Humanities & Social Sciences Communications

REVIEW ARTICLE

https://doi.org/10.1057/s41599-024-04044-8 OPEN

Trust in AI: progress, challenges, and future directions

Saleh Afroogh[1✉], Ali Akbari[2], Emmie Malone[3], Mohammadali Kargar[4] & Hananeh Alambeigi[5]

The increasing use of artificial intelligence (AI) systems in our daily lives through various applications, services, and products highlights the significance of trust and distrust in AI from a user perspective. AI-driven systems have significantly diffused into various aspects of our lives, serving as beneficial "tools" used by human agents. These systems are also evolving to act as co-assistants or semi-agents in specific domains, potentially influencing human thought, decision-making, and agency. Trust and distrust in AI serve as regulators and could significantly control the level of this diffusion, as trust can increase, and distrust may reduce the rate of adoption of AI. Recently, a variety of studies focused on the different dimensions of trust and distrust in AI and its relevant considerations. In this systematic literature review, after conceptualizing trust in the current AI literature, we will investigate trust in different types of human–machine interaction and its impact on technology acceptance in different domains. Additionally, we propose a taxonomy of technical (i.e., safety, accuracy, robustness) and non-technical axiological (i.e., ethical, legal, and mixed) trustworthiness metrics, along with some trustworthy measurements. Moreover, we examine major trust-breakers in AI (e.g., autonomy and dignity threats) and trustmakers; and propose some future directions and probable solutions for the transition to a trustworthy AI.

Introduction

A person's trust in someone or something can determine their behavior, interaction, and acceptance (Siau, 2018). In fact, trust is a crucial factor in accepting and adopting technology in real life. Artificial intelligence (AI) refers to the capability of machines or systems to carry out tasks that typically require human intelligence (Srinivasan, 2019). It has become deeply embedded in our daily lives through a range of applications, services, and products. AI is now a crucial element of contemporary life, taking on a more significant role in our everyday activities (Lockey et al., 2021). AI has achieved significant progress in out-performing conventional solutions in many areas, including health (Itani et al., 2019; Shailaja et al., 2018; Wiens and Shenoy, 2018), autonomous transportation (Qayyum et al., 2020; Schwarting et al., 2018; Sligar, 2020), military (Galán et al., 2022; Roessingh et al., 2017), data

[1] Urban Information Lab, The University of Texas at Austin, Austin, NY 12203, USA. [2] Stanford University School of Medicine, 450 Serra Mall, Stanford, CA 94305, US. [3] Department of Philosophy, Lone Star College in Houston, Houston, TX, USA. [4] Department of Mechanical Engineering, Texas A&M University, College Station, TX, USA. [5] Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX, USA. ✉email: saleh.afroogh@utexas.edu

security (Aljably et al., 2020; Pan et al., 2019), entertainment (Brown and Sandholm, 2018; Mnih et al., 2013; Moravčík et al., 2017; Silver et al., 2016) etc. This has led to the rapid increase of AI-based methods in these areas.

Trust in AI can significantly control the level of this diffusion as distrust may reduce the chance of adoption of AI. Trust in AI can be viewed as "the willingness of people to accept AI and believe in the suggestions, decisions made by the system, share tasks, contribute information, and provide support to such technology" (Siau, 2018). AI can be developed and adopted only if it satisfies the stakeholders' and users' expectations and needs, and that is how the role of trust becomes essential. In general terms, trust is built when the trustor can anticipate the trustee's behavior to know if it matches its desires (Jacovi et al., 2021a). Therefore, individuals, organizations, and societies will only ever be able to realize the full potential of AI if trust can be established in its development, deployment, and use (Thiebes et al., 2021a). Therefore, it is vital to understand the definition, scope, and role of trust in AI technology and determine its influential factors and unique application-dependent requirements.

Trust in AI is not just a non-technical ethical consideration (Ryan, 2020a). Instead, it also includes various domains, including AI performance, transparency and explainability, and compliance with legal and technical regulations. AI is different from other automated systems in the sense that it can learn, and it can behave proactively, unexpectedly, and incomprehensibly for humans (Saßmannshausen et al., 2021). Overall, influential factors of trust in technology could be divided into human-based, context-based, and technology-based factors. No matter what technology the trustee is, the impacts of human-based and context-based factors are more or less similar. For instance, a person with a high-trusting stance would be more likely to accept and depend on new technologies (Siau, 2018). However, the technology-based factors of AI that affect trust are unique and usually more challenging than other technologies, even compared to rule-based automation. That is because, in AI, the system can make new decisions based on training data. Therefore, parameters such as accuracy, reliability, transparency, and explainability of the decision become extremely important to determine the level of trustworthiness of AI.

Recently, many researchers have tried to identify reasons for distrust in AI and improve trust by different means since distrust has hindered the successful adoption of AI technology in various domains. For instance, despite AI's considerable potential in the manufacturing industry, its application still faces the challenge of insufficient trust due to the black-box nature of AI, which introduces difficulties for ordinary users to understand it (Li et al., 2021a). Medical imaging is another domain that can significantly benefit from AI technology. Still, these technologies have not been widely adopted in this area due to a lack of trust by medical practitioners, healthcare stakeholders, and patients, in addition to regulatory, medicolegal, or ethical issues (Z. Song et al., 2021). Similarly, risk-averseness and lack of trust have limited the application and adoption of AI technology in many other domains such as autonomous vehicles (Ajenaghughrure et al., 2020a), customer service chatbots (Adam et al., 2021a), personal assistants (Wu and Huang, 2021a), finance (J. Li et al., 2016; Sarpatwar et al., 2019), depression treatment (Yan and Xu, 2021a), robotics (Lazányi, 2019), and IoT (Hong et al., 2009). Accordingly, trust in AI functions as a driver in AI usage, and distrust is considered a barrier to the development and application of AI systems, and it would negatively affect the stakeholder's perspective toward AI systems in different contexts.

Different dimensions and impacts of trust/distrust in AI are mentioned and discussed in a variety of studies, reports, and case studies in different domains. The current body of knowledge, however, lacks a systematic review of the different dimensions and varying considerations of conceptualization of trust/distrust in AI, and a discussion of the relationships and possible resolutions of these considerations. Therefore, in this study, we conduct a systematic literature review to (1) reveal the different conceptions and theories of trust/distrust in AI, as well as its different types, models, and relevant impacts; (2) discuss the two major classes of technical and axiological trustworthy metrics, relevant frameworks and measurements, as well as distrust origins and motivations, such as autonomy and dignity threat; (3) provide solutions for some problems and considerations that accelerate the transition to a trustworthy and responsible AI.

Our discussion proceeds as follows (see Table 1): *Methodology*, describes the methods used in systematic reviews of studies related to trust in AI. *Findings*, present the findings and results related to the key values and major cords and how they are discussed in the literature. It includes the following subsections: "Different types/models of trust in AI", "Trustworthy AI and its metrics: trustworthy AI", "Distrust in AI and Scary AI".

Trust makers: building/increasing trust in AI. *Discussion*, analytically discusses the major codes and key values and considerations related to trust in AI, as well as the address of the practical value conflicts and the probable tradeoff between the key values and considerations. It includes 12 subsections. Concluding Remarks and Future Directions for trust research in AI are also discussed in the section "Concluding Remarks and Future Directions".

Methodology

We conducted an inclusive and systematic review of academic papers, reports, case studies, and trust frameworks in AI, written in English. Given that there is not a specific database on trust in AI in particular, we used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework to develop a protocol in this review (Fig. 1).

In order to conduct a comprehensive review of the relevant studies, we followed two approaches. First, we manually searched for the most related papers on trust in AI: 19 papers were identified through the online search after the removal of duplicate files. Secondly, we fulfilled a keyword-based search (using the http://scholar.google.com search engine) to collect all relevant papers on the topic. This search was accomplished using the following keyword phrases: (1) "trust + AI" which provided 19 relevant result pages of Google Scholar, (2) "trust + Artificial + Intelligence" for which the first five result pages were reviewed, (3) "trustworthy + AI," for which the first 15 result pages were reviewed; and (4) "trustworthy + Artificial + Intelligence," for which the first 13 result pages of Google Scholar were reviewed. Moreover, the following keywords, "Trust + explainability/transparency/interpretability/empathy/privacy/fairness/accountability/safety/accuracy/robustness + AI /Artificial + Intelligence" and "distrust + AI/ Artificial + Intelligence" were included because of their central role in the research as the major known (based on a preliminary review) considerations of trust in AI. Additionally, the search was suspended within results for each search term due to limited appearances of new relevant papers on the following pages.

The results of the search were 336 relevant papers (which were selected based on the semantical keywords relevancy), out of 1205 (which appeared on the result pages). Afterward, the duplicated papers were eliminated from the analysis. We selected the 329 target papers for this systematic review based on the following two inclusion/exclusion criteria. First, articles that were published in academic journals were included. Second, the dominant topic of the papers (or a significant part of it) was trust in AI. To this end, the papers' main sections were reviewed to understand their

Table 1 A road map of this study.

Section N.	Section Title		Subsection themes
1	**Introduction**		
2	**Methodology**		
3	**Findings**	3.1. Different types/models of trust in AI (Human-Machine interaction)	3.1.1. Theories and definitions of trust in AI
			3.1.2. Trust in types of human-Machine interaction
			3.1.3. Impact of trust/distrust on AI technology acceptance in different domains
		3.2. Trustworthy AI and its metrics: Trustworthy AI (i.e., technical, and non-technical metrics: legal, ethical, mixed)	3.2.1. Trust & explainability / transparency / interpretability
			3.2.2. Trust & empathy in AI
			3.2.3. Trust and privacy
			3.2.4. Trust and fairness in AI
			3.2.5. Trust and accountability in AI
			3.2.6. Trust and technical metrics (safety, accuracy, robustness,)
			3.2.7. Evaluating and measuring/ trustworthiness certificate in AI
			3.2.8. Trustworthy AI Frameworks
		3.3. Distrust in AI and Scary AI	3.3.1. Distrust makers in AI systems
			3.3.2. Surveillance, and manipulation
			3.3.3. Human autonomy/dignity threat
			3.3.4. Distrust and unpredictable futures
			3.3.5. Challenges and barriers to breaking distrust
		3.4. Trust makers: building/increasing trust in AI	3.4.1. Factors that affect trust
			3.4.2. Methods of Building trust in AI
			3.4.3. Case studies and items effects on building trust
4	**Discussion**	4.1. Interaction of technical and non-technical factors of trust and trustworthiness in AI	
		4.2. Non-interchangeability of interpretability, explainability and transparency, and their classification	
		4.3. Trust as a two-way street	
		4.4. Distinction between empathy in human's trust in AI and empathy in AI's trust in human agents	
		4.5. Tradeoff between empathy and privacy	
		4.6. The subjectivity of trust in AI vs. the objectivity of reliable AI	
		4.7. AI privacy and human agent privacy	
		4.8. The developmental problem of 'right to explanation'	
		4.9. Development of direct AI accountability	
		4.10. Challenges of measuring trust and trustworthiness in AI	
		4.11. Trust equity problem in AI	
		4.12. Impossibility of Interpersonal trust in AI systems	
5	**Concluding Remarks and Future Directions**		

dominant topic rather than only relying on the title and papers' keywords.

Findings

The qualitative analysis of the selected papers was performed by four researchers who critically read the papers and who developed the eight major key codes as the building blocks of the categorization of the review result in the next step of this research (Table 2).

Trust in human–machine interaction: typology and parameters

Theories and definitions of trust in AI. Trust is a central component of interaction between people and artificial intelligence (AI) as well as machines and AI since "incorrect" levels of trust

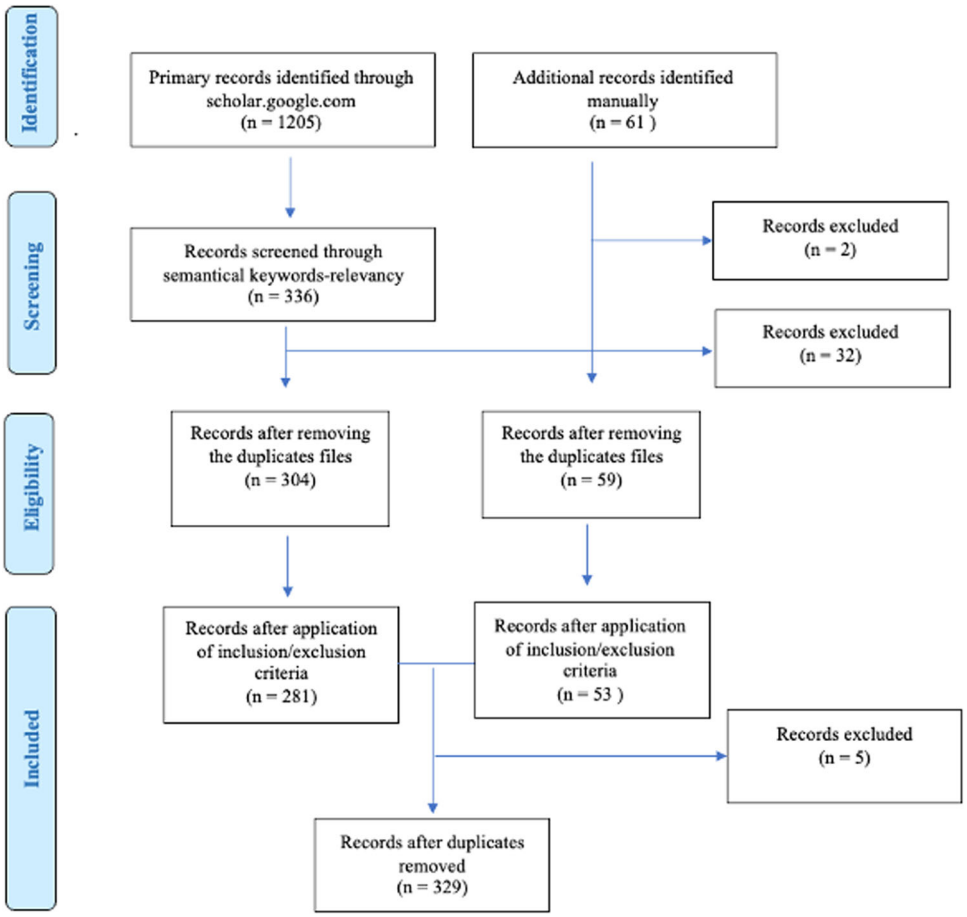

Fig. 1 Developed PRISMA flow diagram for review of trust in AI.

may cause misuse, abuse, or disuse of the technology (Jacovi et al., 2021a). To understand trust's implications and influential factors, we first need to have a formalized definition of trust that is expandable to AI. Generally, trust is defined as a directional transaction between two parties. In this definition, A trusts B if it believes that B will act in its best interest and accepts vulnerability to B's actions (Mayer et al., 1995). Trust is then necessary to predict events by anticipating the impact of actions and behaviors and facilitating collaboration between the parties (Misztal, 2013). Hence, trust is tightly coupled with vulnerability, anticipation, risk, and uncertainty. The trustor needs to anticipate the trustee's behavior to know if it matches its desires. Still, at the same time, the trustor knows that there is a level of uncertainty associated with this anticipation. Therefore, there is a risk of disadvantageous or otherwise undesirable events (Jacovi et al., 2021a).

The critical question is how to adapt the general definition of trust to the notion of AI. AI can be broadly defined as a computer program that can make intelligent decisions (Mccarthy and Hayes, 1969). In the context of AI, the meaning of anticipation in trust changes since the goal of the trustor is not necessarily to anticipate AI's behavior; instead, the trustor needs to anticipate if the model is correct and confident in its decision. This definition can be further expanded based on the theory of contractual trust, which states that the trustor should anticipate or believe that the trustee will stick to a specific contract, which could be any functionality that is deemed useful (Hawley, 2014; Tallant, 2017). In that sense, the former definition would be a specific case of trust in AI, which is the trust in the model's "correctness." In some instances, humans may trust in other functionalities of an AI model rather than its correctness. For example, a classifier trained for medical samples may reveal strong correlations between attributes for one of the classes, demonstrating causation between the attributes, even if the model is not helpful for the original classification task (Lipton, 2019).

From a different perspective, while interpersonal trust is associated with benevolence, integrity, and ability, trust in AI is less relevant to honesty and benevolence since AI systems lack intentionality (Asan et al., 2020a). Trust in AI heavily depends on the user's perception of an AI system's ability, which depends on the quality of the input data, the mathematical problem representation, and the algorithms used in the decision-making. In general, AI systems could be generative, and they could learn, evolve, and permanently change their functional capacities with operational and contextual information (Kessler et al., 2017). As a result, AI-based systems' actions and decisions could become more indeterminate across time, making them more challenging to predict. Consequently, establishing trust between humans and AI-based systems is generally more complex and challenging to understand than interpersonal trust (Thiebes et al., 2021b).

A vital phenomenon here is the fact that trust and trustworthiness are entirely disentangled: pursuing one does not entail following the other, and trust can exist in a model that is not trustworthy, or a trustworthy model does not necessarily gain trust (Gille et al., 2020). For example, in the healthcare domain, a highly complex classifier trained to identify the risk of cardiovascular diseases from a combination of genetics, lifestyle, and metabolic factors may show a high accuracy, meaning that it is trustworthy regarding the correctness; however, this model may not be trusted by the healthcare providers since the logic behind

Table 2 Major and minor codes included in the reviewed papers.

