to appropriate trust. We found three common measurement techniques to measure appropriate trust as perceived, demonstrated, and mixed. In addition, multiple domains and associated tasks were used to study appropriate trust. Furthermore, our analysis of articles revealed four common methods for building appropriate trust, such as transparency, perception, guidelines, and studying the continuum of trust.

In synopsis, the review highlights what approaches exist to build appropriate trust and how successful they seem to be. We discussed the challenges and potential gaps in studying appropriate trust, which presents opportunities for future research such as discord and diversity in defining appropriate trust or a strong focus on capability. Overall, this article provided (a) a comprehensive overview of the current state of research on appropriate trust in AI by studying measures, tasks, methods, and results of those methods; (b) a BIA mapping of appropriate trust and its related concepts; and (c) a set of recommendations for fostering appropriate trust in AI based on current trends and challenges. With these contributions, we can advance our understanding of designing for appropriate trust in human-AI interaction by taking a step closer toward responsible AI [81].

Acknowledgments

We thank Mo, Pei-Yu Chen, Enrico Liscio, Carolina Jorge, Ruben Verhagen and anonymous reviewers for their contribution in iterations of the manuscript.

References

[1] Kumar Akash, Neera Jain, and Teruhisa Misu. 2020. Toward adaptive trust calibration for level 2 driving automation. In *Proceedings of the 2020 International Conference on Multimodal Interaction.* 538–547.

[2] Arjun Akula, Shuai Wang, and Song-Chun Zhu. 2020. CoCoX: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2594–2601.

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

[4] Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem. 2023. Tensions in transparent urban AI: Designing a smart electric vehicle charge point. *AI & Society* 38, 3 (June 2023), 1049–1065. https://doi.org/10.1007/s00146-022-01436-9

[5] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2023. Contestable AI by design: Towards a framework. *Minds and Machines* 33, 4 (Dec. 2023), 613–639. https://doi.org/10.1007/s11023-022-09611-z

[6] Basel Alhaji, Michael Prilla, and Andreas Rausch. 2021. Trust dynamics and verbal assurances in human robot physical collaboration. *Frontiers in Artificial Intelligence* 4 (2021), 703504. https://doi.org/10.3389/frai.2021.703504

[7] Fatemeh Alizadeh, Oleksandra Vereschak, Dominik Pins, Gunnar Stevens, Gilles Bailly, and Baptiste Caramiaux. 2022. Building appropriate trust in human-AI interactions. In *Proceedings of the 20th European Conference on Computer-Supported Cooperative Work (ECSCW'22)*, Vol. 6.

[8] James F. Allen. 1984. Towards a general theory of action and time. *Artificial Intelligence* 23, 2 (1984), 123–154.

[9] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems*. 1–13.

[10] Apple. 2020. Human Interface Guidelines. Retrieved February 14, 2023 from https://developer.apple.com/design/human-interface-guidelines/guidelines/overview/

[11] Jackie Ayoub, Lilit Avetisyan, Mustapha Makki, and Feng Zhou. 2021. An investigation of drivers' dynamic situational trust in conditionally automated driving. *IEEE Transactions on Human-Machine Systems* 52, 3 (2021), 501–511.

[12] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, X. Jessie Yang, Lionel P. Robert, and Dawn M. Tilbury. 2020. Context-adaptive management of drivers' trust in automated vehicles. *IEEE Robotics and Automation Letters* 5, 4 (2020), 6908–6915.

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

[14] Lisanne Bainbridge. 1983. Ironies of automation. In *Analysis, Design and Evaluation of Man–Machine Systems*. Elsevier, 129–135.

[15] Gagan Bansal, Alison Marie Smith-Renner, Zana Buçinca, Tongshuang Wu, Kenneth Holstein, Jessica Hullman, and Simone Stumpf. 2022. Workshop on trust and reliance in AI-human teams (TRAIT). In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA'22)*. ACM, New York, NY, USA, Article 116, 6 pages. https://doi.org/10.1145/3491101.3503704

[16] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[17] David Barnard. 2016. Vulnerability and trustworthiness: Polestars of professionalism in healthcare. *Cambridge Quarterly of Healthcare Ethics* 25, 2 (2016), 288–300.

[18] Jeff A. Bauhs and Nancy J. Cooke. 1994. Is knowing more really better? Effects of system development information in human-expert system interactions. In *Conference Companion on Human Factors in Computing Systems (CHI'94)*. 99–100.

[19] Tom L. Beauchamp. 1995. Moral Prejudices: Essays on Ethics. *Hastings Center Report* 25, 4 (1995), 36–37.

[20] Astrid Bertrand. 2024. *Misplaced Trust in AI: The Explanation Paradox and the Human-Centric Path: A Characterisation of the Cognitive Challenges to Appropriately Trust Algorithmic Decisions and Applications in the Financial Sector*. Ph.D. Dissertation. Institut Polytechnique de Paris. https://theses.hal.science/tel-04661844

[21] Philip Bobko, Leanne Hirshfield, Lucca Eloy, Cara Spencer, Emily Doherty, Jack Driscoll, and Hannah Obolsky. 2022. Human-agent teaming and trust calibration: A theoretical framework, configurable testbed, empirical illustration, and implications for the development of adaptive systems. *Theoretical Issues in Ergonomics Science* 24, 3 (2022), 310–334.

[22] Michael Bratman. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, USA.

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

[24] Meghan Madhavi Burke. 2016. Shraddha: A special kind of Trust. *Healing Arts Centre*. Retrieved September 25, 2024 from https://edu.nl/muppx

[25] Davide Calvaresi, Kevin Appoggetti, Luca Lustrissimini, Mauro Marinoni, Paolo Sernani, Aldo Franco Dragoni, and Michael Schumacher. 2018. Multi-agent systems' negotiation protocols for cyber-physical systems: Results from a systematic literature review. In *Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART'18)*. 224–235.

[26] Cristiano Castelfranchi and Rino Falcone. 1998. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proceedings of the International Conference on Multi Agent Systems*. IEEE, 72–79.

[27] Christiano Castelfranchi and Rino Falcone. 2010. *Trust Theory: A Socio-Cognitive and Computational Model*. John Wiley & Sons.

[28] George Charalambous, Sarah Fletcher, and Philip Webb. 2016. Development of a human factors roadmap for the successful implementation of industrial human-robot collaboration. In *Advances in Ergonomics of Manufacturing: Managing the Enterprise of the Future: Proceedings of the AHFE 2016 International Conference on Human Aspects of Advanced Manufacturing, July 27–31, 2016, Walt Disney World®, Florida, USA*. Advances in Intelligent Systems and Computing, Vol. 490. Springer, 195–206.

[29] Gilad Chen and John E. Mathieu. 2008. Goal orientation dispositions and performance trajectories: The roles of supplementary and complementary situational inducements. *Organizational Behavior and Human Decision Processes* 106, 1 (2008), 21–38.

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

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

[32] Erin K. Chiou and John D. Lee. 2023. Trusting automation: Designing for responsivity and resilience. *Human Factors* 65, 1 (2023), 137–165.

[33] Jin-Hee Cho, Kevin Chan, and Sibel Adali. 2015. A survey on trust modeling. *ACM Computing Surveys* 48, 2 (2015), 1–40.

[34] Sanghyun Choo and Chang S. Nam. 2022. Detecting human trust calibration in automation: A convolutional neural network approach. *IEEE Transactions on Human-Machine Systems* 52, 4 (2022), 774–783.

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

[36] Marvin S. Cohen, Raja Parasuraman, and Jared T. Freeman. 1998. Trust in decision aids: A model and its training implications. In *Proceedings of the 1998 Command and Control Research and Technology Symposium*. 1–37.

[37] Michael G. Collins and Ion Juvina. 2021. Trust miscalibration is sometimes necessary: An empirical study and a computational model. *Frontiers in Psychology* 12 (2021), 690089.

[38] Sven Coppers, Davy Vanacken, and Kris Luyten. 2020. FortNIoT: Intelligible predictions to improve user understanding of smart home behavior. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–24.

[39] David Danks. 2019. The value of trustworthy AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES'19)*. ACM, New York, NY, USA, 521–522. https://doi.org/10.1145/3306618.3314228

[40] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*. Auerbach Publications, 296–299.

[41] Ewart de Visser, Brian Kidwell, John Payne, Li Lu, James Parker, Nathan Brooks, Timur Chabuk, Sarah Spriggs, Amos Freedy, Paul Scerri, and Raja Parasuraman. 2013. Best of both worlds: Design and evaluation of an adaptive delegation interface. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57, 1 (2013), 255–259. https://doi.org/10.1177/1541931213571056

[42] Ewart J. de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. 2014. A design methodology for trust cue calibration in cognitive agents. In *Virtual, Augmented and Mixed Reality: Designing and Developing Virtual and Augmented Environments*, Randall Shumaker and Stephanie Lackey (Eds.). Springer International Publishing, Cham, 251–262.