Major ethical codes	N. of reviewed papers	Minor ethical codes
Theories and definitions of trust in AI	16	Interaction between human and AI, directional transaction, vulnerability acceptance, facilitating collaboration, risk and uncertainty, model's "correctness", confident decision, integrity and ability, user's perception of an AI system's ability, aesthetic of a user interface, behavior and risk anticipation, reliability, perceived trustworthiness, predictive power, over-trust, beneficence, non-maleficence, autonomy, justice, and explicability
Trust in types of human-machine interaction	25	human-machine interactions, machine-human interactions, machine-machine interactions, machines as the host of AI, Robo-advisors, autonomous vehicles, Adversarial attacks, unreliable sources, smart contracts, self-imposed standards, certification, corporate guidelines, governmental regulations
Impact of trust/distrust on AI technology acceptance in different domains	31	economic output, electronic markets, acceptance of AI by physicians, Reduce the wait time in healthcare, algorithmic investment advice, AI-based personal assistants, chatbots, [AI] deception, racist and genocidal ideologies in [AI developers], managers' endorsement, cognitive trust, emotional trust
Trust & explainability/ transparency/interpretability in AI	53	complex opaque concepts, deep neural networks, opaque nature of complex AI algorithms, AI-based decisions, transparency vs. explainability, transparency against overtrusting AI, levels of transparency, dynamic process [of trust building], transparency criteria, [right of rejecting] automated processing, interpretability vs. explainability, [AI's] decision's rationale, human-interpretable, model's inner machinery, pre-model interpretability, intrinsic interpretability, post-hoc interpretability
Trust & empathy in AI	12	subjective process, deep understanding of other people's feelings, non-judgmental, ability to simulate, cognitive ability, accurate inference, empathic accuracy, affective or emotional ability, supportive, benevolent, and compassionate response, other's feelings and thoughts, efficient communication, social bonding, social interactions, people's mental states, user's expectations, agent's credibility and trust, empathic cultural-aware agents, social values and norms, behavioral and motivational levels, observation and detection of social signals, empathic action and interaction, stakeholders' viewpoints, similar tastes and ratings
Trust and privacy	20	personal information, detailed information, customer privacy empowerment, privacy-by-design, data minimization, controllability, transparency, easy-to-use privacy function, data confidentiality, different levels of privacy, a pessimism problem,
Trust and fairness in AI	9	Algorithmic bias, discrimination, perceived fairness, induced fairness, equal treatment, regulations, Algorithmic unfairness, minority and marginalized groups, User biases, perceptions of harm and injustice, reported wrongdoing, rules and regulations Implementation

its decision is not clear. A different example is trust in an untrustworthy AI. For instance, there is a correlative but not a causal relationship between high-quality visual interface (GUI) and trustworthy AI models. If the cause of the user's trust is the model GUI, then the model's ability to make correct predictions will not affect this trust (Jacovi et al., 2021a). Even an untrustworthy AI model with poor performance could be trusted merely because of the good GUI. Ghassemi et al. showed a case where the interface can increase doctors' confidence in a tool, despite not significantly increasing the AI's accuracy (Ghassemi et al., 2018).

Table 2 (continued)

Trust and accountability in AI	5	legal framework, public trust in AI, reliability of models, minimum standards, Transparent explanations, empathy, privacy concerns
Trust and technical metrics (safety, accuracy, robustness,) in AI	22	vulnerability of the user, technical elements of trust, Reliability, security, lineage, system accuracy, weight of advice, high-stakes decisions, calibrating the trust in AI systems, distrust AI, over-trust AI, zero-touch security, diversity, segregating malicious nodes, cybersecurity, Bayesian-based trust model, human multi-robot team, biosignals
Evaluating and measuring/ trustworthiness certificate in AI	17	psycho-physiological approaches, empirical approaches, theoretical methods, accuracy and safety guidelines, qualitative evaluations, experimental constraints, trust-theoretical model, transparent AI systems, vulnerability and risk assessments, physiological model, multi-dimensional metrics, ndividual and team performance scores, situation awareness, philosophical evaluation of trust, complexity and inexplicability, algorithmic auditing, customization of AI certification, resilience, agility, satisfaction, efficiency, data protection, predictability, believability
Trust Frameworks in AI	22	ethical AI systems, flexibility, accurate incorporation of the data, privacy protection, ethics by design, ethics in design, ethics for design, algorithmic bias, worldwide health, fairness, accountability, transparency, behavioral patterns, invitation of trust, culture factors, explainability, multi-level framework, robot autonomy, sociological framework, unwarranted varieties of trust, warranted trustworthy AI, Human agency, technical robustness and safety, data governance, privacy, diversity, societal and environmental wellbeing, trust measurement, justice and fairness, non-maleficence
Distrust in AI	26	Scary AI, faster than human beings, malevolent artificial intelligence, alignment problem, surveillance, privacy, hackability, loss of human control, [AI's] uses in war, applications in healthcare, potential consequences of AI for the economy, Distrust makers in AI, surveillance & manipulation, human autonomy, dignity thread, and unpredictable futures, distrust breaker, optimal level of transparency, bias propagation
Trust makers: Building/increasing trust in AI	33	technical and axiological factors, AI personality, anthropomorphism, reputation, transparency, team-related factors, context-related factors, individual-related factors, human traits, actual capabilities of the AI, human agency and oversight, accountability, marketing, over-trust, technical method for building trust, global and local explainability, local justifications, interactive visualization, sharing transparency, standard or technical regulation, non-expert end-users, experts' endorsement, AI risk-mitigating practices, certification/accreditation, performance metrics, responsibility of AI system, interpretations of these standards, Value-based trust, positive values, Good will, biases elimination, ethics-washing, ethical guidelines, marketing communication, Human-related factors (expertise, culture, personal traits), AI-related factors (accuracy

It is also shown that a model is more trustworthy when the observable decision process of the model matches user priors on what this process should be. This is equivalent to, for example, a doctor that is considered more trustworthy because they are citing various respectable studies to justify their claims. The relationship between actual trustworthiness, which is a characteristic of the trustee, and perceived trustworthiness, which is a characteristic of the trustor, and its influence on the trust has been modeled (Schlicker and Langer, 2021). The central assumption is that the actual trustworthiness cannot be accessed directly and is therefore inferred via cues to form a user's perceived trustworthiness. Cues are observable pieces of

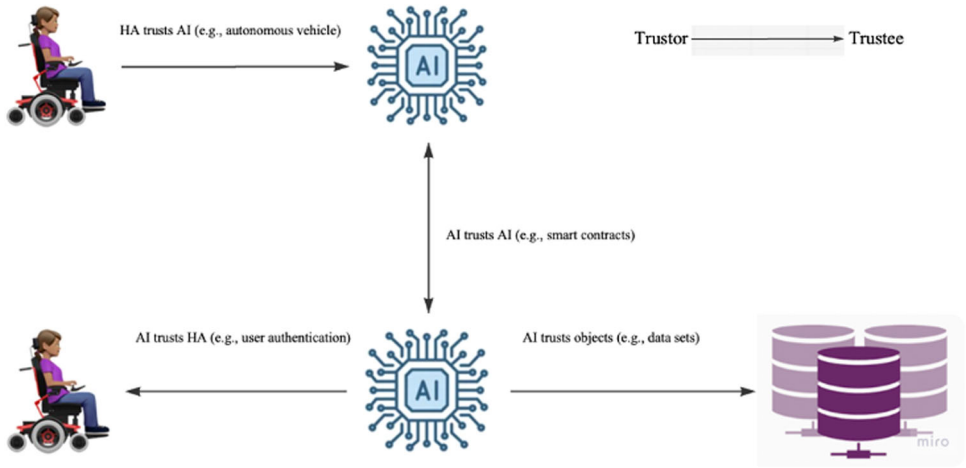

Fig. 2 Types of human, object, and AI trust interaction.

information such as the esthetic of a user interface, information about the inputs a system uses, single outputs that a system produces, the displayed predictive power of a classifier, information about uncertainty accompanying a classification output, a stated or communicated rationale for the system's recommendation, or the logo of a company. These cues could significantly affect the perceived trustworthiness and the trust consequently. Therefore, engineering and misleading cues could lead to over-trust.

Accurate assessment of actual trustworthiness that leads to realistic perceived trustworthiness is affected by four factors, including relevance, availability, detection, and utilization, where relevance and availability are associated with the trustee (e.g., an AI system), and detection and utilization are associated with the trustor (e.g., the user) (Schlicker and Langer, 2021). A relevant cue can be any information regarding a system's predictive accuracy in a task. Relevant cues provide information related to a system's performance, which is considered a facet of trustworthiness (J. D. Lee and See, 2004). Availability means that only the cues that are accessible to the trustor can be leveraged to assess trustworthiness. Detection refers to the fact that the trustor must detect the relevant and available cues, and the utilization means that the trustors must be able to correctly interpret the relevant, available, and detected cues toward the estimation of trustworthiness (Schlicker and Langer, 2021).

For an AI system to be perceived as trustworthy, five principles need to be fulfilled, including beneficence, non-maleficence, autonomy, justice, and explicability (Dosilovic et al., 2018). Beneficence refers to the development, deployment, and use of AI that is beneficial to humanity and the planet and respects fundamental human rights. On the other hand, non-maleficence means the development, deployment, and use of AI such that it avoids bringing harm to people. The autonomy principle is not directly related to trusting beliefs, but it helps mitigate integrity and reliability risks by balancing human- and machine-led decision-making. Justice is also a broad term that covers the utilization of AI to amend past inequities like discrimination and biases and the creation of shareable and subsequent distribution of benefits. Finally, explicability requires the creation of explainable and interpretable AI models while maintaining high levels of performance and accuracy from a practical perspective and creating accountable AI from an ethical perspective (Thiebes et al., 2021a). Different guidelines proposed for building ethical and trustworthy AI have addressed different combinations of these principles.

Trust in types of human–machine interaction. Recent technological breakthroughs in artificial intelligence and machine learning have generated a surge of interest in the usage of AI technology in daily life. Many applications, such as healthcare, autonomous vehicles, financing, and marketing, benefit from the development of AI technology. Nowadays, AI is tied to many daily tasks, and different types of interaction between humans and machines occur everyday. Figure 2 shows different types of interactions, including human–machine, machine-human, and machine-machine interactions. Herein, we refer to machines as the host of AI technology. Therefore, machines have the capability of reasoning and decision-making based on the data.

Human–machine interaction is the most common type of interaction with AI, in which the trustor is the human user and the trustee is the AI system. For example, in healthcare, AI could process medical images to diagnose cancer, and the trustor would be physicians who adjust or base their decision on the outcome of the AI model (Asan et al., 2020a). In another example, Robo-advisors make investment advice (Ajenaghughrure et al., 2020b; Szeli, 2020a). In autonomous vehicles, human drivers interact with the AI-based driver in safety-critical situations (Ajenaghughrure et al., 2020b). Humans also often use AI-powered personal assistants and chatbots (Fan et al., 2021a; Wu and Huang, 2021b). Examples of human-AI interactions are endless, and trust is essential to facilitate this interaction. Machine-human interaction is a more special case that has not been widely addressed in the literature. The AI systems need to acquire information from human users to update their algorithms. In this scenario, the AI system needs to ensure acquiring data or annotations from trustworthy resources. Adversarial attacks or unreliable sources of information could lead to poor performance of the AI systems. In addition, in many privacy-critical applications such as healthcare, the AI system needs to identify and authorize trusted human users before sharing the data. In this case, user authorization would be essential. Finally, machine-machine interactions are becoming more popular in the light of technology development. For example, sensor networks and IoT (internet of things) strongly rely on machine-machine interactions. In addition, the domains of electronic financing, smart contracts, cryptocurrency, and smart vehicles require an extensive amount of interaction between different AI systems in which trust is paramount since adversarial attacks in these cases are highly possible (M. Wang et al., 2020).

Due to the prevalence of human-AI interactions and the fact that it was the first type of interaction since AI emerged, this is the most widely studied topic in the literature. In particular, factors such as access to knowledge, transparency, explainability, certification, as well as self-imposed standards and guidelines are

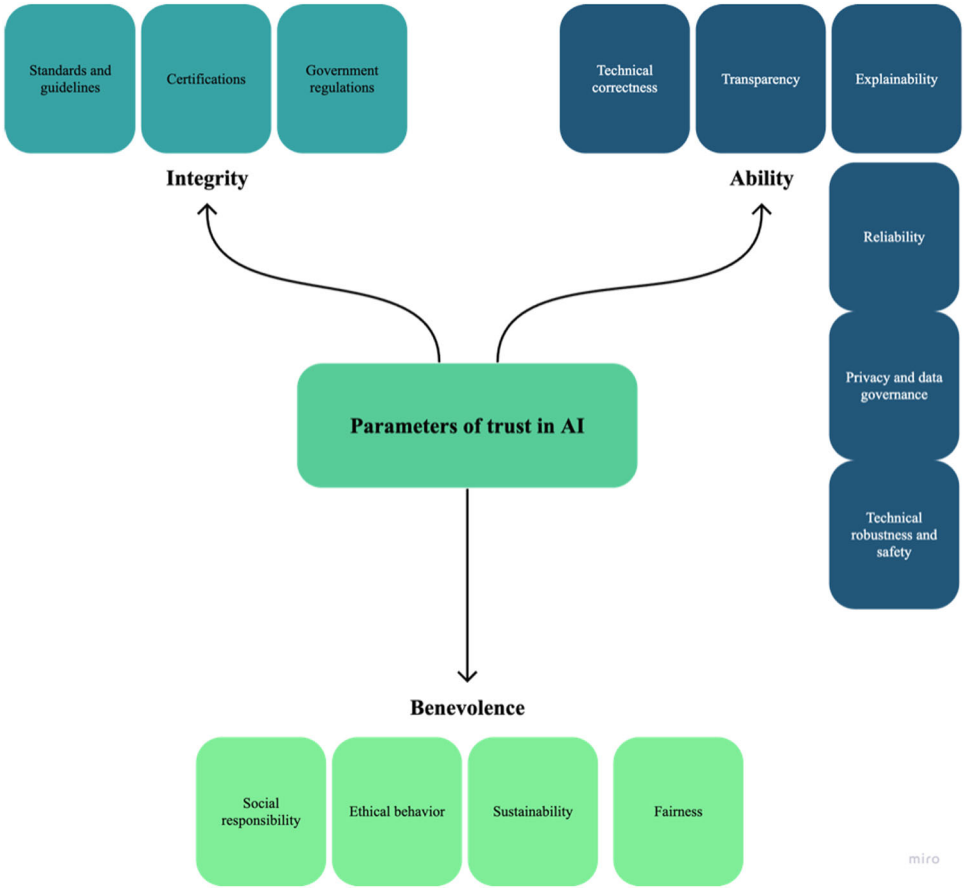

Fig. 3 Parameters of trust in AI.

important to build trust in human–machine interactions. Figure 3 summarizes some of the most important factors of trust and their belonging category based on benevolence, integrity, and ability (Bedué and Fritzsche, 2021). All these factors matter in promoting trust in human–machine interactions as the human is the trustor. However, in other types of interactions, mostly integrity factors and data governance matter, while transparency and explainability are less important than technical correctness and integrity parameters.

The safety and efficiency of human–machine collaboration depend on the perceived trust of the human trustor. Over-trusting the AI system may cause serious safety issues. A method of adaptive trust calibration was developed by detecting the inappropriate calibration status via monitoring the user's reliance behavior and cognitive cues to prompt the user to reinitiate trust calibration. This becomes significant in applications such as military coalition operations, where data is limited and often of low quality. These problems can be mitigated by taking steps that allow rapid trust calibration so that decision-makers understand the AI system's limitations and likely failures and can calibrate their trust in its outputs appropriately. An AI service can achieve this by being both interpretable and uncertainty-aware (Okamura and Yamada, 2020a).

Although considering all these factors could increase the trustworthiness of AI systems, in the case of human–machine interaction, personal traits, especially emotions, play an important role in developing trust as the trustor is a human. It was shown that the humanness of AI applications is an important basis for trusting bonds in human–machine interactions (Troshani et al., 2021). People may express conflicting concerns about the lack of empathy in an AI decision where some may see it as a positive aspect that increases trust in the process by keeping human emotions in check, while others think of AI's lack of empathy and morality disqualify it for making higher-stakes decisions (Ashoori and Weisz, 2019). It was argued that personality often overrides any external influence on trust (Sharan and Romano, 2020). Researchers found that different people might have different perceived trust throughout the process of engaging with AI systems (Tutul et al., 2021a). Based on these findings, open individuals trust the AI decision more than non-open individuals. Moreover, trust in AI changes over time when the trustor is a human. In many human–machine interactions, there is a need for collaboration between humans and AI agents. For example, in automated driving, the automated driving agent may release the control for the human driver to take over in certain critical situations. In this case, there are important questions that affect this interaction. First, how should functions between humans and machines be allocated? Answering this question requires technical and contextual knowledge. For example, when the automated driving agent faces a critical situation where it cannot handle or is uncertain about its decision, it should initiate the transfer of control (McDonald et al., 2019). In another example, when an AI system is not certain about its decision due to noisy data or lack of training, it can interact with the human user to verify its decision or obtain more training (Akbari and Jafari, 2020). The second question is who is doing the allocation? The AI or the human user? Third, who can authorize an allocation?

There are several risks associated with human-AI interactions, mostly related to the AI agent's performance, fairness, and transparency. In the absence of transparency and explainability of the AI, the human does not have enough information to form a

judgment regarding the chosen decision. Moreover, undesirably biased recommendations could make humans accountable for unethical or legally uncompliant decisions (Abbass, 2019a). Bias can exist in many shapes and forms (Mehrabi et al., 2021)—data-related biases such as measurement bias, historical bias, population bias, longitudinal data fallacy, and social bias, algorithm-related biases such as linking bias, omitting important and influential variables from model, and algorithmic biases, and results interpretation-related biases, such as aggregation bias, user interaction bias, and evaluation bias.

AI–AI interaction is an emerging paradigm. For example, smart and connected vehicles have gradually stepped into our daily lives, and they generally rely on vehicular networks to generate and exchange traffic-related messages. Malicious accidents could result from the untrusted content of vehicle navigation and autonomous systems. Blockchain and artificial intelligence (AI) empowered trust management systems were developed for trust evaluation, where it was leveraged to filter the information gained from other smart vehicles (Pan et al., 2020; Zhang et al., 2021). In another study, an Intelligent Trust Collaboration Network System (ITCN) was developed to collect data through collaboration with mobile vehicles and Unmanned Aerial Vehicle (UAV), in which there is a score determining trust associated with each vehicle (Guo et al., 2022). Distributed systems of software agents are another example of AI–AI interaction where the AI agents cooperate in helping their users find services provided by different agents. Examples are prevalent in blockchain, cryptocurrency, and smart contracts (Ahmed and Aura, 2018; al Khalil et al., 2017; Albizri and Appelbaum, 2021; Beck et al., 2016; P. A. Ryan, 2017). In this scenario, the agents need to ensure that the service providers they select are trustworthy. Because the agents are autonomous and there is no central trusted authority, the agents help each other to determine the trustworthiness of the service providers they are interested in. A trust network is a multiagent system where each agent potentially rates the trustworthiness of another agent (Y. Wang and Singh, n.d.).