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

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

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

[46] Chadha Degachi, Siddharth Mehrotra, Mireia Yurrita Semperena, Evangelos Niforatos, and Myrthe Tielman L. 2024. Practising appropriate trust in human-centred AI design. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. https://doi.org/10.1145/3613905.3650825

[47] Morton Deutsch. 1958. Trust and suspicion. *Journal of Conflict Resolution* 2, 4 (1958), 265–279.

[48] Pierre P. Duez, Michael J. Zuliani, and Greg A. Jamieson. 2006. Trust by design: Information requirements for appropriate trust in automation. In *Proceedings of the 2006 Conference of the Center for Advanced Studies on Collaborative Research*. 9–es.

[49] Shemuel Noah Eisenstadt and Luis Roniger. 1984. *Patrons, Clients and Friends: Interpersonal Relations and the Structure of Trust in Society*. Themes in the Social Sciences. Cambridge University Press.

[50] Fredrick Ekman, Mikael Johansson, and Jana Sochor. 2017. Creating appropriate trust in automated vehicle systems: A framework for HMI design. *IEEE Transactions on Human-Machine Systems* 48, 1 (2017), 95–101.

[51] Mica R. Endsley and David B. Kaber. 1999. Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics* 42, 3 (1999), 462–492. https://doi.org/10.1080/001401399185595

[52] Thorsten M. Erle and Michael K. Zürn. 2020. Illusory trust: Kanizsa shapes incidentally increase trust and willingness to invest. *Journal of Behavioral Decision Making* 33, 5 (2020), 671–682.

[53] Anthony M. Evans, Ursula Athenstaedt, and Joachim I. Krueger. 2013. The development of trust and altruism during childhood. *Journal of Economic Psychology* 36 (2013), 82–95.

[54] Rino Falcone and Cristiano Castelfranchi. 2001. Social trust: A cognitive approach. In *Trust and Deception in Virtual Societies.* Springer, 55–90.

[55] Edward A. Feigenbaum. 1971. Computer Professionals against ABMs: Newsletters and press clippings: 1970–71—Organization of computer experts calls ABM project a dangerous mistake. *Stanford Libraries.* Retrieved September 25, 2024 from https://exhibits-lb.stanford.edu/cs/catalog/pc764bb9418

[56] Andrea Ferrario and Michele Loi. 2022. How explainability contributes to trust in AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT'22).* ACM, New York, NY, USA, 1457–1466. https://doi.org/10.1145/3531146.3533202

[57] Gavin D. Flood. 1996. *An Introduction to Hinduism.* Cambridge University Press.

[58] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. In *Proceedings of the 2007 International Symposium on Collaborative Technologies and Systems.* IEEE, 106–114.

[59] Markus Freitag and Paul C. Bauer. 2016. Personality traits and the propensity to trust friends and strangers. *Social Science Journal* 53, 4 (2016), 467–476.

[60] Gary D. Friedman and Thomas McCarthy. 2020. Employment law red flags in the use of artificial intelligence in hiring. *ABA.* Retrieved September 25, 2024 from https://www.americanbar.org/groups/business_law/resources/business-law-today/2020-october/employment-law-red-flags/

[61] Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. 1999. The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages.* Lecture Notes in Computer Science, Vol. 1555. Springer, 1–10.

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

[63] Felix Gille, Anna Jobin, and Marcello Ienca. 2020. What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine* 1 (2020), 100001.

[64] Nicole Gillespie. 2011. Measuring trust in organizational contexts: an overview of survey-based measures. In *Handbook of Research Methods on Trust*, F. Lyon, G. Mollering, and M. Saunders (Eds.). Edward Elgar Publishing, 175–188.

[65] Nicole Gillespie, Steven Lockey, Caitlin Curtis, Javad Pool, and Ali Akbari. 2023. *Trust in Artificial Intelligence: A Global Study.* The University of Queensland and KPMG Australia.

[66] Jane Goudge and Lucy Gilson. 2005. How can trust be investigated? Drawing lessons from past experience. *Social Science & Medicine* 61, 7 (2005), 1439–1451.

[67] Gregory M. Gremillion, Jason S. Metcalfe, Amar R. Marathe, Victor J. Paul, James Christensen, Kim Drnec, Benjamin Haynes, and Corey Atwater. 2016. Analysis of trust in autonomy for convoy operations. In *Micro-and Nanotechnology Sensors, Systems, and Applications VIII*, Vol. 9836. SPIE, 356–365.