AI-Object interaction is the final paradigm; however, it has barely been discussed in the literature so far. For instance, a self-driving vehicle's AI system needs to trust stop signs; or an AI system, which is used in disaster management requires recognition of un/trustworthy social network datasets during disasters. In addition to physical objects, a trustee, in this paradigm, would also include mental objects, such as theories, thoughts, and algorithms that an evaluative AI system must handle. There are also complex objects at work such as institutions and systems, as trustees.(Afroogh, 2022).

Impact of trust/distrust on AI technology acceptance in different domains. AI is one of the most-discussed technology trends in research and practice today and is estimated to deliver an additional global economic output of around 13 trillion dollars by the year 2030 (Bughin et al., 2018). Various domains benefit from AI as it helps decision-makers and provides vital services for end-users. AI has significantly affected domains such as healthcare (Benda et al., 2021; Hui et al., 2021; Jacobs et al., 2021; D. K. D. Kim and Kim, 2021a), finance (Chandra et al., 2010; Kumar et al., 2021; Zierau et al., 2021), personal assistance (Fan et al., 2021b; Pitardi and Marriott, 2021a; Zierau et al., 2020), autonomous vehicles (Ajenaghughrure et al., 2020a), etc. The importance of trust in AI extends to other areas as well. Electronic markets, for example, are increasingly augmented with AI-based systems such as customer service chatbots (Adam et al., 2021b). Likewise, several cloud providers recently began offering "AI as a Service," referring to web services for organizations and individuals interested in training, building, and deploying AI-based systems (Dakkak et al., 2019).

Trust is paramount for the well-functioning of healthcare systems and, consequently, for the acceptance of AI by physicians and within healthcare more broadly (Gille et al., 2015). Transparency and explainability are the most important factors of trust in healthcare systems (Caspers, 2021; Gille et al., 2020). Other concerns that can decrease physicians' trust in AI include, among others, the low number of randomized clinical trials to test the performance of AI systems, the lack of transparency of information flows within AI applications, the risk of inequity and discrimination introduced by algorithmic biases, and insufficient regulatory clarity (Nagendran et al., 2020; Vollmer et al., 2018). In addition, limited public literacy about AI negatively affects trust in healthcare (Gille et al., 2020).

In a study investigating the trust of the end-users in detecting heart diseases from ECG signals acquired by smartwatches and detecting cancer from skin photos, it was found that several participants appreciated the more holistic perspective of a human doctor compared to the limited focus of an AI-powered app. Moreover, they mentioned the social impact of visiting and talking to a human doctor, which is missing in AI-driven apps. Many participants raised concerns regarding the overall technical feasibility of these AI-driven diagnostics, and they did not want to be notified about life-threatening health complications through an app. Finally, although some participants found the AI-driven diagnostics helpful in reducing the wait time, they mentioned that they would first test the performance of such an AI-based app themselves (Baldauf et al., 2020). These findings show the distrust in AI-driven diagnostics systems, leading to lower acceptance of technology. In order to replace or supplement human diagnosis from physicians and healthcare professionals, it may not be enough for the AI diagnosis system to be just accurate as an accurate diagnosis without justification or explanation might be ignored. The format and the timing of explanation play important roles in regulating trust in healthcare systems (Lui and Lamb, 2018). Algorithmic analysis of pathology images is one of the most promising and advanced applications of AI in healthcare (Meyer, 2021). However, few AI systems are currently being used in the field, and it is uncertain to what extent pathologists will adopt AI and rely on its recommendations.

Another example of the role of trust as a driving factor for technology adoption is found in the financial sector. While AI already enjoys a high level of trust in some areas (entertainment, navigation), only half of people trust algorithmic investment advice (Szeli, 2020b, 2020a). Many human investors would rather trust a human prediction than an algorithmic prediction (Diab et al., 2011), a phenomenon known as algorithm aversion because humans are more tolerant if a human is mistaken than if it is an algorithm. When relying on AI algorithms to manage investment, humans' loss tolerance is highest when humanized algorithms manage portfolios—e.g., by giving the algorithm a human name (Szeli, 2020a). Several banks have leveraged chatbots for interaction with customers in the financial sector. One of the advantages of AI systems to gain customers' trust is the inherent absence of self-interest. Nevertheless, humans are still preferred to advise customers concerning complex financial products such as equity derivatives. Humans are also preferred when customers wish to complain or discuss a complicated matter or situation. A common criticism of chatbots and robots is that they cannot empathize (Lui and Lamb, 2018).

AI-based personal assistants, chatbots, and coaches are other domains in which trust in AI directly impacts technology adoption. AI-based voice-assistant systems (VAS) are used for various purposes in daily lives. It was found that interaction quality (e.g., information and system quality) and trust are critical factors influencing the adoption of AI-based VASs (O.-K. D. Lee et al., 2021). Although the current AI systems do not have an

internal drive to misbehave, lack of transparency in these systems may seem to indicate deception to some users (Kaplan et al., 2021). For example, Microsoft's AI bot, "Tay," was meant to learn to chat by communicating with internet users. Instead, due to toxic influences, it began spouting racist and genocidal ideologies, which resulted in distrust and users' outrage that forced Microsoft to suspend the system (*The Racist Hijacking of Microsoft's Chatbot Shows How the Internet Teems with Hate|Paul Mason| he Guardian*, n.d.).

Despite the significant potential of AI in the manufacturing industry, its application still faces the challenge of insufficient trust. Research on how users trust AI in an organization such as a manufacturing company is rare. In the organizational context, the decision of trust in AI is not completely personal; instead, users must consider the institutional influences of the company, the leader, or peers before they make the final trust decision. Research showed that factors such as gender, age, education, and position had no significant effect on organizational trust. Instead, the support from the top manager acts as the endorsement to ensure that the AI is qualified and that the AI-related project will be successful, which enhances trust in the organization (J. Li, Zhou et al., 2021b).

Robotics is another important field empowered by AI, which requires trust between humans and machines. Influential factors of trustworthiness in the context of social robots were investigated (Y. Song and Luximon, 2020). Robot-relevant issues (e.g., the characteristics and performance of the robot), human-relevant issues (the specific need, propensity to trust, personality, comfort, self-confidence, attitude, memory, attention, expertise, competency, workload, prior experience, and situation awareness), and scenario relevant issues (task application, task complexity, multi-tasking requirement, physical environment, in-group membership, culture, communication, team collaboration, etc.) were found significant factors, among which robot-relevant issues are the most significant factors influencing people's trustworthiness evaluation towards human-robot interaction. In conclusion, it was shown that cognitive trust and emotional trust are positively related to the intention to adopt an AI-based recommendation system as a decision aid, where cognitive trust has a stronger effect. Moreover, emotional and cognitive trusts were found correlated (Shi et al., 2021).

Trustworthy AI and its metrics: technical, and non-technical metrics. The increasing use of artificial intelligence (AI) in various industries, including healthcare, has raised concerns about its trustworthiness. Trustworthy AI is critical for ensuring that AI systems are reliable, safe, and ethical. In this context, several technical and non-technical metrics have been proposed to evaluate the trustworthiness of AI systems in healthcare. Focusing on many clinical research in AI and robotic abilities for the diagnosis of diseases or rehabilitation assistance revealed the significance of discussion about trust definition in the clinical decisions or suggestions and factors that can improve the clinicians and technical trust to the AI (Kellmeyer et al., 2018; Shafiei et al., 2018) (Asan et al., 2020). In addition, it is needed to focus on some non-technical (e.g., ethical or legal) foundation for autonomous AI requirements (such as maximizing traceability of patients and ongoing monitoring of real-world performance), which can be found in (Abramoff, 2021) and references therein. On the other hand, there are some articles that clarify why we can't use AI in medicine as a trusty system (DeCamp and Tilburt, 2019), and/or use it with some limitations because AI reliability is insufficient (Hatherley, 2020) that can enable better interpretation (Cabitza et al., 2021). In categorization which was carried out by Kush R. Varshney (Varshney, 2022), accuracy and safety are preliminary metrics for AI, reliability includes fairness and robustness, and also, transparency (which includes interpretability and explainability) and value alignment are essential for human-AI interaction. He also itemized the principles in ethics guidelines as (1) privacy, (2) fairness and justice, (3) safety and reliability, (4) transparency, and (5) social responsibility and beneficence.

There are some factors in the other industrial domains to achieve trust such as using standard definitions, a system for complaints declaration, and independent rating services (Arnold et al., 2019a). Since trust in new technologies and AI is one of the human concerns, many companies and agencies established formal validation and verification of autonomous robot's software (Ingrand, 2019), surveyed trust metrics and modeling (DiLuoffo and Michalson, 2021; Khavas et al., 2020; Salem and Dautenhahn, 2015), and provided Robotics and AI Roadmap (Robotics Australia Group, 2022; Villani, 2018; White Paper on Artificial Intelligence: A European Approach to Excellence and Trust, 2020; Devitt et al., 2021; Cihon et al., 2021a) to carry out about algorithmic ethics and human-AI interaction in the field of autonomous robotic systems. Furthermore, (Barrué, 2021) provided more information about ethics guidelines and approaches employed in European surveys. A comprehensive research about trust and relationships in AI studied robot imitation, understanding, and AI-human interaction concerning concepts in biology, neuroscience, social psychology, and sociology using outcome matrices as a tool for robot's interactions (C. Wagner, 2009).

In what follows, we elaborate on 7 non-technical and 3 technical metrics for trustworthy AI, as well as some measurement models and frameworks for trust/trustworthiness in AI.

Trust & explainability/transparency/interpretability. In solving complex problems, most AI methods are based on the direct use of complex opaque concepts such as deep neural networks (Meske and Bunde, 2020a). There are some metrics to measure the performance of AI models, such as testing the model on a test set or cross-validation score. However, in most complex AI-based solutions, even the developer has limited access to the mechanism in which the model processes the input data. This opaque nature of complex AI algorithms in turning the input into output is referred to as "black-box" AI (Das and Rad, 2020; Scharowski and Brühlmann, 2020; von Eschenbach, 2021). The trustworthiness of these algorithms has been questioned by many ethical, technical, and engineering communities (Das and Rad, 2020; von Eschenbach, 2021). The pervasive use of deep neural networks in which the number of input features sometimes exceeds thousands of nodes has exacerbated these concerns (Andrulis et al., 2020). Accordingly, AI scientists in recent years have focused on a branch of AI called Explainable AI (XAI), which aims to add explanation, transparency, and interpretation to AI-based decisions by shedding light on the opaque nature of AI methods (Shaban-Nejad et al., 2021a). Studies have shown that XAI can increase the trust of the end-user in AI-based decisions (Zolanvari et al., 2021).

Transparency is one of the fundamental ethical principles in creating trust in users toward AI decisions (Lockey et al., 2021). Although transparency and explainability have been usually categorized under the same ethical principle (Jobin et al., 2019a), it is essential to distinguish between these two different topics before extensive interchangeable misuse of them. Explanations seek broader goals, and transparency (explaining how clearly the system reached the answer) is one of them (Pieters, 2011a; Roth-Berghofer and Cassens, 2005). Research studies have shown that transparency averts overtrusting AI (A. R. Wagner et al., 2018). However, other types of explanations, such as justification, might

lead to users' overtrust by representing manipulative information (Langer et al., 1978). Also, researchers have warned that too much focus on transparency, especially at the early stages of an AI product, can damage innovations (Weller, 2017). In addition, it is worth mentioning that different stakeholders look for different facts in an AI model. Thus, the level of transparency reported to different stakeholders might be different (Felzmann et al., 2019; Varshney, 2019). In Felzmann et al., (2019), the authors divide the stakeholders into five big categories, including developer, regulator, deployer, user, and society in general, and talk about how much detail of transparency they are looking for. Sometimes, even the required level of transparency within one category of stakeholders might be different, for instance, depending on their social geography (Robinson, 2020) or their personality (Gretton, 2018). As a result, defining context-based transparency criteria is difficult to achieve (Weller, 2017).

Building trust is dynamic (Alam, 2020), ranging from initial trust to ongoing trust (W. Wang and Siau, 2018). Needless to say, explanations are a hand-in-hand partner in this dynamic process (Pieters, 2011a). However, the pervasive prevalence of using opaque deep neural networks in AI in recent years has challenged the explainability of the models and, thus, the perceived trustworthiness of the users. These black-box networks are complex and opaque in terms of operation (von Eschenbach, 2021), and even sometimes, the developer has limited access to how they operate. Their complexity is also the underlying reason for their outstanding performance in outperforming conventional solutions and other AI models (Meske and Bunde, 2020b). Thus, there is a tradeoff between the desired accuracy and the level of transparency and explainability, meaning that models with the clearest explanations, such as decision trees, may not have a good performance, while those that are the most accurate, such as deep learning based models, are the least explainable (Agarwal et al., 2021; Holzinger et al., 2017). XAI models have been given momentum recently to mitigate this tradeoff and to open the black box by providing transparency and explainability to AI-based models (Ferrario and Loi, n.d.) in different areas, including medical domains (Muddamsetty et al., 2021; Pawar et al., 2020), robotics (Sakai and Nagai, 2022), autonomous transportation (Glomsrud et al., 2019), and stocks (Carta et al., 2021; Gite et al., 2021; Yang et al., 2018). Recent data regulation set by European Union (EU), known as General Data Protection Regulation (GDPR), has attracted more attention to this field. GDPR recognizes the right of EU citizens not to accept decisions made solely based on automated processing (Thelisson, 2017), which incentivize XAI.

Interpretability is another crucial aspect in increasing users' trust in AI-based decisions (Schmidt and Biessmann, 2019). It should be noted that interpretability and explainability share some common goals, yet they are two different things and should not be used interchangeably (Rudin, 2019). Interpretability aims at making the decision's rationale understandable for the stakeholders, i.e., the relationship between the cause (input) and effect (output) is human-interpretable (Doshi-Velez and Kim, 2017). But explainability is a deeper concept that is concerned not only with the system's inference but also with the model's inner machinery, i.e., how the model works and the way the model is trained (R. R. Hoffman et al., 2018; Masis, 2021). Thus, interpretability can be categorized as a subset of explainability (Gilpin et al., 2018). Kamath et al. (Kamath and Liu, 2021) further categorize the interpretability of XAI methods into three stages: (a) pre-model interpretability, (b) intrinsic interpretability, and c) post-hoc interpretability.

Pre-model interpretability emphasizes the importance of understanding the dataset through exploratory data analysis, data visualization, and feature engineering before model selection (Nandi and Pal, 2022; Okay et al., n.d.; Silva et al., 2019) and the fact that there is no "go-to" model that fits all datasets. Intrinsic interpretability refers to techniques that are intrinsically interpretable due to their structure (Ai and Narayanan, 2021; Pintelas et al., 2020; Stiglic et al., 2020). It ranges from basic models such as decision trees to advanced ones such as explainable boosting machines. Post-hoc interpretable methods refer to methods that utilize the power of complex black-box models in accurate predictions and try to add global or local interpretability to their decisions (Du et al., 2019; Madsen et al., 2021; Molnar et al., 2019; Peake and Wang, 2018). Similar to transparency, some researchers proposed that the level of interpretability of a system should depend on the category of the entity working with the system, e.g., operators, executors, examiners, and the system should reveal a set of suitable measures of interpretability based on their relation to the system (Tomsett et al., 2018). Besides, a consensual definition for interpretability in AI and how to quantify it has not yet been reached (Carvalho et al., 2019; Doshi-Velez and Kim, 2017; Molnar, 2020). Despite these challenges, it is hoped that interpretability will soon reach a state of readiness(Molnar et al., 2020).

Trust & empathy in AI. Empathy is often considered a crucial factor in building trust in all cases, particularly in relation between human users and AI systems (e.g., see Gamer et al., 2010). It is defined as a subjective process in which one comes to have a deep understanding of other people's feelings (M. L. Hoffman, 2000), particularly in a non-judgmental way (Wiesman, 1996), or an "ability to simulate how others subjectively experience a situation and how they regulate elicited emotions" (Gebhard et al., 2021). In fact, empathy involves two main components: the cognitive ability to make an accurate inference of what others think and feel (empathic accuracy) and the affective or emotional ability to make a supportive, benevolent, and compassionate response to their thoughts and feelings (Feng et al., 2004a; Ickes, 1993). A distinction has been drawn between cognitive and behavioral aspects of both trust and empathy, where the latter involves the relevant agent's behaviors, such as a robot's ability to safely lift a patient, and the former involves the agent's cognitive abilities, such as the ability to provide accurate information, make proper inferences, or exhibit an understanding of the other's feelings and thoughts. In these two aspects, trust is essential for efficient communication, and empathy is key to social bonding, which facilitates social interactions, and hence, helps build trust between people, and between persons and AI systems, since social agents are more trusted if they show an understanding of people's mental states (Gebhard et al., 2021). Moreover, empathy has links to expectation: if the AI system can act in accordance with its user's expectations, the user will probably form a trust in the system (Gebhard et al., 2021). Empathic accuracy, as an essential element of empathy, has an impact on the agent's credibility and trust (Brave et al., 2005). Gebhard and colleagues (Gebhard et al., 2021) have identified the following requirements for "empathic cultural-aware agents"; that is, those that take social values and norms into consideration: explainability on both behavioral and motivational levels; observation and detection of social signals such as smiles, facial expressions, and postures; interpretation of utterances and simulation; empathic action and interaction (showing respect for the values and norms of others); adaptability to individuals (showing respect for individual aspects such as their levels of hearing or their dialects). Empathy also has an indirect role in trust through accountability: an agent can be trusted if, among other things, it is accountable, and accountability requires consideration of the stakeholders' viewpoints and needs; that is, empathy (R. Srinivasan and San Miguel González, 2022). Another

significant link between trust and empathy is established by findings about user similarity and trust. For example, users tend to trust online recommendations based on preferences by other users with similar tastes, which are extracted from similar ratings or online purchases and the like (Ziegler and Lausen, 2004). Similarly, it is found that users tend to trust agents with values similar to their own (Mehrotra et al., 2021a). The necessity of transformations in the notions of empathy and trust in patient-doctor relations in the age of AI-based treatments or patient-online communities has been highlighted ((Kerasidou, 2020; Zhao et al., 2013; Montemayor et al., 2021). The link between empathy and trust in the case of online or AI-based services is discussed in Bock et al. (Bock et al., 2020) and Yoon and Lee (Yoon and Lee, 2021).

Trust and privacy. There are obvious tradeoffs between trust and privacy. The idea is articulated in terms of negative associations between online privacy concerns and trust (Araujo et al., 2020a; van Dyke et al., 2007a; Olivero and Lunt, 2004; Reuben, 2018): the higher privacy concerns the lower trust (Culnan and Armstrong, 1999). Privacy is defined as self-determination of when, how, and how much one's personal information or personally identifiable information is communicated to others; that is, privacy is one's control over identity privacy (information leading to identification of a specific person), location privacy (information from which one's location can be identified), communication privacy (confidentiality of one's information), access privacy (control of access privileges), and data processing privacy ("information about the information flow in processing and dissemination of data")(Mehri and Tutschku, 2017). Control over personal information is deemed important in many definitions of privacy (van Dyke et al., 2007b). In addition to privacy concerns, people's assumptions about their self-efficacy in protecting their data play a role in their trust in AI (Araujo et al., 2020b). To provide better services and to keep and attract customers, AI companies need a plethora of information, which depends on detailed information from their customers, but this raises privacy concerns on the part of the customers (van Dyke et al., 2007b). As a result, they lose their trust, and as a consequence, they become reluctant to share detailed or accurate personal information, which in turn lessens the value that the companies were supposed to gain from personal information (Olivero and Lunt, 2004). Moreover, this tends to result in avoidance of online shopping (van Dyke et al., 2007b; Zarifis et al., 2021). An increase in trust leads to increased and more accurate information sharing and decreased perceived risk (Culnan and Armstrong, 1999); (Kok and Soh, 2020). Van Dyke and colleagues propose "customer privacy empowerment" to increase trust and encourage information sharing. The idea is to give customers greater control over how, when, and how much personal information is used (van Dyke et al., 2007b; Spreitzer, 1995). The European Community published the General Data Protection Regulation in 2016, according to which service providers are required to answer user questions about the location of the data, whether their information might be read by others, whether their information is traced, or whether they can revoke permission to use their personal information (Mehri and Tutschku, 2017). This could be done through transparency on the part of the producer or service provider by giving information on the lineage and providence of the product (Arnold, Bellamy et al., 2019b), by giving explanations of the product or service (Pieters, 2011b), or by the social presence and social attributes of the AI system, as in voice assistants (Pitardi and Marriott, 2021b). This is an external legal guarantee for protection of privacy, but "privacy-by-design" embeds privacy requirements in the system's design; e.g., by data minimization, controllability, transparency, easy-to-use privacy function, data confidentiality, technical quality of data, and limited use of data (Mehri and Tutschku, 2017). A significant point here is that since varying degrees of trust are needed in different contexts, different levels of privacy (high, medium, and low) might be required (Mehri and Tutschku, 2017). Technical solutions have been proposed for the tension between privacy and trust (Kok and Soh, 2020), such as secure two-party computation techniques based on homomorphic encryption (Guo et al., 2017), certain blockchain implementations (Sarpatwar et al., 2019), a model that only warrants cooperative AI systems to receive high-fidelity information (Hale et al., 2019), and anonymous authentication and attack tracking (Lu et al., 2019). Some studies show that people are more likely to share information when the technology in question offers a much-needed function or is pleasurable (Ostherr et al., 2017); (Pitardi and Marriott, 2021b), while they are reluctant to do so when it comes to scientific surveys and interviews. Richards and Woodrow (Richards and Hartzog, 2015) suggest that extant privacy laws have "a pessimism problem" in that they excessively focus on harms from privacy infringements and make too much of people's ability to opt out of possibly harmful data practices. In contrast, they propose that privacy should be seen as what enables trust in major information relationships, in which way value is created for all parties to information exchange by establishing a sustainable data relationship.

Trust and fairness in AI. Algorithmic discrimination and bias tend to hinder trust in AI, while perceived fairness or justice enhances trust (Sullivan et al., 2022; Zhou et al., 2021); that is, to trust a system, users should be assured that it can act justly or in an unbiased manner toward all groups (Bartneck et al., 2021). Fairness in AI is equal treatment or equitability of an AI decision about various groups of users (Zhou et al., 2021). Algorithmic unfairness, in many cases, is caused by failure to develop AI systems based on a fair training of data or a fair design of the relevant machine-learning model (Zhou et al., 2021). For instance, in AI-based medical diagnostic systems, discriminations and biases might arise when little or no information from black-skinned and other minorities is fed into the system during its development (Noor, 2020). When it comes to perceived fairness and trust, it was found that people see human decisions as fairer and thus more trustworthy than algorithmic decisions in the case of human (as opposed to mechanical) tasks (M. K. Lee and Rich, 2021a; Hobson et al., 2021). This is not true, however, of minority and marginalized groups: they tend to perceive algorithmic decisions as trustworthy as human decisions (M. K. Lee and Rich, 2021b). Moreover, it was found that perceived fairness is positively related to induced fairness, which is to say that a high degree of induced fairness culminates in a high degree of perceived fairness by people, and the latter is in turn positively related to user trust (Zhou et al., 2021). User biases might also affect their perceptions of trust in an explainable AI system: differences in user trust have been found between malignant and benign diagnoses of an AI system (Branley-Bell et al., 2020). It is also found that perceptions of harm and injustice as well as reported wrongdoing are positively related to uncanniness, which in turn negatively influences trust in an AI agent (Sullivan et al., 2022). Implementation of rules and regulations is deemed necessary for achieving fair, trustworthy AI systems (Kerasidou, 2021a).

Trust and accountability in AI. Contemporary research on accountability in AI and machine learning is mainly focused on defining the rights of human stakeholders, obligations of developers, and ways to enforce them. This approach, known as "offloading," has led to the central concept of the "right to explanation," which demands that AI systems provide

justifications for their actions. This focus on accountability as answerability has led to the development of a regulatory framework and a system design approach that ensures that the AI system can provide the right kind of answers. Thus, the primary aim of accountability in AI and machine-learning research is to define the rights and obligations of stakeholders and to build AI systems capable of providing satisfactory explanations for their actions.

Previous research has identified the need for a robust legal framework for establishing and maintaining trust in artificial intelligence (Leonard, 2018a; Millar et al., 2018; Nalepa et al., 2019). While one aspect of public trust in AI is the reliability of models and the individual recommendations of those models, willingness to trust (and the underlying trustworthiness that willingness tracks) is situated in the context of public trust in institutions (Nalepa et al., 2019). This suggests a two-pronged approach in which researchers work to improve trust in individual models and recommendations and also work to develop a system of minimum standards, verification, and accountability. With regards to the first prong (that of trust in models and recommendations), one component is developing standards of explanation (Shaban-Nejad et al., 2021b). Transparent explanations and accountability are a prerequisite for trust in individual decision recommendations.

The primary focus of contemporary research in accountability in AI research is on offloading questions. That is, the question is one of clarifying exactly what rights human stakeholders have, what obligations AI developers have, and how governments, developers, and watchdogs can enforce this scheme of rights and obligations (Blacklaws, 2018; Bovens et al., 2014; Smith-Renner et al., 2020). Within this project, the so-called "right to explanation" has been central (Ahn et al., 2021; Alam and Mueller, 2021; Ausloos et al., 2020; Binns, 2018; Bovens et al., 2014; Buçinca et al., 2021; Doshi-Velez et al., 2016; Feldman et al., 2019; Smith-Renner et al., 2020; Sperrle et al., 2020; Spiegelhalter, 2020). In one sense, this focus makes sense for a discussion of accountability. It has been argued that accountability is either, at its core or in part, a matter of answerability (Han and Perry, 2020; Williams et al., 2022). If we take the demands of answerability literally in this way, then an accountable system of artificial intelligence will justify its actions (Williams et al., 2022). We can then proceed from the offloading project of accounting for a right to good explanation (whatever we take that to mean) and a regulatory framework guaranteeing it to the agent-building project of designing a system capable of providing the right kind of answers.

Trust and technical metrics (safety, accuracy, robustness). In Jacovi et al. (2021b), a precise discussion is presented regarding the nature of trust in AI, as well as the prerequisites and goals of the cognitive mechanism of trust. Their model, based on interpersonal trust, considers both the vulnerability of the user and their ability to accurately assess the impact of AI decisions. Several technical aspects of trust are proposed to be addressed in AI systems, including reliability, safety (encompassing fairness and explainability), security, and lineage (Bore et al., 2018); (Arnold, Bellamy et al., 2019a). Other studies have shown that providing human-meaningful explanations regarding the system's accuracy can influence user understanding and subsequently enhance trust in AI performance (Nourani et al., 2019; N. Wang et al., 2015). Additionally, these findings suggest that accuracy is a more influential factor than explainability when it comes to improving user trust (Papenmeier et al., 2019).

When it comes to making high-stakes decisions, particularly in fields such as law, medicine, and the military, trust and reliance on AI systems become more challenging. In order to address this, Tomsett et al. (2020) explain the concept of trust calibration, which involves making AI systems interpretable and uncertainty-aware. By incorporating interpretability and awareness of uncertainty, trust in AI systems can be better calibrated. Another model, as described in Y. Zhang et al., (2020), compares the decisions made by AI systems and humans in their respective tasks to determine when to trust or distrust the AI. This model helps establish guidelines for understanding the appropriate level of trust to place in AI systems. Additionally, (Okamura and Yamada, 2020b) present an adaptive trust calibration approach for human-AI interaction to analyze instances of over-trust in AI.

In Xu et al. (2021), a novel sparse decision-making model is proposed that integrates trust and information rating. This model takes into account both trust and the quality of information when making decisions. Several articles propose various mechanisms to increase trust, such as supplier's declaration of conformity (SDoC) for AI services (Bore et al., 2018) or the use of FactSheets (Arnold, Piorkowski et al., 2019), which are filled out by both AI service providers and users. These mechanisms aim to enhance transparency and accountability, thereby fostering trust in AI systems. In the context of trusting the evolution of 5G internet services, a conceptual zero-touch security and trust architecture has been proposed (Carrozzo, 2020). This architecture aims to ensure secure and trusted communication in the 5G network. Additionally, it has been suggested that combining diversity (utilizing network nodes with different characteristics) and trust (immunity from failures and attacks) can enhance the structural robustness of sparse networks (Abbass, 2019b).

To address trust and knowledge sharing in graph models, a blockchain-based approach has been introduced (J. Li, Wu, et al., 2021). This method facilitates the sharing of trusted knowledge, isolates malicious nodes, and prevents knowledge pollution, thereby promoting reliable information exchange. The utilization of AI has demonstrated increased efficiency in various tasks [e.g., Rahman et al. (Mizanoor Rahman et al., 2016; Maurtua et al., 2017)]. However, it should be noted that trust in AI is not guaranteed in the realm of cybersecurity. Nonetheless, it is argued that trust can play a role in improving the design, development, and deployment of AI systems (Taddeo et al., 2019).

Several articles have developed structured models with mathematical definitions to explore various aspects of trust. These models include parameters such as trust system space, maximal and intuitive attacker models, and robustness properties (Muller et al., 2014). Other models focus on advisors hiding or minimizing their true observations (N. Wang et al., 2015), a Bayesian-based trust model for human multi-robot teams (Fooladi Mahani et al., 2020), and factors related to both humans and robots (Khavas et al., 2020) to discuss and formulate trust robustness. In Vodrahalli et al. (2021), a psychological metric called the weight of advice (WoA) is employed to analyze human-AI interactions when advice is provided by both human and AI sources. The study reveals that participants' behaviors are similar in both cases, but the level of trust varies depending on the topic of the advice. Another study aims to evaluate human trust in AI by examining the relationship between psycho-physiological states, such as biosignals or physiological signals, and trust and cognitive load (Gupta et al., 2019). The study explores how physiological indicators can provide insights into the level of trust individuals have in AI systems. Overall, these studies contribute to the understanding of trust by utilizing mathematical models, psychological metrics, and physiological signals to examine the robustness and dynamics of trust in various human-AI interactions.

Evaluating and measuring/trustworthiness certificate in AI. To assess the trust between humans and AI and establish accuracy

and safety guidelines for AI-assisted decision-making, numerous psycho-physiological approaches (e.g., (Ajenaghughrure et al., 2019; S. Bhatti et al., 2021)) and empirical approaches (e.g., (Chandra, 2010; Oh et al., 2019; Okamura and Yamada, 2020b)), supported by theoretical methods, have been proposed. These approaches are aided by the use of questionnaires, experimental protocols, qualitative evaluations, and other evaluation techniques. However, several challenges can affect the validity and accuracy of these investigations. Firstly, it may be difficult to encompass all trust factors within questionnaires, experimental protocols, and qualitative evaluations. Additionally, the diverse designs and models of trust, coupled with the dynamic nature of trust influenced by experimental constraints and the methods employed by the AI system, present challenges in developing comprehensive guidelines and protocols (R. Hoffman et al., 2021) (Vereschak et al., 2021). The complexities surrounding the evaluation of trust are further explored in (Hurlburt, 2017a), offering insights into the associated problems.

In Chandra (2010), a trust-theoretical model analyzes consumer trust in mobile payment (m-payment) services, shedding light on user trust in m-payment systems. The verification and validation of autonomous systems are debated in (Cho et al., 2015; Lyons et al., 2017), with a focus on transparency and sharing awareness between designers, testers, and users in the development of transparent AI systems. Cho et al. (Cho et al., 2016a) identify key attributes of trustworthiness (such as reliability, safety, resilience, and agility) in relation to trust. The framework of ontology-based trustworthiness considers vulnerability, errors, and the relationships between these factors to establish a threshold of confidence for AI systems.

In Cho et al. (2019a), the quality evaluation of computer-based systems is conducted using the aforementioned metrics, incorporating vulnerability and risk assessments. The study aims to identify future research directions and enhance the metrics and methodologies employed. The theoretical aspects of trust in AI within the manufacturing industry are explored in (J. Li, Zhou, et al., 2021c), which categorizes trust into three levels: organization, group, and individual. These levels involve factors such as management commitment, authoritarian leadership, and trust in AI promoters.

A psycho-physiological model for assessing user trust in AI is proposed by Ajenaghughrure et al. (2019), seeking to determine which user signals provide accurate assessments of trust. In evaluating trust in human-AI interaction, (Schmidt et al., 2020a; 2020b) find that participants prefer physical interaction and embodiment with AI rather than relying solely on voice control. Another study introduces multi-dimensional metrics, including user satisfaction, to assign a trust score to an AI system. This trust score encompasses factors such as job efficiency and effectiveness, understanding, control, and data protection (J. Wang and Moulden, 2021).

Hoffman et al. (R. Hoffman et al., 2021) discuss trust scales in AI and emphasize two key aspects: trust in the output and reliance on machine advice. They provide a comprehensive review of various trust assessment scales and suggest that recommended scales should focus on the predictability, reliability, efficiency, and believability of AI systems. (S. Bhatti et al., 2021) employs behavioral and physiological measures, such as individual and team performance scores, team situation awareness, and process measures, to evaluate human-AI interaction and trust in AI. The study reveals that trust levels differ between human-human and human-AI interactions, and interestingly, trust scores for human-human interaction increase in degraded scenarios, in contrast to human-AI interaction.

(S. S. Lee, 2021a) conducted an influential study that offers comprehensive and well-structured explanations regarding the philosophical evaluation of trust. The research puts forth a rational argument stating that trust in AI is impossible due to its complexity and inexplicability. However, the study highlights the importance of value-based trust, which can be derived intuitively from the information obtained through decision-making algorithms and their implications, such as diagnosis accuracy and safeguarding. The significance of AI certification is also discussed, emphasizing the inclusion of ethical principles and mandatory conformity assessment. This certification process aims to enhance algorithmic auditing, facilitate customization of AI certification, and establish educational programs addressing AI and its safety concerns. Applying ISO standards to the quality and security management of AI in specific processes is seen as a valuable approach to ensuring AI's reliability and safety (Cihon et al., 2021b).

Trustworthy AI frameworks. Banavar (Banavar, 2016) developed a framework centered around secure and morally sound AI systems that aim to cultivate trust through repeated interactions. Nonetheless, he emphasized the importance of algorithmic accountability, adaptability, precise integration of data, algorithms, and AI systems, as well as safeguarding privacy within this broader understanding of trust. Given that the ethical framework is crucial for human–AI interactions, as discussed by Cihon et al. (2021b), Hauer (2021) has adopted this framework to examine a model for human-robot interactions. It was emphasized that AI decisions are influenced by human judgment. The relationship between ethics and AI can be summarized into three aspects: ethics by design, ethics in design, and ethics for design (Dignum, 2018). In a study by (Jobin et al., 2019b), five ethical principles—transparency, justice and fairness, non-maleficence, responsibility, and privacy—were highlighted to encourage a global convergence in integrating AI system guidelines. Additionally, another related study proposed five foundational principles—beneficence, non-maleficence, autonomy, justice, and explicability—to develop a data-based framework for trustworthy AI (Thiebes et al., 2021c). The ethics of AI in global health, as explained in (Kerasidou, 2021b), revolve around the metrics of explainability, algorithmic bias, and trust, raising important questions regarding value, fairness, and trust. For a comprehensive review of AI ethics guidelines, the practical implementation of AI and ethics, and advancements in AI ethics, interested readers are referred to (Hagendorff, 2020).

Through an examination of various machine-learning approaches in air traffic management, researchers (Hernandez et al., 2021) devised an explainable framework aimed at enhancing trust in AI. Their automated method operates by leveraging existing guidelines and incorporating user feedback to bridge the gap between research transparency and practical explainability. In a separate study, (Shaban-Nejad et al., 2021a) focused on the explainable AI framework in public health and medicine domains, emphasizing the metrics of fairness, accountability, transparency, and ethics. Furthermore, these four factors are deemed crucial in obtaining a social license and fostering trust in data (Leonard, 2018b).