[68] Tricia A. Griffin, Brian Patrick Green, and Jos V. M. Welie. 2023. The ethical agency of AI developers. *AI and Ethics.* Published Online, January 9, 2023. https://doi.org/10.1007/s43681-022-00256-3

[69] Yaohui Guo and X. Jessie Yang. 2021. Modeling and predicting trust dynamics in human–robot teaming: A Bayesian inference approach. *International Journal of Social Robotics* 13, 8 (2021), 1899–1909.

[70] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53, 5 (2011), 517–527.

[71] Katherine Hawley. 2014. Trust, distrust and commitment. *Noûs* 48, 1 (2014), 1–20.

[72] Katherine Hawley. 2017. Trustworthy groups and organizations. In *The Philosophy of Trust.* Oxford Academic, 230–250.

[73] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 2013. Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications.* 210–217.

[74] Sarita Herse, Jonathan Vitale, Benjamin Johnston, and Mary-Anne Williams. 2021. Using trust to determine user decision making and task outcome during a human-agent collaborative task. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction.* 73–82.

[75] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434.

[76] Robert R. Hoffman. 2017. A taxonomy of emergent trusting in the human–machine relationship. In *Cognitive Systems Engineering: The Future for a Changing World.* CRC Press, Boca Raton, FL, USA, 137–164.

[77] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 4 (2019), e1312.

[78] Sandy H. Huang, Kush Bhatia, Pieter Abbeel, and Anca D. Dragan. 2018. Establishing appropriate trust via critical states. In *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'18)*. IEEE, 3929–3936.

[79] Ronald Hurst and Leslie R. Hurst. 1982. *Pilot Error: The Human Factors*. Jason Aronson.

[80] IBM. 2020. IBM Design for AI. Retrieved February 14, 2023 from https://www.ibm.com/design/ai/

[81] IEEE. 2017. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems. Retrieved September 25, 2024 from https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

[82] Brett Israelsen, Peggy Wu, Katharine Woodruff, Gianna Avdic-McIntire, Andrew Radlbeck, Angus McLean, Patrick "Dice" Highland, Thomas "Mach" Schnell, and Daniel "Animal" Javorsek. 2021. Introducing SMRTT: A structural equation model of multimodal real-time trust. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI'21 Companion)*. 126–130.

[83] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 624–635.

[84] Theodore Jensen, Yusuf Albayram, Mohammad Maifi Hasan Khan, Ross Buck, Emil Coman, and Md. Abdullah Al Fahim. 2018. Initial trustworthiness perceptions of a drone system based on performance and process information. In *Proceedings of the 6th International Conference on Human-Agent Interaction*. 229–237.

[85] Theodore Jensen, Mohammad Maifi Hasan Khan, and Yusuf Albayram. 2020. The role of behavioral anthropomorphism in human-automation trust calibration. In *Artificial Intelligence in HCI*. Lecture Notes in Computer Science, Vol. 12217. Springer, 33–53.

[86] Theodore Jensen, Mohammad Maifi Hasan Khan, Md. Abdullah Al Fahim, and Yusuf Albayram. 2021. Trust and anthropomorphism in tandem: The interrelated nature of automated agent appearance and reliability in trustworthiness perceptions. In *Proceedings of the 2021 Designing Interactive Systems Conference*. 1470–1480.

[87] Wolfgang Jentner, Rita Sevastjanova, Florian Stoffel, Daniel A. Keim, Jürgen Bernard, and Mennatallah El-Assady. 2018. Minions, sheep, and fruits: Metaphorical narratives to explain artificial intelligence and build trust. In *Proceedings of the Workshop on Visualization for AI Explainability at IEEE*.

[88] Craig J. Johnson, Mustafa Demir, Nathan J. McNeese, Jamie C. Gorman, Alexandra T. Wolff, and Nancy J. Cooke. 2021. The impact of training on human–autonomy team communications and trust calibration. *Human Factors*. Published Online, October 1, 2021.

[89] Christopher Jones. 2020. Law enforcement use of facial recognition: Bias, disparate impacts on people of color, and the need for federal legislation. *North Carolina Journal of Law & Technology* 22 (2020), 777.

[90] Karen Jones. 2018. The politics of credibility. In *A Mind of One's Own*. Routledge, 154–176.

[91] C. Centeio Jorge, Siddharth Mehrotra, M. L. Tielman, and C. M. Jonker. 2021. Trust should correspond to trustworthiness: A formalization of appropriate mutual trust in human-agent teams. In *Proceedings of the 2021 22nd International Trust Workshop*.

[92] Audun Jøsang and Stéphane Lo Presti. 2004. Analysing the relationship between risk and trust. In *Proceedings of the International Conference on Trust Management*. 135–145.