A computing architecture was developed in order to explore the similarities between human and AI decision-making processes using an effective trust model (DAngelo et al., 2015). Within their theoretical framework, the Naïve Bayes method was utilized to classify final decisions, taking into account behavioral patterns derived from human-AI interactions. Furthermore, a theoretical framework was presented to generalize trust antecedents in AI-based conversational agents across various contexts (W. Wang, 2021a). In industrial settings, a normative framework known as the "invitation of trust" was employed, considering cultural factors and focusing on conversational AI (J. Kim, 2021).

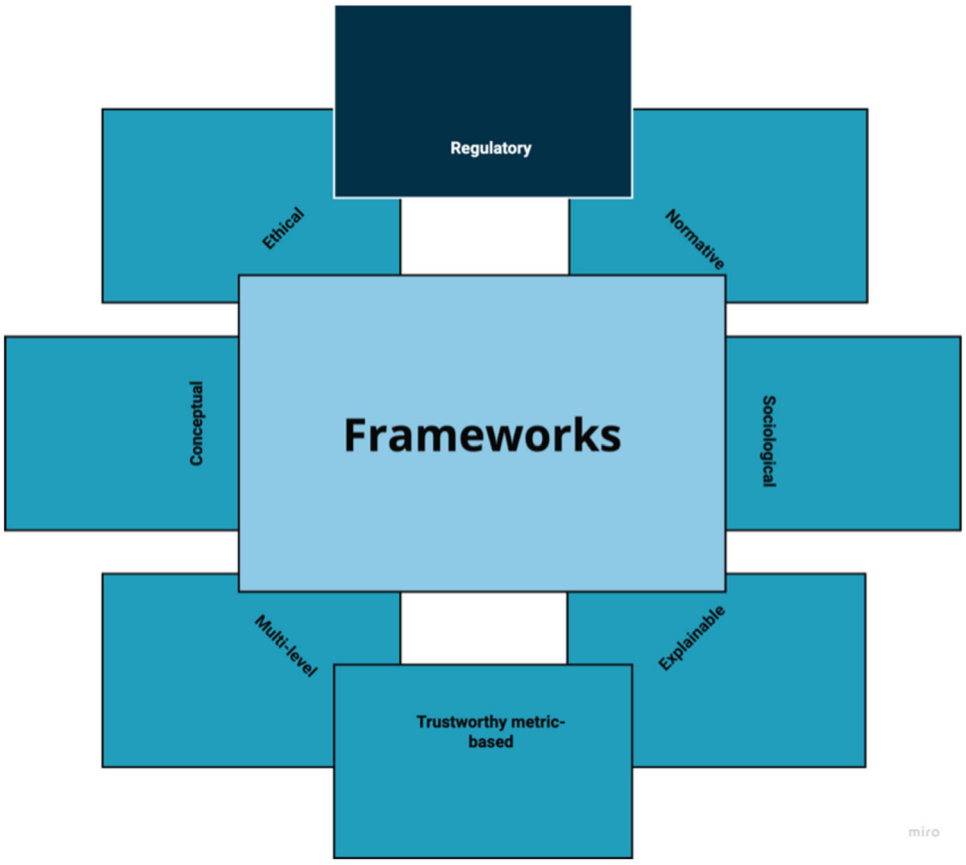

Fig. 4 Different frameworks that can be employed in trustworthy AI.

Autoregressive models were applied to extensive datasets, encompassing dialogs, and negotiations, to facilitate an ethical and automated training process for AI chat systems. Trust in technology and its acceptance were the subjects of a theoretical study that aimed to address the risks associated with AI usage, emphasizing the importance of user training (Eigenstetter, 2020). Additionally, The National Institute of Standards and Technology (NIST) is actively working on a risk management framework tailored for artificial intelligence (AI, 2023). This framework aims to equip organizations with best practices and a common language to effectively handle the risks linked to AI across the entire organizational life cycle.

In a conceptual framework presented in (Guckert et al., 2021), the authors initially addressed the issue of trust in AI systems when the processes involved are not adequately understandable and traceable. They then conducted an analysis using two different datasets related to urban logistics planning and heart arrhythmias. The purpose of this analysis was to demonstrate how the identification of patterns can enhance trust in human–AI interactions, emphasizing the importance of examining the results and conducting thorough inspections.

The explainability and trustworthiness of AI pose numerous challenges, making it difficult to make decisions regarding the acceptance of outcomes, as discussed in (Pickering, 2021). To address this, the paper proposes three scenarios: tracking contacts, analyzing big data, and conducting research during public health emergencies. These scenarios aim to establish a consent-based trustworthiness process. Additionally, the framework proposed in (Cho et al., 2016b, 2019b) is designed to operate based on trustworthy metrics mentioned in the preceding subsection.

In Lyons et al. (2017), a concise overview of different frameworks addressing the differentiation between human labor and AI labor is provided, along with the importance of understanding how AI systems operate. Furthermore, (Kaur et al., 2021) offer a brief review of the principles outlined by the European Union for trustworthy AI, summarizing the approaches and requirements for establishing trustworthiness in such systems. A redefined multi-level framework for robot autonomy in human-AI interactions is presented in (Beer et al., 2014a), aiming to provide guidelines on how different levels of robot autonomy can impact variables such as acceptance and reliability. (M. Ryan, 2020b) argues that instead of trust, the concept of reliance should be used when referring to AI, as AI lacks emotional states and responsibility for its actions.

Within a sociological framework, Jacovi et al. (2021b) introduce a trust model that revolves around two key factors: the vulnerabilities of the user and the ability to predict the consequences of AI decisions. The article defines concepts such as contractual trust, warranted trust, and unwarranted trust, offering a formalism for designing trustworthy AI that is based on warranted trust.

The European approach towards establishing global trust involves the development of ethics guidelines within a regulatory framework for AI. These guidelines are designed to foster an ecosystem of trust that includes policy, fundamental rights, and consumer rights. They outline seven essential factors that are vital for ensuring trustworthy AI: 1-Human agency and oversight, 2-Technical robustness and safety, 3-Privacy and data governance, 4-Transparency, 5-Diversity, non-discrimination and fairness, 6-Societal and environmental well-being, 7-Accountability (EUROPEAN COMMISSION, 2020). To put these trust measurement approaches into practice, several frameworks have been suggested, which can be seen in Fig. 4. These frameworks act as guidelines for effectively integrating and assessing the key elements of trust in AI systems.

Distrust in AI and scary AI. Distrust in artificial intelligence tends to be attributed to a variety of reasons. At the macro-level, there are general fears about the power of a machine with artificial general intelligence (AGI) (Baum, 2017; Bostrom, 2014; Hurlburt, 2017b). AGIs are hypothetical artificial intelligence systems possessing broad plastic intelligence (like our own) as opposed to task-specific algorithms. These concerns worry that because an AGI would be able to process information faster than its human counterparts and could have access to the full domain of human knowledge available on the internet, they would be able to outcompete their human creators (Bostrom, 2014). Outside of fears of a malevolent artificial intelligence, however, others worry about the so-called "alignment problem" of how we, as a society, and researchers working on artificial intelligence research could ensure that an AGI's interests and values would align with our own (Abbass, 2019c; Bostrom, 2014; Gabriel, 2020; Prasad, 2019; Taylor et al., 2016). Nevertheless, fears that lead to distrust in artificial intelligence are not limited to concerns about AGI. Distrust in AI is attributable based on a variety of factors including (but not limited to) concerns about surveillance, privacy, hackability, autonomous technologies in the defense and transportation sector, and the potential impact of decisions made either directly by or informed by artificial intelligence algorithms (Akkara and Kuriakose, 2020; Tschopp, 2019). According to one study, the top concerns of those who were distrustful of artificial intelligence were (in order) its uses in war, loss of human control, issues of privacy, applications in healthcare, and potential consequences of artificial intelligence for the economy (Tschopp, 2019). Following the concerns about security and defense applications of artificial intelligence, some have argued that we are right to be distrustful of these uses and that decision-making by these algorithms should be heavily circumscribed (requiring human input) (Taddeo et al., 2019). Accordingly, distrust in artificial intelligence tends to increase as the stakes of decision-making increase (Ajenaghughrure et al., 2020). Given the high stakes for patients in using artificial intelligence to make diagnoses or suggest treatments, considerable attention has been paid to how to reduce distrust in healthcare settings (Alam and Mueller, 2021; Asan et al., 2020; Feldman et al., 2019; Ross, 2020). However, it should be noted that trust in artificial intelligence in healthcare settings can sometimes outpace trust in human doctors and that this effect is gendered, which raises its own ethical concerns about the uses of the technology in a healthcare environment (D. K. D. Kim and Kim, 2021b). Finally, there are more general moral concerns about particular applications of artificial intelligence, which roughly map on to concerns about alignment problems for AGI. Here, some worry that, when facing novel circumstances, an artificial intelligence program might exploit vulnerabilities in order to achieve its goals rather than report them (Hurlburt, 2017b). Whereas we might expect (for better or worse) a human competitor to be bound by moral considerations when facing novel circumstances, the response to which is underdetermined by the rules of the competition, a computer may not be so reliable. Whether or not we are right to think that humans are, on average, trustworthy under conditions of competition, it is harder to justify trust in artificial intelligence systems that have not been trained or instructed on what to do under those circumstances.

Distrust makers in AI systems. In what follows, we will elaborate on three major classes of distrust makers in AI systems: surveillance & manipulation, human autonomy & dignity threat, and unpredictable futures (see Fig. 5).

Surveillance and manipulation. As mentioned above, issues of surveillance and privacy are among the top concerns for those who distrust artificial intelligence (Tschopp, 2019). In this case, Applications of artificial intelligence are distrusted because their widespread use might make information about private individuals susceptible to surveillance or data theft (Chen and Wen, 2021). In the case of surveillance, distrust in artificial intelligence carries over from distrust in the parties employing the technology (Chen and Wen, 2021; Jackson and Panteli, 2021; Spiegelhalter, 2020). Distrust of artificial intelligence, in this domain, can range from concern over whether the company or government using the algorithm is trustworthy in what they say about its use to concern over what is said by the algorithm and why it came to that recommendation (Spiegelhalter, 2020). These issues suggest that distrust in artificial intelligence is tied to distrust in a particular instance's developers and users. Accordingly, trust in artificial intelligence is a composite of trust in the program itself and in the general scientific and institutional community around artificial

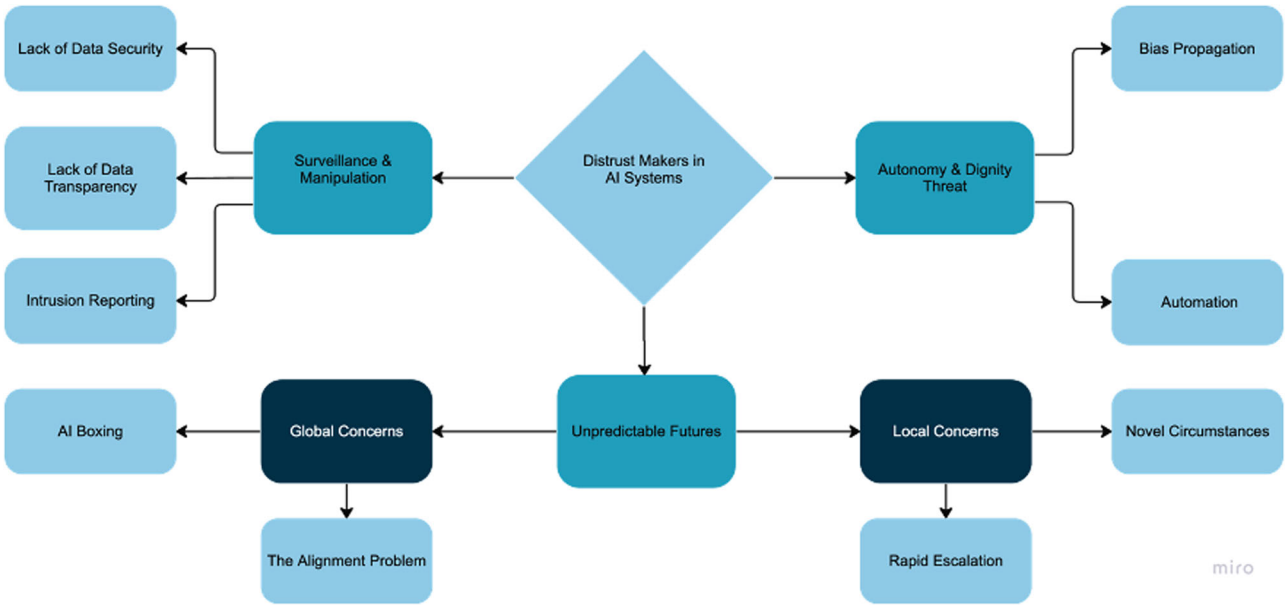

Fig. 5 Three major classes of distrust makers in AI systems.

intelligence (Chen and Wen, 2021). Further still, distrust in artificial intelligence can also be rooted in distrust of government, even in private applications. This has led some to argue for a layered model in which users first come to trust the government which regulates artificial intelligence developers, and then trust the corporate culture, interest, and oversight within the companies that serve as artificial intelligence developers, before coming to trust specific applications or recommendations made by specific algorithms (Jackson and Panteli, 2021). As for manipulation, distrust in artificial intelligence can be founded on concerns about cybersecurity. For instance, a demonstration by researchers revealed that hacking into the dataset of an artificial intelligence program used in a healthcare setting could lead to widespread false detection of cancerous lesions (Akkara and Kuriakose, 2020). The potential human cost of systematic misdiagnoses is raised as another contributor to distrust of artificial intelligence systems. Concerns like these have motivated some work on creating artificial intelligence systems, which detect and report outside modification (Abbass, 2019c). However, we should also recognize that corrupted datasets are not always the result of outside manipulation, and might merely be incomplete, unbalanced, small, or inaccurate (Hurlburt, 2017b).

Human autonomy/dignity threat. When it comes to issues of autonomy and dignity, the most prevalent concerns about AI are either (1) that these algorithms will only reify and propagate existing biases and inequities or (2) that artificial intelligence will supplant human agency in part or in total. That is, distrust in AI in these cases is grounded in skepticism about artificial intelligence's ability to preserve the dignity of all humans and/or human dignity as such. In the first case, considerable attention is and should be paid to the lessons that machine-learning (ML) algorithms learn, which might inherit our own societies' biases (Asan et al., 2020; Sperrle et al., 2020). As an example, an algorithm designed to predict individual recidivism rates so as to help inform sentencing and parole decisions in the criminal justice system might propagate an existing social bias on the basis of skin tone (Hurlburt, 2017b). If police are more likely to patrol and make arrests in predominantly black neighborhoods, then the dataset coded with an increased rate of recidivism will likely follow suit. Thus, even if the artificial intelligence is not specifically looking at race, its dataset will have encouraged it to associate facts about race with facts about recidivism. With regards to human dignity in general, and the issue of supplanting human agency, the issue is among the highest rated concerns of those distrustful of AI (Tschopp, 2019). Outside of the general concern, distrust in artificial intelligence can be rooted in domain-specific intrusions. For instance, despite not always trusting AI, people do sometimes trust artificial intelligence algorithms more than humans (including in healthcare and governance) (Ingrams et al., 2021; D. K. D. Kim and Kim, 2021b). The degree to which workers in particular domains find meaning in their work is the degree to which they might perceive the influence of artificial intelligence as pernicious. This effect is amplified when human patients (for instance) come to trust algorithms more than they trust human doctors. The insult to dignity is only made deeper when this imbalance of trust is distributed unevenly along demographic lines (as is the case for female doctors) (D. K. D. Kim and Kim, 2021b). Potential solutions to this problem have been proposed, and they often focus on bringing human users into relationship with the artificial intelligence (Beer et al., 2014b; Tschopp, 2019).

Distrust and unpredictable futures. Another issue contributing to distrust in AI is unpredictability. Concerns about unpredictability come in global or local varieties. For instance, global concerns include things like those mentioned above, over the large-scale implications for society of machines utilizing artificial generalized intelligence (Baum, 2017; Bostrom, 2014; Hurlburt, 2017b). One species of these worries (again, as mentioned above) concerns the two-pronged project of (1) making sure that an artificial generalized intelligence system has values and interests aligned with those of humans (Abbass, 2019c; Bostrom, 2014; Gabriel, 2020; Prasad, 2019; Taylor et al., 2016) and (2) determining how we could be assured that we succeeded in ensuring this alignment. Likewise, others have suggested that mitigation strategies should be put in place using so-called "AI boxing" in order to ensure that large-scale social damage is avoided in cases where researchers erroneously believe they have succeeded at both projects (Chalmers, 2010). Meanwhile, local concerns include worries about the unpredictability of artificial intelligence systems in specific recommendations or under specific circumstances. One version of this is those, brought up above, of so-called "dirty tricks", in which novel circumstances are exploited rather than reported (Hurlburt, 2017b). This need not involve circumstances of competition, as without training on how to handle novel cases, machine-learning algorithms might simply use personal data in ways that human analysts might not (Johnson, 2020; Leta Jones et al., 2018). Another version of these worries includes uses of AI technology in autonomous weapon systems (AWS) (Johnson, 2020). In these cases, AI behavior under novel circumstances (including the so-called "drone swarming") might lead to conflict escalation too rapidly for humans to intervene so as to avert unsafe or catastrophic outcomes (Johnson, 2020).

Challenges and barriers to breaking distrust. There are many challenges and barriers to reducing distrust in artificial intelligence systems. For one, trust in these systems requires some amount of transparency (Sperrle et al., 2020; Tutul et al., 2021b). However, what this norm of transparency entails is less clear. For instance, users and stakeholders are often unmotivated to engage in explanations (Buçinca et al., 2021). In healthcare settings, the utility of these explanations is time-sensitive (Alam and Mueller, 2021). Explanations of how an artificial intelligence system arrived at a particular recommendation are helpful at critical times, but providing explanations at non-critical times tends to diminish trust overall (Alam and Mueller, 2021). Likewise, how the explanations are delivered will alter how effective they are at increasing trust. Written narrative explanations are not often as effective as visual representations and local explanations of specific reasons for recommendations tend to be more effective than global explanations of the artificial intelligence system making the recommendation (Alam and Mueller, 2021). Still, there is skepticism about how helpful explanations are in general (Ahn et al., 2021; Feldman et al., 2019). As Feldman points out, patients are trusting of healthcare interventions in medical science despite often not understanding the underlying mechanisms at work (Feldman et al., 2019). We might wonder why we should think that the relative distrust in artificial intelligence is attributable to a lack of explanatory transparency. One challenge is then settling on the right set of conditions for satisfying the optimal level of transparency. Clarifying the conditions required for a good and helpful explanation is a project currently being undertaken (Buçinca et al., 2021; Sperrle et al., 2020; Spiegelhalter, 2020). Likewise, there is a challenge posed by trying to actually assess how trustworthy an algorithm is. Developing models of trustworthiness is an ongoing project (Jiang et al., 2018; Skopik et al., 2009; Spiegelhalter, 2020). Finally, while there are barriers and challenges to establishing trust in AI, there is also a problem of arriving at the optimal level of trust. People often over-rely on artificial intelligence systems, trusting them too much (Buçinca et al., 2021). For instance, the driver of an autonomous car in

Florida crashed into a truck because they had over-trusted the artificial intelligence system steering the car (Hurlburt, 2017). They stopped paying attention to the road and began watching a film during the drive. On top of the challenge of measuring trustworthiness, we then also have the challenge of finding the optimal level of trust and developing interventions, which can push users in that direction as well.

Trust makers: building/increasing trust in AI. Trust is the essential component for humans to accept AI technology and adopt it in different domains. Hence, technology owners and developers seek strategies to either increase the trustworthiness of their AI systems or enhance end-users' trust. As mentioned in the section "Methodology", trust and trustworthiness are two different phenomena, and one does not necessarily grant the other. Various technical and axiological factors could increase the trustworthiness of AI models, while the literature has paid more attention to technical factors such as explainability and accuracy to enhance trustworthiness. Although trust could be improved as trustworthiness increases, there exist specific trust engineering techniques that only focus on building trust without considering the features of the AI model and its trustworthiness. In this section, we first introduce the overall factors that could affect the trust and trustworthiness of AI to understand the required basis for building trust. We will then review methods of building trust in AI as used in previous research. Finally, we will introduce a few case studies in the different domains and explore their respective influential factors for building trust.

Factors that affect trust. The factors that impact trust in AI systems could be categorized as technical and axiological factors. From another perspective, these factors could be divided into human-related, AI-related, and context-related factors, where the latter is mostly related to particular requirements of a specific application and the developers' characteristics. Among the technical AI-related factors that influence trust, transparency and explainability have been widely investigated since black-boxes are generally less trustworthy (Ashoori and Weisz, 2019).

Among human-related factors, understanding the technology, expertise, culture, and personal traits have been found significant (Kaplan et al., 2021). There are conflicting results about the effect of gender, where it was found effective in (Kaplan et al., 2021) but not significant in (Khalid et al., 2016). However, education and age do not play an important role in building trust. Among AI-related factors, performance and reliability have been significant, along with AI personality, anthropomorphism, reputation, and transparency. Finally, team-related factors and risk of the task have been found significant among the context-related factors, where higher risk leads to lower trust (Kaplan et al., 2021). Although several studies have talked about the individual-related factors, findings of one survey study with 226 participants to measure the relative advantage of AI-based advisory over human experts in the context of financial planning suggested that the implementation of human traits was negligible while the ability to test the service noncommittal was superior (Mesbah et al., 2019). While the AI-related factors mostly focus on improving the capabilities of the AI system, the other factors could change trust even when the capabilities of the system and its trustworthiness have not changed. This could lead to over-trust or under-trust, as shown in Fig. 6, where the former could cause damages and the latter leads to less adoption of the AI systems (Asan et al., 2020; Alambeigi et al., 2021). Under trust happens when due to axiological factors such as lack of documentation or good reputation of the developers, the level of trust is lower than the actual capabilities of the AI system. On the other hand, over-trust

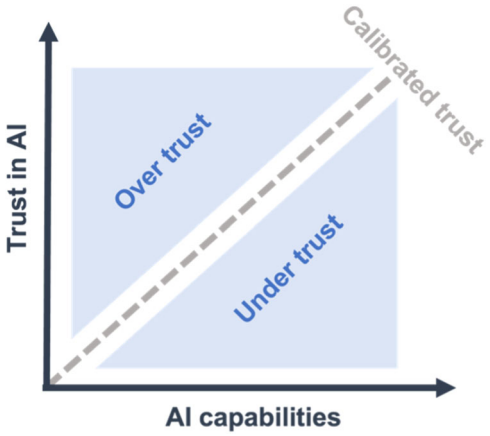

Fig. 6 Calibrated trust axis (Asan et al., 2020).

happens when the trust is higher than the system's actual capabilities, which could happen due to misperceived trustworthiness. In fact, some of the axiological factors, such as reputation, human agency, oversight, and accountability, could be engineered in true branding, marketing, or other ways. These methods could then lead to over-trust if the AI's ability is not aligned with those factors. Other categorizations of the influential factors have also been proposed in the literature based on subgroups of human-related, AI-related, and context-related factors (Wang, 2021b).

Methods of building trust in AI. Different methods for building trust have been suggested in the literature, where these methods focus on technical factors to enhance trustworthiness or axiological factors that focus on trustors' characteristics to enhance trust (Siau, 2018). The former methods consider aspects such as model performance, transparency, and explainability. The latter, however, focuses on building trust through accessibility, preparing comprehensive documentation and regulations. From a technical perspective, for an AI system to be trustworthy, technology creators should ensure that the data acquired, processed, and fed into the algorithm is accurate, reliable, consistent, relevant, bias-free, and complete. Similarly, the selected, trained, and tested algorithm should be explainable, interpretable, transparent, bias-free, reliable, and useful (Srinivasan, 2019).

There is an extensive body of literature discussing the explainability and interpretability of AI as one of the most important factors that affect the trustworthiness of AI (Caspers, 2021; Ezer et al., 2019; Jacovi et al., 2021; Li et al., 2020; Mahbooba et al., 2021; Sengupta and Chandrashekhar, 2021; Srinivasan, 2019; Yan and Xu, 2021b). Explainability becomes extremely important in critical applications such as healthcare, where the decision made by the AI would not be reliable unless it is justified based on common medical knowledge. It is worth noting that a bad explanation for trust may fail to create trust. In other words, too little detail does not explain trust, and too much detail becomes confusing for users to trust (Pieters, 2011c). Another technical method for building trust is to provide confidence level and the AI's decision. When the AI system reports its confidence in its decision, it allows its users to judge how reliable this decision is, and it significantly increases trust (Bruzzese et al., 2020).

Explainability can be provided on two levels, including global and local explainability. The former explains the overall behavior of an AI model, while the latter explains its decision process in response to a specific input. It was shown that the global explanations about the process had no impact on immediate satisfaction and trust but improved later judgments of

understanding about the AI. On the other hand, local justifications were found effective, but their effect is time-sensitive. For instance, during a critical situation or when AI was making errors, local justifications were very effective and powerful explanations (Lui and Lamb, 2018).

Presentation of the results and explanation of AI systems could also affect trust. AI systems need to meet a certain level of performance criteria, they need to be explainable and interpretable, they need to consider fairness and biases in their design and evaluation. However, the way that an AI system communicates its results with human agents has a direct effect on trust. A study showed that interactive visualization is a technology that helps to increase trust in AI systems (Oelke et al., n.d.). Another study found that users had significantly more trust in the explanations that were presented by a human agent (Miller et al., 2017). In the healthcare domain, it was shown that visual and example-based explanations integrated with rationales had a significantly better impact on patient satisfaction and trust than no explanations or with text-based rationales alone (Lui and Lamb, 2018).

One of the non-technical methods of building trust through generating and sharing transparent, clear, and comprehensive documentation is the supplier's declaration of conformity (SDoC) (Hind et al., 2018). SDoC for AI increases trust by focusing on providing cues to the trustors to understand the system's characteristics better to assess if they will get what they expect from the AI system. The availability of accurate and relevant cues is necessary for the trustworthiness of the AI system to be perceived correctly (Schlicker and Langer, 2021). SDoC is a transparent, standardized, but often not legally required document used to describe the lineage of a product along with the safety and performance testing it has undergone. SDoC gains trust since it shows the process or service conforms to a standard or technical regulation. This document contains sections on performance, safety, and security. It also explains how the system was created, trained, and deployed, along with what scenarios it was tested on, how it will respond to non-tested scenarios, guidelines that specify what tasks it should and should not be used for, and any ethical concerns of its use (Hind et al., 2018). This level of transparency and detail concerning every aspect of the system, especially the evaluation process, helps increase trust, but mostly for expert users who know how to interpret the metrics provided in the fact sheet. However, this technique may discriminate against patients from low literacy backgrounds who are less used to interpreting statistical risks (Lee, 2021b). Therefore, in addition to SDoC, there is a need for expert agencies to assess these documents so that non-expert end-users can rely on their assessment. In this case, the experts' endorsement can only function on a principle of value-based trust since this endorsement provides no extra functional information. Trust could be more prominent when this expert agency is the government or a well-known regulatory entity in which people trust. A diverse group of stakeholders could develop and define standards for promoting trust, as well as AI risk-mitigating practices through greater industry self-governance, and adherence to such standards could be verified, specifically through certification/accreditation (Roski et al., 2021).

Some believe that the public distrust in AI originates from the under-development of a regulatory ecosystem that would guarantee AI's trustworthiness (Knowles and Richards, 2021). They argue that being accountable to the public through elaborating rules for AI and developing resources for enforcing these rules is what will ultimately make AI trustworthy enough. Based on this theory, building public trust in AI is not simply a case of creating explainable AI or standardizing various performance metrics for AI components. Instead, public trust requires some authority that urges organizations to take ethical responsibilities seriously and to validate their interpretations of these standards.

Value-based trust also suggests presenting evidence to persuade users that their positive values, for example, inclusion, confidentiality, good-will, are encompassed by developers and governing bodies. This would, in practice, be achieved through the elimination of biases (Lee, 2021b). This could be achieved by requiring either the expert agency or the AI developers to show evidence of previous ethical conduct in data privacy and usage.

To improve trustworthiness, Roszel et al. proposed 20 guidelines that provide clarity on different influential factors, namely, efficacy, reliability, safety, and responsibility of a given AI system (Roszel et al., 2021). However, it is important to mention that overstating the role of ethics in corporations' policy, culture, and advertisements, known as ethics-washing, in order to avoid or escape governmental regulations and convince and reassure customers to keep with the company's products or services hurts trust (Peukert and Kloker, 2020). In fact, if users perceive that a company is only pretending to comply with ethical guidelines, they may build less trust in that company. These findings suggest that companies should be aware of the issues of ethics-washing regarding their AI services and should try to avoid ethics-washing in their marketing communication.

Case studies and items effects on building trust. Leveraging the aforementioned methods of building trust also depends on the unique requirements and context of different application domains. In the field of marketing, it is crucial to understand how consumer adoption of the information generated by AI can be improved. One of the important aspects of building trust in marketing is understanding the psychological factors that influence consumers' acceptance of AI-generated information and how to induce more favorable consumers' responses about their AI-generated information and marketing. In this domain, the axiological factors could outweigh technical ones such as model performance and explainability. The relationship between number presentation details and users' trust in AI-based marketing was investigated to improve trust as the authors believed that the number preciseness of the recommendation information would critically influence consumer responses to AI technology (Kim et al., 2021). The results of this study showed that the use of a precise (vs. imprecise) information format leads to higher trust. Moreover, when the product's objective quality is high (vs. how), information preciseness strongly influences consumers' trust and purchase intentions. Also, interestingly, when the accuracy of the information is low (vs. high), information preciseness has a stronger influence on consumers' responses. This is a perfect example of the importance of customizing trust-building methodology based on the unique requirements of the application domain rather than using a general solution.

Even for the well-known and general solutions for building trust, such as improving transparency and explainability, the implementation of these methods could significantly regulate their impact on trust. For example, in the case of explainability, the impact of virtual agents on the perceived trustworthiness of autonomous intelligent systems was discussed (Weitz et al., 2019). It was found that the integration of virtual agents into explainable AI interaction design leads to an increase of trust in the autonomous intelligent system in the particular application of speech recognition. Overall, users had significantly more trust in the explanations that were presented by the agent. The users found the system to be less deceptive, more trustworthy, and less worrying when the explanation results were presented by the agent. This is a great example of using context-based factors to improve trust rather than focusing on the technical aspects of explainability.

Recently, there has been a considerable amount of interest in blockchain technologies. In this area, technical factors of trust and models of trust are important since AI–AI interaction is prevalent in this domain. A platform where consumers and data providers can transact data and/or models and derive value was proposed considering trust complications, given that preserving trust during these transactions is a paramount concern (Sarpatwar et al., 2019). This study focused on the use of blockchain technology in the field of transfer learning, where a consumer entity wants to acquire a large training set from different private data providers that match a small validation dataset provided by the consumer. Data providers expect a fair value for their contribution, and the consumer also wants to maximize their benefit. To gain consumers' trust, this platform focused on AI-based factors. They implemented a distributed protocol on a blockchain that provides guarantees on privacy and consumer benefit that plays a crucial role in addressing the issue of fair value attribution and privacy in a trustable way.

Some studies tried to understand all the human-related and AI-related requirements of trust in AI-driven chatbots (Zierau et al., 2020). This study leveraged surveys from end-users and experts and came up with several guidelines and design principles to enhance trust. Another study in the area of human-agent interaction found that human and agents' value similarity plays a significant role in increasing trust. In other words, people base their trust judgments on whether they feel that the system shares similar goals, thoughts, values, and opinions (Mehrotra et al., 2021b). Finally, to objectively evaluate the factors of trustworthiness of an AI system, a system called Cortex Certifai was developed to assess aspects of robustness, fairness, and interpretability of pre-trained AI models without requiring access to its internal model parameters (Henderson et al., 2020). Cortex Certifai generates various reports along these axes and only requires query access to the model and an "evaluation" dataset. Using these reports, stakeholders can understand, monitor, and build trust in their AI systems.

Discussion

Despite all the progress in AI, we are in an era of AI similar to when James Watt had to develop the concept of horsepower to help market his steam engine and convince people to buy it. AI is promising a new revolution in technology and has provided excellent results. Nonetheless, AI is complex, and its complexity is an integral part of it. Thus, as James Watt developed the concept of horsepower to gain people's trust in his steam engine, the AI community seeks to adapt trust-building concepts in AI, such as explainability, interpretability, and transparency to gain people's trust. Although all of these concepts serve to gain people's trust in AI, it is essential to consider the differences between them and the means to improve each.

Interaction of technical and non-technical factors of trust and trustworthiness in AI. Trust is an essential component in accepting and adopting AI technology in different domains. Although the general definition of trust between humans can be used to define trust in AI, there are unique factors that define trust in AI as a challenging problem. Humans need to ensure that their desires, needs, and rights are fulfilled by AI; these expectations could be related to AI's performance, reliability, and explainability. It is important to consider the differences between trust and trustworthiness and the means to improve each. While trustworthiness mostly refers to the ability of the AI system and targets technical factors, trust could be triggered by other non-technical factors such as reputation or documentation. Trust in the domain of AI can be defined in the interaction between human and AI, AI and human, and AI and AI, each of which has some unique requirements beyond the common basic factors of trust. Trust is a critical determinant of the successful adoption of AI technology. A large reason for the lack of adoption of AI models in different domains is the fact that the users are risk-averse and do not implicitly trust AI models. For users to depend on Google Assistant for weather forecasts, they must have confidence in the information provided. A lack of trust has greatly restricted the use of AI in areas like healthcare, self-driving cars, finance, education, personal assistants, chatbots, etc.

Understanding influential factors of trust is important, but there is an unmet need to understand the relationship between these factors in each domain. In addition to well-known parameters such as performance, explainability, transparency, compliance with certain regulations and standards, and ethical concerns, several other challenges, such as bias, discrimination, and privacy, need to be addressed to enhance trust and technology adoption further. The most important prerequisite is to identify the factors, their relationship, and how they interact to build trust and make an AI system trustworthy. Doing so requires well-designed experiments and comprehensive models of trust that consider quantitative and qualitative aspects of modeling trust in various domains.

Non-interchangeability of interpretability, explainability, and transparency, and their classification. While interpretability sheds light on the relationship between the cause (input) and the effect (output), explainability goes a step further and explains the inside of the AI system and its inner workings. Finally, transparency ensures that these explanations are transparent and clear. Unfortunately, sometimes researchers use these concepts interchangeably. One of the main reasons is the lack of a consensus definition for these concepts in AI. Moreover, it should be noted that different stakeholders need different levels of information. Therefore, the need to classify the degree of transparency, explainability and interpretability for different categories of stakeholders is inevitable. For example, researchers have shown through a behavioral experiment that giving excessive transparency will confuse the user and negatively impact trust (Schmidt et al., 2020a; 2020b).

Trust as a two-way street. Another downside of offering too much transparency, which deserves further exploration, is that it can allow a malicious user to exploit the system. In other words, trust operates in both directions; users need to trust the AI, but the AI also needs to trust the user. Lastly, a famous quote says, "trust is like a piece of paper. Once it is crumpled, it can never be perfect again." Given the presence of giants in the AI industry, such as Google and Facebook, any trust-destructive decisions by big AI companies undermine public trust in artificial intelligence regardless of all academic efforts on XAI. One of the concerns people have when using AI-based solutions is the reliability and safety of AI products. As a result, in addition to academic efforts, the need to establish an institution composed of neutral AI experts without any political, regional, and surveilling biases that oversee the decisions of AI companies and evaluate their products from the perspective of safety and reliability is suggested.