[93] Poornima Kaniarasu, Aaron Steinfeld, Munjal Desai, and Holly Yanco. 2013. Robot confidence and trust alignment. In *Proceedings of the 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI'13)*. IEEE, 155–156.

[94] Arnon Keren. 2014. Trust and belief: A preemptive reasons account. *Synthese* 191, 12 (2014), 2593–2615.

[95] Siddartha Khastgir, Stewart Birrell, Gunwant Dhadyalla, and Paul Jennings. 2018. Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation Research Part C: Emerging Technologies* 96 (2018), 290–303. https://doi.org/10.1016/j.trc.2018.07.001

[96] H. E. Knee and J. C. Schryver. 1989. *Operator Role Definition and Human System Integration*. Technical Report. Oak Ridge National Laboratory, Oak Ridge, TN, USA.

[97] Roderick M. Kramer and Tom R. Tyler. 1995. *Trust in Organizations: Frontiers of Theory and Research*. SAGE Publications, Los Angeles, CA, USA.

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

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

[100] Vivian Lai, Han Liu, and Chenhao Tan. 2020. “Why is “Chicago” deceptive?” Towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[101] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.

[102] Himabindu Lakkaraju and Osbert Bastani. 2020. “How do I fool you?”: Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.

[103] Kathryn Ann Lambe, Gary O’Reilly, Brendan D. Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: A systematic review. *BMJ Quality & Safety* 25, 10 (2016), 808–820.

[104] Christian Lebiere, Leslie M. Blaha, Corey K. Fallon, and Brett Jefferson. 2021. Adaptive cognitive mechanisms to maintain calibrated trust and reliance in automation. *Frontiers in Robotics and AI* 8 (2021), 652776.

[105] John D. Lee and Neville Moray. 1994. Trust, self-confidence, and operators’ adaptation to automation. *International Journal of Human-Computer Studies* 40, 1 (1994), 153–184.

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

[107] Roy J. Lewicki and Barbara B. Bunker. 1996. Developing and maintaining trust in work relationships. In *Trust in Organizations: Frontiers of Theory and Research*. SAGE Publications, Los Angeles, CA, USA, 114–139.

[108] Yugang Li, Baizhou Wu, Yuqi Huang, and Shenghua Luan. 2024. Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-AI trust. *Frontiers in Psychology* 15 (2024), 1382693.

[109] Q. Vera Liao and S. Shyam Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1257–1268.

[110] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), Article 408, 45 pages. https://doi.org/10.1145/3479552

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

[112] Fergus Lyon, Guido Mšllering, and Mark N. K. Saunders. 2015. *Handbook of Research Methods on Trust*. Edward Elgar Publishing.

[113] Stefanie M. Faas, Johannes Kraus, Alexander Schoenhals, and Martin Baumann. 2021. Calibrating pedestrians’ trust in automated vehicles: Does an intent display in an external HMI support trust calibration and safe crossing behavior? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[114] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI ’23)*. ACM, New York, NY, USA, 1–19. https://doi.org/10.1145/3544548.3581058

[115] Stephen Marsh and Mark R. Dibben. 2005. Trust, untrust, distrust and mistrust—An exploration of the dark(er) side. In *Trust Management*. Lecture Notes in Computer Science, Vol. 3477. Springer, 17–33.

[116] Emmanuel Martinez and Lauren Kirchner. 2021. The Secret Bias Hidden in Mortgage-Approval Algorithms. Retrieved September 25, 2024 from https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms

[117] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *Academy of Management Review* 20, 3 (1995), 709–734.

[118] Maranda McBride and Shona Morgan. 2010. Trust calibration for automated decision aids. *Institute for Homeland Security Solutions* 2010 (2010), 1–11.

[119] Patricia L. McDermott and Ronna N. ten Brink. 2019. Practical guidance for evaluating calibrated trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications, Los Angeles, CA, USA, 362–366.

[120] Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. 2018. Does ACM’s code of ethics change ethical decision making in software development? In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE’18)*. ACM, New York, NY, USA, 729–733. https://doi.org/10.1145/3236024.3264833

[121] Siddharth Mehrotra. 2021. Modelling trust in human-AI interaction: Doctoral consortium. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS’21)*. 1826–1828.

[122] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. Integrity-based explanations for fostering appropriate trust in AI agents. *ACM Transactions on Interactive Intelligent Systems* 14, 1 (2024), Article 4, 36 pages. https://doi.org/10.1145/3610578 Just Accepted.