Distinction between empathy in human's trust in AI and empathy in AI's trust in human agents. When it comes to the relation between empathy and trust, a distinction should be made between the role of empathy in people's trust in AI systems and its role in their trust in other persons in computer mediated or online exchanges and communications. Most of the works

reviewed above are focused on the former, while some, including (Feng et al., 2004b), are focused on the latter.

Tradeoff between empathy and privacy. When empathy is most effective in trust-building, it might involve violations of privacy. This raises issues of the tradeoff between the empathy-trust link and privacy. For example, for patients to receive empathic AI-based treatments, they might have to share certain private data. A recent example is Amazon's Alexa's ability to mimic any person's voice,[1] which might enhance empathy and trust. However, such encroachments on privacy might, in turn, undermine trust. In future work, it might be studied how privacy-breaching empathy might be designed in AI systems (for medical, commercial, educational, and other purposes) so that a circle of mistrust does not ensue.

The subjectivity of trust in AI vs. the objectivity of reliable AI. Trust is a subjective or psychological phenomenon (it is a matter of one's confidence, say, in an AI system), in contrast to reliability, which is an objective probabilistic phenomenon (a matter of whether the system discharges its function properly). This implies that a company might do things (such as creating enjoyment and fun or other presentations), which can attract people's trust, without it being reliable enough. This would result in undue trust or overtrust in an AI system, disposing the user to act carelessly with regard to their private information (Kok and Soh, 2020). On the other hand, the subjective character of trust means that a system's reliability does not suffice to attract people's trust and convince them to share their private information. Developers of the system need to add features to gain the required degree of trust.

One might suggest that fairness has a role in enhancing a system's reliability or trustworthiness (as an objective phenomenon), which is a necessary requirement of trust as a psychological state. A system with (almost) no bias is more reliable than a biased system that fails to do justice to all groups of users. Aside from the positive relation between objective fairness and trust, different requirements might be in place for perceived fairness, which is also positively related to trust, including transparency of the data fairly fed into the system or the system's explainability. So, one might explore transparency, explainability, and the like as bridges between objective fairness and perceived fairness.

AI privacy and human agent privacy. There are factors that mediate between trust and privacy, such as an explanation of the workings of an AI system or an online provider or transparency, which can engender trust by ensuring that people's privacy is protected. This means that the link between trust and privacy might involve other factors as well, which need to be explored. Moreover, an AI product has private features, the disclosure of which might help gain the user's trust, such as its lineage and provenance. So privacy is a two-way street: both the producer and the user should exchange private data to make the relation work for both.

The developmental problem of "right to explanation". It is not enough to guarantee users a "right to explanation" (Cakir, 2020). Likewise, a right to explanation might create a perverse incentive according to which AI developers are encouraged to produce suboptimal models that are easier to explain in order to avoid the investment necessary to produce an optimal AI system, which is explicable (Doshi-Velez et al., 2016). Another factor leading to this potential perverse incentive is liability concerns on the part of developers (Doshi-Velez et al., 2016). If, as some have suggested, "trustworthiness is… a kind of reliability," then we can distinguish trust in AI from trust in the institutional system that AI emerges from (McLeod, 2020). While these are separate issues and establishing and maintaining the trustworthiness of each requires different kinds of solutions, trust in the latter will increase trust in the former. Likewise, we can further distinguish trust in the larger institutional systems according to whether the concern is about whether that system will reliably produce AI systems that (1) provide accurate recommendations, (2) provide equitable outcomes, (3) will use data responsibly, and/or (4) will be held accountable for failures of trust.

Development of direct AI accountability. This gestures at a central problem for accountability in AI. For human agents, trust in decision-making and the explanations for our decisions go hand in hand with accountability, but for artificial intelligence systems, decision-making and (potentially) explanation of those decisions rest with the AI, but accountability rests with the developers. For humans, we may trust other humans because we deem their motivations and intentions reliable. At least a part of this is a tendency to act so as to avoid punishment. A straightforward way to avoid punishment is to avoid punishable behavior. This leads to humans acting cautiously when trusted. It is not clear what accountability for an AI looks like. Yet, without a vision of what it might mean to hold an artificial intelligence system accountable, we have one less tool for establishing the reliability of behavior necessary for trust. In this way, accountability will rest with punishable developers until a theory of direct AI accountability is developed. This will, in turn, engender a perverse incentive for AI developers to avoid liability. Being predictably accurate is often insufficient to establish or warrant trust in humans. Attributions of trustworthiness often require a deeper concern for the basis of this reliability of behavior. This allows us to distinguish a lucky run of correct responses from one brought about because of some reliable mechanism for arriving at correct responses. Since machine learning is often thought of as a "black box", we may be left only with incentives like punishment avoidance as a potential mechanism for establishing trustworthiness. This leaves a considerable gap in the research when it comes to articulating not just how we might hold researchers and developers accountable for their use and design of AI, but whether it might be possible to hold AI directly accountable (and what this scheme might look like). A robust legal framework will require aligning explanation and accountability at the agential level.

Challenges of measuring trust and trustworthiness in AI. There are many challenges for defining trust and related metrics, including the dynamics of AI, both in terms of moment-by-moment developments and in terms of AI dependence on culture. There are also challenges in measuring trust. Not all of these challenges can be completely solved, even with the use of questionnaires, surveys, and protocols. If we look at the issue of human selection from a philosophical point of view, we see that the right choice for human beings is also challenging and has different dimensions. It is a person's choices that set him/her apart from the others. Now, how can we define the right choice for implementation in an AI to lead to trust? How would different dimensions (e.g., trust in AI) be concerned without having contradictions in their principles? Ethically, there are similar challenges to evaluating trust. First, without considering different cultures, one must explore what could be the universal ethical principles that can convince everyone to trust in AI. Next, how can these ethical principles be more in tune with a nation's culture in order to achieve greater value in trust assessment? Many

cases, by empirical or experimental methods only through trial and error, can greatly acquaint us with principles. Psychophysiological methods that have been used to evaluate trust in human-to-human relationships can also be helpful. Ultimately, all of these approaches should be able to work together with the help of theories to give the correct measures for trust assessment. Evaluating these principles should be used at three levels of organization, group, and individual to give a score. These scores can depend on the chosen approach. They include the efficiency of AI and its effectiveness in the task, user understanding, proper interaction between humans and AI, control, and data protection. There are other scales to measure, including trust in output and reliance on AI advice, which are also related to efficiency and predictability. Team and individual performance scores, team awareness, and metrics related to this process can also reveal differences between human-human trust and human-AI trust. We must not forget to consider the weak metrics of the AI system (such as vulnerabilities, errors, and risk assessment) along with the other mentioned metrics. It seems that the ethical framework is the most crucial and serious one (among the various frameworks shown in Fig. 4) in the codification of principles that can be used for guidelines and protocols by evaluating trust scores. Of course, many frameworks that have been employed in scientific texts are not able to satisfy us that we can work in a singular framework. In fact, it can be imagined that the illustrated frameworks are like nodes of a regular graph that are all interconnected. If the connection of one node to the other is disconnected, certainly not all aspects of the study will be considered. We know that not all of these can be included in an article with several authors, but they can be defined as a national or international project for different teams or organizations to find metrics and measure trust in AI. In this project, the evaluated parameters should be categorized in each framework from the most influential metric to the least important. In this way, a universal, comprehensive reference for frameworks and methods of measuring trust in AI via related metrics is designed for all AI manufacturers and their users. This reference includes all principles, protocols, roadmaps, and guidelines for producing and using AI and also trust them.

Trust equity problem in AI. In sum, there are a variety of kinds of concerns about AI that result in distrust. While concerns about AGIs have garnered significant academic and popular attention, so-called "weak AI" is not free from concern. While there is a sizable and growing literature on the reasons contributing to distrust, and on what kinds of explanations count as transparent in a way that encourages trust, there are still many ethical issues raised by considerations of trust in artificial intelligence. First, further research is needed on managing distrust in AI such that automation occurs in equitable ways. Without significant planning and foresight, the adoption of AI systems as alternatives to human-centered resources runs the risk of disproportionately affecting human competitors to AI from marginalized groups. For instance, if patients trust artificial intelligence programs more than female doctors but not more than male doctors, then the widespread introduction of artificial intelligence could exacerbate professional inequalities in healthcare between men and women. This means that trust in artificial intelligence systems might ultimately determine (in whole or in part) whether automation occurs primarily in industries dominated by otherwise marginalized groups, or (within industries) primarily as a replacement for jobs previously held by marginalized people. Thus, ensuring a basic level of trust necessary for the adoption of the technology may not be ethically adequate. Equitable adoption of artificial intelligence entails establishing a robust public sense of trust beyond a minimal threshold. Inversely, further research is needed to determine if the negative impacts of distrust are distributed equitably. If overreliance on artificial intelligence recommendations is domain-specific (such that users incorrectly assume that the AI is correct in its recommendation at different levels in different applications or domains), then the externalities associated with this misplaced trust might be distributed inequitably among the stakeholders in that decision process. In this way, concerns like those raised about bias propagation in criminal justice applications of AI might be mediated by judges' and lawyers' willingness to grant unearned trust in these systems. This issue, of trust equity in artificial intelligence (which concerns the relationship between trust in AI and equity in AI), demands significant further attention.

The previous discussion highlights the necessity for two significant recommendations for future research. First, researchers should develop and adopt an artificial intelligence trust equity framework. Such a framework would further identify the ways in which trust in artificial intelligence relative to human counterparts is distributed along the lines of demographic data about those human counterparts. This framework would also allow for targeted interventions to appropriately increase or decrease distrust in AI so as to ensure that the effects of artificial intelligence adoption are equitable with regard to economic impact and the concern for human dignity that are wrapped up in automation within the workplace. What a successful targeted intervention looks like is likely to be domain-dependent and specific to the particular trust inequities it is designed to target. Second, a complete trust equity framework requires further clarification of the conditions for trustworthiness and for inappropriate trust. This sets up a feedback loop in which solving the ethical issues which arise over equity in AI requires research on trust in AI and solving the issue of trust in AI adequately requires research on trust inequity. This suggests the utility of adopting an intersectional approach to analyzing these problems (see Fig. 7).

Impossibility of interpersonal trust in AI systems. The widely accepted assumption is that AI systems cannot have intentions; that is, they cannot intend their functions to be directed at certain goals. On the other hand, a main constituent of human trust is

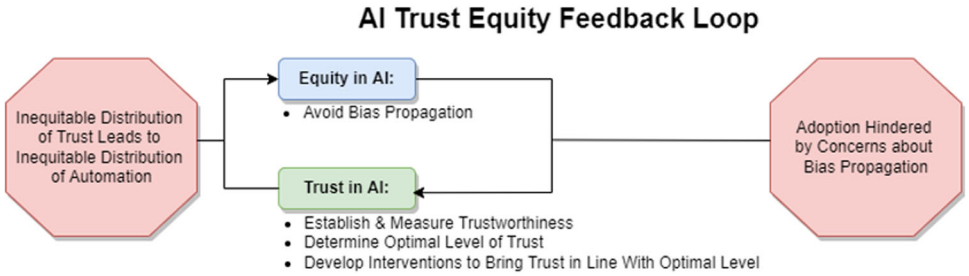

Fig. 7 AI trust equity feedback loop.

benevolence or honesty, in the sense that we trust in a human agent only when they exhibit honesty and good intentions. For this reason, it can arguably be said that it is not possible to have interpersonal trust in AI systems, since they lack intentions, which is why they cannot exhibit honesty and benevolence, which are necessary components of interpersonal trust. In human interactions, honesty or benevolence provides assurances for a trustful relationship, and the absence of this component in the case of human–AI relations makes it more difficult to create trust.

Concluding remarks and future directions

Consideration of trust in AI is one of the requirements of developing technologies in the fields of theorizing about AI and designing robots, human-AI interaction, and training their designers and users. In papers under our review, we were able to gain a general grasp of factors that would be employed as a metric to work on trust in AI. There are also some basic challenges that must be addressed in future research. To create AI algorithms and products or related technology, in the initial step, we must take the necessary precautions about the care of human and their satisfaction. Moreover, we must be very careful in formulating laws and standardizing AI and related technologies in design and exploitation for all users. These basic principles should be followed by determining the appropriate parameters for product quality remotely or by communication with the user. The implementation of these universal principles is possible only in a pervasive and comprehensive system that can be seen and tracked at any time all over the world. This system must be able to control the growing algorithms and production of technologies, as well as the implementation of principles in their codifications. One of these principles should be assigned to trust definition. This is a challenge that we would not be able to achieve just by considering one or more of the above-mentioned metrics.

The next step is to consider the dynamics of AI and technology, which may sometimes conflict with principles written for earlier developments or pre-defined metrics. This may not be a dangerous product in terms of human logic. Therefore, in such cases, the pervasive and comprehensive system must notify all lawmakers of the principles change or modify the previous principles for the new ones. Another example of dynamism is the dependence of AI and technology on different cultures. As we know, different cultures have different protocols, standards, and laws. What one culture deems right may be interpreted as obscenity for another. Therefore, the produced AI should have flexibility within the same boundaries as planned. This flexibility can also be achieved through statistics extracted from questionnaires, interviews, and surveys. Since one of these principles would be related to the concept of trust in AI and its parameterization, we must inevitably work with metrics of trust. Different approaches are employed to give various metrics for trust in AI. A preliminary metric that is necessary to trust in AI is reliability manifested in outputs and proper performance. Furthermore, depending on the purpose of the system, transparency of the system implementation would be in the higher order of significance. Transparency includes explainability, interpretability, and accountability. As mentioned, explainability and interpretability take a higher order of importance than accuracy. The metric of safety consisting of fairness, as well as the metric of security consisting of respect for privacy (data protection), are the other factors for trust in AI. The other influential metrics are the provenance (in some texts referred to as lineage) and automation of AI. Among these metrics, reducing and/or removing vulnerabilities and errors are very crucial and must be considered in research. By developing the AI, there would be factors characterized as metrics of trust in the future.

Building trust in AI requires understanding AI-related, human-related, and context-related factors that affect trust in a certain domain. It must be noted that some factors are application-dependent and should be evaluated in the context of the problem at hand. Transparency, explainability, and performance of the AI are amongst the most important technical AI-related factors that play critical roles in building trust in most application domains. These factors mainly increase the trustworthiness of the AI system. However, for the AI system to be trusted by the users, the AI's trustworthiness must be truly perceived by them. This requires certain cues to be provided to the users, which could be done through proper documentation. Other axiological factors for building trust, especially human-related ones, could be engineered to enhance trust without the need to improve the trustworthiness of AI.

An important need to ensure calibrated trust and avoid over or under-trust is to design standards and regulations that could be overseen by trustable agencies such as the government. This approach could increase trust even among those with less technical knowledge. However, little research has been done into building trust in the growing context of AI–AI interaction. There is an unmet need to design models for calibrating trust in AI–AI interaction as this type of interaction has unique and different requirements compared to human-AI interaction. In the case of AI–AI interaction, parameters such as transparency and explainability have no impact on building trust, while other

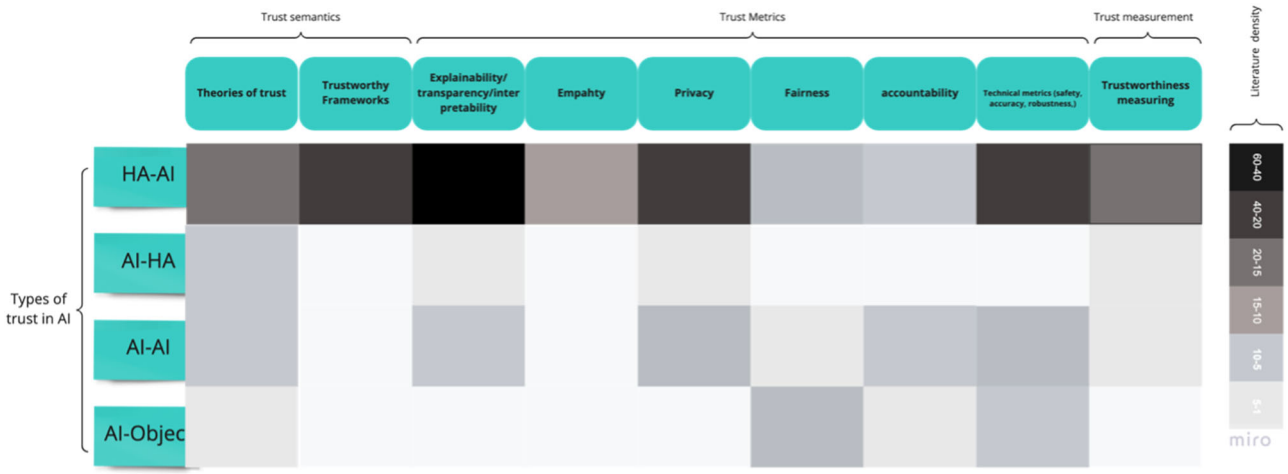

Fig. 8 Heatmap of the current work distribution on trust semantics, metrics, and measurement.

aspects such as security and reliability become important. In addition, from a technical perspective, there is a need to build robust models against adversarial attacks. Trust models also need to be able to determine malicious resources and calibrate their trust dynamically when interacting with multiple sources.

Finally, after the qualitative literature review, based on the number of reviewed papers and quantitative analysis, we determined that different research eras have not received equal attention. Figure 8 shows what has been done regarding trust-related research in AI, in its four major classes. There are some areas that have received very little or no attention in the literature and may be fertile areas for future research. Some other areas might be open questions for long term.

Data availability

All the necessary data for this study are included, and there are not any supplementary datasets.

Received: 2 April 2024; Accepted: 30 October 2024;
Published online: 18 November 2024

Note

1 Reuters, "Amazon has a plan to make Alexa mimic anyone's voice." URL: https://www.reuters.com/technology/amazon-has-plan-make-alexa-mimic-anyones-voice-2022-06-22/.