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

[124] Tim Miller. 2022. Are we measuring trust correctly in explainability, interpretability, and transparency research? *arXiv preprint arXiv:2209.00651* (2022).

[125] Tim Miller. 2023. Explainable AI is dead, long live Explainable AI! Hypothesis-driven decision support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

[126] Alexander G. Mirnig, Philipp Wintersberger, Christine Sutter, and Jürgen Ziegler. 2016. A framework for analyzing and calibrating trust in automated vehicles. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 33–38.

[127] Agata Mirowska. 2020. AI evaluation in selection: Effects on application and pursuit intentions. *Journal of Personnel Psychology* 19, 3 (2020), 142–149.

[128] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligence Systems* 11, 3-4 (Sept. 2021), Article 45, 45 pages. https://doi.org/10.1145/3387166

[129] Xiaomin Mou. 2019. *Artificial Intelligence: Investment Trends and Selected Industry Uses*. Brief Report. International Finance Corporation.

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

[131] Luke Munn. 2023. The uselessness of AI ethics. *AI and Ethics* 3, 3 (Aug. 2023), 869–877. https://doi.org/10.1007/s43681-022-00209-w

[132] Mohammad Naiseh, Reem S. Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Nudging through friction: An approach for calibrating trust in explainable AI. In *Proceedings of the 2021 8th International Conference on Behavioral and Social Computing (BESC'21)*. IEEE, 1–5.

[133] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Explainable recommendation: When design meets trust calibration. *World Wide Web* 24, 5 (Sept. 2021), 1857–1884. https://doi.org/10.1007/s11280-021-00916-0

[134] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys* 55, 13s (July 2023), Article 295, 42 pages. https://doi.org/10.1145/3583558

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

[136] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 427–436.

[137] David J. Niedober, Nhut T. Ho, Gina Masequesmay, Kolina Koltai, Mark Skoog, Artemio Cacanindin, Walter Johnson, and Joseph B. Lyons. 2014. Influence of cultural, organizational and automation factors on human-automation trust: A case study of Auto-GCAS engineers and developmental history. In *Human-Computer Interaction: Applications and Services*. Lecture Notes in Computer Science, Vol. 8512. Springer, 473–484.

[138] Bart Nooteboom. 2013. Trust and innovation. In *Handbook of Advances in Trust Research*, Reinhard Bachmann and Akbar Zaheer (Eds.). Edward Elgar Publishing, Northampton, MA, USA, 106–122.

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

[140] Onora O'Neill. 2002. *Autonomy and Trust in Bioethics*. Cambridge University Press.

[141] Jeroen Ooge and Katrien Verbert. 2022. Explaining artificial intelligence with tailored interactive visualisations. In *Companion Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI'22 Companion)*. ACM, New York, NY, USA, 120–123. https://doi.org/10.1145/3490100.3516481

[142] Scott Ososky, David Schuster, Elizabeth Phillips, and Florian G. Jentsch. 2013. Building appropriate trust in human-robot teams. In *Proceedings of the 2013 AAAI Spring Symposium Series*.

[143] Elinor Ostrom. 1998. A behavioral approach to the rational choice theory of collective action: Presidential address, American Political Science Association, 1997. *American Political Science Review* 92, 1 (1998), 1–22.

[144] Margit E. Oswald and Corina T. Ulshöfer. 2017. Cooperation and distrust—A contradiction? In *Social Dilemmas, Institutions, and the Evolution of Cooperation*, Ben Jann and Wijtek Przepiorka (Eds.). De Gryter, 357–372.

[145] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews* 5 (2016), 1–10.

[146] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjorn Hrobjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery* 88 (2021), 105906.

[147] Google PAIR. 2019. People + AI Guidebook. Retrieved February 14, 2023 from https://pair.withgoogle.com/guidebook

[148] Raja Parasuraman and Evan A. Byrne. 2003. Automation and human performance in aviation. In *Principles and Practice of Aviation Psychology.* CRC Press, Boca Raton, FL, 311–356.

[149] Raja Parasuraman and Christopher A. Miller. 2004. Trust and etiquette in high-criticality automated systems. *Communications of the ACM* 47, 4 (2004), 51–55.

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

[151] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–52.

[152] Sareh Pouryousefi and Jonathan Tallant. 2022. Empirical and philosophical reflections on trust. *Journal of the American Philosophical Association* 9, 3 (2022), 1–21.

[153] Anand S. Rao and Michael P. Georgeff. 1995. BDI agents: From theory to practice. In *Proceedings of the 1st International Conference on Multiagent Systems (ICMAS'95).* 312–319.

[154] P. L. Patrick Rau, Ye Li, and Dingjun Li. 2009. Effects of communication style and culture on ability to accept recommendations from robots. *Computers in Human Behavior* 25, 2 (2009), 587–595.

[155] John K. Rempel, John G. Holmes, and Mark P. Zanna. 1985. Trust in close relationships. *Journal of Personality and Social Psychology* 49, 1 (1985), 95.

[156] René Riedl. 2022. Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. *Electronic Markets* 32, 4 (2022), 2021–2051.

[157] Jens Riegelsberger, M. Angela Sasse, and John D. McCarthy. 2005. The mechanics of trust: A framework for research and design. *International Journal of Human-Computer Studies* 62, 3 (2005), 381–422.

[158] Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of Management Review* 23, 3 (1998), 393–404.

[159] Siby Samuel, William J. Horrey, and Donald L. Fisher. 2015. A predictive model of driver response in an autonomous environment. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 59. SAGE Publications, Los Angeles, CA, USA, 1671–1675.

[160] Tracy Sanders, Kristin E. Oleson, Deborah R. Billings, Jessie Y. C. Chen, and Peter A. Hancock. 2011. A model of human-robot trust: Theoretical model development. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 55. SAGE Publications, Los Angeles, CA, USA, 1432–1436.

[161] Kristin E. Schaefer, Jessie Y. C. Chen, James L. Szalma, and Peter A. Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors* 58, 3 (2016), 377–400.

[162] John Schaubroeck, Simon S. K. Lam, and Ann Chunyan Peng. 2011. Cognition-based and affect-based trust as mediators of leader behavior influences on team performance. *Journal of Applied Psychology* 96, 4 (2011), 863.

[163] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI'23).* ACM, New York, NY, USA, 410–422. https://doi.org/10.1145/3581641.3584066

[164] Nadine Schlicker and Markus Langer. 2021. Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In *Proceedings of Mensch und Computer 2021 (MuC'21).* 325–329.

[165] Kamran Shafi. 2017. A machine competence based analytical model to study trust calibration in supervised autonomous systems. In *Proceedings of the 2017 9th International Conference on Advanced Computational Intelligence (ICACI'17).* IEEE, 245–252.

[166] Gagan Deep Sharma, Anshita Yadav, and Ritika Chopra. 2020. Artificial intelligence and effective governance: A review, critique and research agenda. *Sustainable Futures* 2 (2020), 100004.

[167] T. B. Sheridan. 1987. Supervisory control. In *Handbook of Human Factors*, G. Salvendy (Ed.). Wiley-Interscience, 1243–1268.

[168] Thomas B. Sheridan. 2019. Extending three existing models to analysis of trust in automation: Signal detection, statistical parameter estimation, and model-based control. *Human Factors* 61, 7 (2019), 1162–1170.

[169] Venkatesh Sivaraman, Leigh A. Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, trust, or negotiate: Understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI'23)*. ACM, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581075

[170] Robert D. Sorkin. 1988. FORUM: Why are people turning off our alarms? *Journal of the Acoustical Society of America* 84, 3 (1988), 1107–1108. https://doi.org/10.1121/1.397232

[171] Robert D. Sorkin, Barry H. Kantowitz, and Susan C. Kantowitz. 1988. Likelihood alarm displays. *Human Factors* 30, 4 (1988), 445–459.

[172] Sonia Sousa, David Lamas, José Cravino, and Paulo Martins. 2024. Human-centered trustworthy framework: A human–computer interaction perspective. *Computer* 57, 3 (March 2024), 46–58. https://doi.org/10.1109/mc.2023.3287563

[173] Randall D. Spain, Ernesto A. Bustamante, and James P. Bliss. 2008. Towards an empirically developed scale for system trust: Take two. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 52. SAGE Publications, Los Angeles, CA, USA, 1335–1339.

[174] Randall Steeb and Steven C. Johnston. 1981. A computer-based interactive system for group decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics* 11, 8 (1981), 544–552.

[175] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access* 9 (2021), 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

[176] Jonathan Tallant. 2017. Commitment in cases of trust and distrust. *Thought: A Journal of Philosophy* 6, 4 (2017), 261–267.

[177] Jiliang Tang, Xia Hu, and Huan Liu. 2014. Is distrust the negation of trust? The value of distrust in social media. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. 148–157.

[178] M. Susan Taylor and Thomas J. Bergmann. 1987. Organizational recruitment activities and applicants' reactions at different stages of the recruitment process. *Personnel Psychology* 40, 2 (1987), 261–285.

[179] Randy L. Teach and Edward H. Shortliffe. 1981. An analysis of physician attitudes regarding computer-based clinical consultation systems. *Computers and Biomedical Research* 14, 6 (1981), 542–558.

[180] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2021. Trustworthy artificial intelligence. *Electronic Markets* 31 (2021), 447–464.

[181] Myrthe L. Tielman, Catholijn M. Jonker, and M. Birna Van Riemsdijk. 2019. Deriving norms from actions, values and context. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 2223–2225.

[182] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI'22)*. ACM, New York, NY, USA, Article 160, 17 pages. https://doi.org/10.1145/3491102.3517732

[183] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* 1, 4 (2020), 100049.

[184] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 272–283.

[185] Anna-Sophie Ulfert, Eleni Georganta, Carolina Centeio Jorge, Siddharth Mehrotra, and Myrthe Tielman. 2023. Shaping a multidisciplinary understanding of team trust in human-AI teams: A theoretical framework. *European Journal of Work and Organizational Psychology* 33, 2 (2023), 158–171.

[186] David Callisto Valentine, Iskander Smit, and Euiyoung Kim. 2021. Designing for calibrated trust: Exploring the challenges in calibrating trust between users and autonomous vehicles. *Proceedings of the Design Society* 1 (2021), 1143–1152.

[187] Kees Van Dongen and Peter-Paul van Maanen. 2006. Under-reliance on the decision aid: A difference in calibration and attribution between self and aid. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. SAGE Publications, Los Angeles, CA, USA, 225–229.

[188] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.

[189] Francesco Walker, Anika Boelhouwer, Tom Alkim, Willem B. Verwey, and Marieke H. Martens. 2018. Changes in trust after driving level 2 automated cars. *Journal of Advanced Transportation* 2018, 1 (2018), 1–9.

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

[191] Lu Wang, Greg A. Jamieson, and Justin G. Hollands. 2009. Trust and reliance on an automated combat identification system. *Human Factors* 51, 3 (2009), 281–291.

[192] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. 109–116. https://doi.org/10.1109/hri.2016.7451741

[193] Xinru Wang and Ming Yin. 2021. Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 318–328.

[194] E. L. Wiener. 1981. Complacency: Is the term useful for air safety. In *Proceedings of the 26th Corporate Aviation Safety Seminar*, Vol. 117. 116–125.

[195] Philipp Wintersberger, Dmitrijs Dmitrenko, Clemens Schartmüller, Anna-Katharina Frison, Emanuela Maggioni, Marianna Obrist, and Andreas Riener. 2019. S(C)ENTINEL: Monitoring automated vehicles with olfactory reliability displays. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI'19)*. ACM, New York, NY, USA, 538–546. https://doi.org/10.1145/3301275.3302332

[196] 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*. 1–16.

[197] Toshio Yamagishi, Satoshi Akutsu, Kisuk Cho, Yumi Inoue, Yang Li, and Yoshie Matsumoto. 2015. Two-component model of general trust: Predicting behavioral trust from attitudinal trust. *Social Cognition* 33, 5 (2015), 436–458.

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

[199] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI'23)*. ACM, New York, NY, USA, Article 14, 14 pages. https://doi.org/10.1145/3544548.3581393

[200] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI'17)*. ACM, New York, NY, USA, 408–416. https://doi.org/10.1145/2909824.3020230

[201] Mireia Yurrita, Agathe Balayn, and Ujwal Gadiraju. 2023. Generating process-centric explanations to enable contestability in algorithmic decision-making: Challenges and opportunities. *arXiv:2305.00739* (May 2023). https://doi.org/10.48550/arXiv.2305.00739

[202] Zahra Zahedi, Sarath Sreedharan, and Subbarao Kambhampati. 2023. A mental model based theory of trust. *arXiv preprint arXiv:2301.12569* (2023).

[203] Qiaoning Zhang, Matthew L. Lee, and Scott Carter. 2022. You complete me: Human-AI teams and complementary expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI'22)*. ACM, New York, NY, USA, Article 114, 28 pages. https://doi.org/10.1145/3491102.3517791

[204] Richong Zhang and Yongyi Mao. 2014. Trust prediction via belief propagation. *ACM Transactions on Information Systems* 32, 3 (2014), 1–27.

[205] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*'20)*. ACM, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[206] Bing Zhu, André Habisch, and John Thøgersen. 2018. The importance of cultural values and trust for innovation—A European study. *International Journal of Innovation Management* 22, 2 (2018), 1850017.

Received 23 October 2023; revised 7 June 2024; accepted 20 August 2024