References

Abbass HA (2019a) Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. Cogn Comput 11(2):159–171

Abbass HA (2019b) Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. Cogn Comput 11(2):159–171. https://doi.org/10.1007/s12559-018-9619-0

Abbass HA (2019c) Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. Cogn Comput 11(2):159–171. https://doi.org/10.1007/s12559-018-9619-0

Abramoff MD (2021) Autonomous artificial intelligence safety and trust. In: Grzybowski A (ed.). Artificial intelligence in ophthalmology. Springer International Publishing. pp. 55–67

Adam M, Wessel M, Benlian A (2021a) AI-based chatbots in customer service and their effects on user compliance. Electron Mark 31(2):427–445

Adam M, Wessel M, Benlian A (2021b) AI-based chatbots in customer service and their effects on user compliance. Electron Mark 31(2):427–445

Afroogh S (2022) A probabilistic theory of trust concerning artificial intelligence: can intelligent robots trust humans? AI and Ethics. https://doi.org/10.1007/s43681-022-00174-4

Agarwal R, Melnick L, Frosst N, Zhang X, Lengerich B, Caruana R, Hinton GE (2021) Neural additive models: interpretable machine learning with neural nets. Adv Neural Inf Process Syst 34. https://arxiv.org/abs/2004.13912

Ahmed AS, Aura T (2018) Turning trust around: smart contract-assisted public key infrastructure. 2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), 104–111. IEEE

Ahn D, Almaatouq A, Gulabani M, Hosanagar K (2021) Will we trust what we don't understand? Impact of model interpretability and outcome feedback on trust in AI. https://doi.org/10.48550/arXiv.2111.08222

AI NIST (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0) https://doi.org/10.6028/NIST.AI.100-1

Ai Q, Narayanan RL (2021) Model-agnostic vs. model-intrinsic interpretability for explainable product search. Proceedings of the 30th ACM International Conference on Information & Knowledge Management. CIKM '21, November 1–5, 2021, Virtual Event, Australia, pp 5–15 https://dl.acm.org/doi/10.1145/3459637.3482276

Ajenaghughrure IB, Sousa SC da C, Lamas D (2020) Risk and trust in artificial intelligence technologies: a case study of autonomous vehicles. 13th International Conference on Human System Interaction. IEEE, pp. 118–123

Ajenaghughrure IB, da Costa Sousa SC, Lamas D (2020a) Risk and trust in artificial intelligence technologies: a case study of Autonomous Vehicles. 2020 13th International Conference on Human System Interaction (HSI). IEEE, pp 118–123. https://ieeexplore.ieee.org/document/9142686

Ajenaghughrure IB, da Costa Sousa SC, Lamas D (2020b) Risk and trust in artificial intelligence technologies: a case study of Autonomous Vehicles. 2020 13th International Conference on Human System Interaction (HSI). IEEE, pp. 118–123

Ajenaghughrure IB, Sousa SC, Kosunen IJ, Lamas D (2019) Predictive model to assess user trust: a psycho-physiological approach. The 10th Indian Conference, pp 1–10. https://doi.org/10.1145/3364183.3364195

Akbari A, Jafari R (2020) Personalizing activity recognition models through quantifying different types of uncertainty using wearable sensors. IEEE Trans Biomed Eng 67(9):2530–2541. https://doi.org/10.1109/TBME.2019.2963816

Akkara JD, Kuriakose A (2020) Commentary: Artificial intelligence for everything: can we trust it? Indian J Ophthalmol 68(7):1346–1347. https://doi.org/10.4103/ijo.IJO_216_20

al Khalil F, Butler T, O'Brien L, Ceci M (2017) Trust in smart contracts is a process, as well. International Conference on Financial Cryptography and Data Security. 510–519

Alam L (2020) Investigating the impact of explanation on repairing trust in ai diagnostic systems for re-diagnosis. Michigan Tech Digital Commons

Alam L, Mueller S (2021) Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. BMC Med Inform Decis Mak 21(1):1–15. https://doi.org/10.1186/s12911-021-01542-6

Alambeigi H, Smith A, Wei R, McDonald A, Arachie C, Huang B (2021) A novel approach to social media guideline design and its application to automated vehicle events. Proc Hum Factors Ergonom Soc Annu Meet 65(1):1510–1514

Albizri A, Appelbaum D (2021) Trust but verify: the oracle paradox of blockchain smart contracts. J Inf Syst 35(2):1–16

Aljably R, Tian Y, Al-Rodhaan M (2020) Preserving privacy in multimedia social networks using machine learning anomaly detection. Secur Commun Netw 2020:1–14

Andrulis J, Meyer O, Schott G, Weinbach S, Gruhn V (2020) Domain-level explainability—a challenge for creating trust in superhuman AI strategies. http://arxiv.org/abs/2011.06665

Araujo T, Helberger N, Kruikemeier S, de Vreese CH (2020a) In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Society 35(3):611–623. https://doi.org/10.1007/s00146-019-00931-w

Araujo T, Helberger N, Kruikemeier S, de Vreese CH (2020b) In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Society 35(3):611–623. https://doi.org/10.1007/s00146-019-00931-w

Arnold M, Bellamy RKE, Hind M, Houde S, Mehta S, Mojsilović A, Nair R, Ramamurthy KN, Olteanu A, Piorkowski D, Reimer D, Richards J, Tsay J, Varshney KR (2019a) FactSheets: increasing trust in AI services through supplier's declarations of conformity. IBM J Res Dev 63(4/5):6:1–6:13. https://doi.org/10.1147/JRD.2019.2942288

Arnold M, Bellamy RKE, Hind M, Houde S, Mehta S, Mojsilović A, Nair R, Ramamurthy KN, Olteanu A, Piorkowski D, Reimer D, Richards J, Tsay J, Varshney KR (2019b) FactSheets: increasing trust in AI services through supplier's declarations of conformity. IBM J Res Dev 63(4/5):6:1–6:13. https://doi.org/10.1147/JRD.2019.2942288

Arnold M, Piorkowski D, Reimer D, Richards J, Tsay J, Varshney KR, Bellamy RKE, Hind M, Houde S, Mehta S, Mojsilovic A, Nair R, Ramamurthy KN, Olteanu A (2019) FactSheets: increasing trust in AI services through supplier's declarations of conformity. IBM J Res Dev 63(4/5):6:1–6:13. https://doi.org/10.1147/JRD.2019.2942288

Asan O, Bayrak AE, Choudhury A (2020) Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 22(6):1–7. https://doi.org/10.2196/15154

Asan O, Bayrak AE, Choudhury A et al. (2020a) Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 22(6):e15154

Ashoori M, Weisz JD (2019) In AI we trust? Factors that influence trustworthiness of ai-infused decision-making processes. http://arxiv.org/abs/1912.02675

Ausloos J, Zaman B, Geerts D, Valcke P, Dewitte P (2020) Algorithmic transparency and accountability in practice. Interdisciplinariteit in Actie. https://www.researchgate.net/publication/339747504_Algorithmic_Transparency_and_Accountability_in_Practice_ATAP_A_Study_into_Automated_N

Baldauf M, Fröehlich P, Endl R (2020). Trust me, i'm a doctor-user perceptions of AI-driven apps for mobile health diagnosis. ACM International Conference Proceeding Series, 167–178. https://doi.org/10.1145/3428361.3428362

Banavar G (2016) What it will take for us to trust AI. Harvard Business Review

Barrué C (2021) A European Survey on AI and Ethics. AI4EU working group

Bartneck C, Lütge C, Wagner A, Welsh S (2021) Trust and fairness in AI systems. In: Bartneck C, Lütge C, Wagner A, Welsh S (eds.). An introduction to ethics in robotics and AI. Springer International Publishing. pp. 27–38

Baum SD (2017) A survey of artificial general intelligence projects for ethics, risk, and policy. https://www.emerald.com/insight/content/doi/10.1108/jeim-06-2020-0233/full/html

Beck R, Stenum Czepluch J, Lollike N, Malone S (2016) Blockchain–the gateway to trust-free cryptographic transactions. Conference: Proceedings of the Twenty-Fourth European Conference on Information Systems (ECIS)

Bedué P, Fritzsche A (2021) Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. J Enterprise Inf Management. https://doi.org/10.1108/JEIM-06-2020-0233
Beer JM, Fisk AD, Rogers WA (2014a) Toward a framework for levels of robot autonomy in human-robot interaction. J Hum Robot Interact 3(2):74–99. https://doi.org/10.5898/JHRI.3.2.Beer
Beer JM, Fisk AD, Rogers WA (2014b) Toward a framework for levels of robot autonomy in human-robot interaction. J Hum Robot Interact 3(2):74–99. https://doi.org/10.5898/jhri.3.2.beer
Benda NC, Reale C, Ancker JS, Ribeiro J, Walsh CG, Lovett Novak L (2021) Purpose, PRocess, Performance: Designing for Appropriate Trust of AI in healthcare position paper
Bhatti S, Demir M, Cooke NJ, Johnson CJ (2021) Assessing communication and trust in an ai teammate in a dynamic task environment. 2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS), 1–6. https://doi.org/10.1109/ICHMS53169.2021.9582626
Binns R (2018) Algorithmic accountability and public reason. Philos Technol 31(4):543–556. https://doi.org/10.1007/s13347-017-0263-5
Blacklaws C (2018) Algorithms: Transparency and accountability. Philos Trans R Soc A Math Phys Eng Sci 376(2128). https://doi.org/10.1098/rsta.2017.0351
Bock DE, Wolter JS, Ferrell OC (2020) Artificial intelligence: disrupting what we know about services. J Serv Mark 34(3):317–334. https://doi.org/10.1108/JSM-01-2019-0047
Bore NK, Kiran Raman R, Markus IM, Remy SL, Bent O, Hind M, Pissadaki EK, Srivastava B, Vaculin R, Varshney KR, Weldemariam K (2018) Promoting distributed trust in machine learning and computational simulation via a blockchain network. https://arxiv.org/abs/1810.11126
Bostrom N (2014) Superintelligence: paths, dangers, strategies (First). Oxford University Press
Bovens M, Goodin RE, Schillemans T, Bovens M, Schillemans T (2014) Meaningful accountability. In: The Oxford handbook of public accountability. Oxford University Press
Branley-Bell D, Whitworth R, Coventry L (2020) User trust and understanding of explainable AI: exploring algorithm visualisations and user biases. In: Kurosu M (ed.). Springer International Publishing. pp. 382–399
Brave S, Nass C, Hutchinson K (2005) Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. Int J Hum Comput Stud 62:161–178. https://doi.org/10.1016/j.ijhcs.2004.11.002
Brown N, Sandholm T (2018) Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. Science 359(6374):418–424
Bruzzese T, Gao I, Dietz G, Ding C, Romanos A (2020, April 25) Effect of confidence indicators on trust in AI-generated profiles. Conference on Human Factors in Computing Systems—Proceedings. https://doi.org/10.1145/3334480.3382842
Buçinca Z, Malaya MB, Gajos KZ (2021) To trust or to think. Proc ACM Hum Comput Interact 5(CSCW1):1–21. https://doi.org/10.1145/3449287
Bughin J, Hazan E, Lund S, Dahlström P, Wiesinger A, Subramaniam A (2018) Skill shift: Automation and the future of the workforce. McKinsey Glob Inst 1:3–84
Cabitza F, Campagner A, Datteri E (2021) To err is (only) human. Reflections on how to move from accuracy to trust for medical AI. In: Ceci F, Prencipe A, Spagnoletti P (eds.). Springer International Publishing. pp. 36–49
Cakir C (2020) Fairness, accountability and transparency—trust in AI and machine learning. In: Bhatti SA, Christi S, Datoo A, Indjic D (eds.). The LEGALTECH BOok: the legal technology handbook for investors, entrepreneurs and FINTECH Visionaries (First). Wiley. pp. 35–38
Carrozzo G (2020) AI-driven zero-touch operations, security and trust in multi-operator 5G networks: a conceptual architecture. European Conference on Networks and Communications (EuCNC). IEEE
Carta SM, Consoli S, Piras L, Podda AS, Recupero DR (2021) Explainable machine learning exploiting news and domain-specific lexicon for stock market forecasting. IEEE Access 9:30193–30205
Carvalho DV, Pereira EM, Cardoso JS (2019) Machine learning interpretability: a survey on methods and metrics. Electronics 8(8):832
Caspers J (2021) Translation of predictive modeling and AI into clinics: a question of trust. Eur Radiol 31(7):4947–4948
Chalmers DJ (2010) The singularity: a philosophical analysis. J Conscious Stud 17:7–65
Chandra S (2010) Evaluating the role of trust in consumer adoption of mobile payment systems: an empirical analysis. Commun Assoc Inf Syst 27(29):561–588
Chandra S, Srivastava SC, Theng Y-L (2010) Evaluating the role of trust in consumer adoption of mobile payment systems: an empirical analysis. Commun Assoc Inf Syst 27. https://doi.org/10.17705/1cais.02729
Chen YNK, Wen CHR (2021) Impacts of attitudes toward government and corporations on public trust in artificial intelligence. Commun Stud 72(1):115–131. https://doi.org/10.1080/10510974.2020.1807380
Cho J-H, Chan K, Adali S (2015) A survey on trust modeling. ACM Comput Surv 48(2):1–40. https://doi.org/10.1145/2815595
Cho J-H, Hurley PM, Xu S (2016a) Metrics and measurement of trustworthy systems. MILCOM 2016—2016 IEEE Military Communications Conference, 1237–1242. https://doi.org/10.1109/MILCOM.2016.7795500
Cho J-H, Hurley PM, Xu S (2016b) Metrics and measurement of trustworthy systems. MILCOM 2016—2016 IEEE Military Communications Conference, 1237–1242. https://doi.org/10.1109/MILCOM.2016.7795500
Cho J-H, Xu S, Hurley PM, Mackay M, Benjamin T, Beaumont M (2019a) STRAM: measuring the trustworthiness of computer-based systems. ACM Comput Surv 51(6):128:1–128:47. https://doi.org/10.1145/3277666
Cho J-H, Xu S, Hurley PM, Mackay M, Benjamin T, Beaumont M (2019b) STRAM: measuring the trustworthiness of computer-based systems. ACM Comput Surv 51(6):128:1–128:47. https://doi.org/10.1145/3277666
Cihon P, Kleinaltenkamp MJ, Schuett J, Baum SD (2021a) AI certification: advancing ethical practice by reducing information asymmetries. IEEE Trans Technol Soc 2(4):200–209. https://doi.org/10.1109/TTS.2021.3077595
Cihon P, Kleinaltenkamp MJ, Schuett J, Baum SD (2021b) AI certification: advancing ethical practice by reducing information asymmetries. IEEE Trans Technol Soc 2(4):200–209. https://doi.org/10.1109/TTS.2021.3077595
Culnan MJ, Armstrong PK (1999) Information privacy concerns, procedural fairness, and impersonal trust: an empirical investigation. Organ Sc 10(1):104–115
Dakkak A, Li C, de Gonzalo SG, Xiong J, Hwu WM (2019) TrIMS: Transparent and isolated model sharing for low latency deep learning inference in function-as-a-service. IEEE International Conference on Cloud Computing, CLOUD, 2019-July, pp. 372–382. IEEE
DAngelo G, Rampone S, Palmieri F (2015) An artificial intelligence-based trust model for pervasive computing. 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), pp. 701–706
Das A, Rad P (2020) Opportunities and challenges in explainable artificial intelligence (XAI): a survey. http://arxiv.org/abs/2006.11371
DeCamp M, Tilburt JC (2019) Why we cannot trust artificial intelligence in medicine. Lancet Digit Health 1(8):e390. https://doi.org/10.1016/S2589-7500(19)30197-9
Devitt SK, Horne R, Assaad Z, Broad E, Kurniawati H, Cardier B, Scott A, Lazar S, Gould M, Adamson C, Karl C, Schrever F, Keay S, Tranter K, Shellshear E, Hunter D, Brady M, Putland T (2021) Trust and Safety. http://arxiv.org/abs/2104.06512
Diab DL, Pui S-Y, Yankelevich M, Highhouse S (2011) Lay perceptions of selection decision aids in US and non-US samples. Int J Selection Assess 19(2):209–216
Dignum V (2018) Ethics in artificial intelligence: introduction to the special issue. Ethics Inf Technol 20(1):1–3. https://doi.org/10.1007/s10676-018-9450-z
DiLuoffo V, Michalson WR (2021) A survey on trust metrics for autonomous robotic systems. http://arxiv.org/abs/2106.15015
Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. Preprint at arXiv
Doshi-Velez F, Kortz M, Budish R, Bavitz C, Gershman S, O'Brien D, Scott K, Shieber S, Waldo J, Weinberger D, Weller A, Wood A (2016) Accountability of AI under the law: the role of explanation. http://arxiv.org/abs/1606.06565
Dosilovic FK, Brcic M, Hlupic N (2018) Explainable artificial intelligence: a survey. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018—Proceedings. pp. 210–215. https://doi.org/10.23919/MIPRO.2018.8400040
Du M, Liu N, Hu X (2019) Techniques for interpretable machine learning. Commun ACM 63(1):68–77
Eigenstetter M (2020) Ensuring trust in and acceptance of digitalization and automation: contributions of human factors and ethics. International Conference on Human-Computer Interaction
European Commission (2020) White paper on artificial intelligence: a European approach to excellence and trust
Ezer N, Bruni S, Cai Y, Hepenstal SJ, Miller CA, Schmorrow DD (2019) Trust engineering for human-AI teams. Proc Hum Factors Ergonom Soc Annu Meet 63(1):322–326. https://doi.org/10.1177/1071181319631264
Fan M, Zou F, He Y, Xuan J (2021a) Research on users' trust of Chatbots driven by AI: an empirical analysis based on system factors and user characteristics. 2021 IEEE Int Conf Consum Electron Comput Eng ICCECE 2021:55–58. https://doi.org/10.1109/ICCECE51280.2021.9342098
Fan M, Zou F, He Y, Xuan J (2021b) Research on users' trust of Chatbots driven by AI: an empirical analysis based on system factors and user characteristics. 2021 IEEE Int Conf Consum Electron Comput Eng ICCECE 2021:55–58. https://doi.org/10.1109/ICCECE51280.2021.9342098
Feldman R, Aldana E, Stein K (2019) Artificial intelligence in the health care space: how we can trust what we cannot know. Stanford Law Policy Rev 30. https://repository.uclawsf.edu/cgi/viewcontent.cgi?article=2755&context=faculty_scholarship
Felzmann H, Villaronga EF, Lutz C, Tamò-Larrieux A (2019) Transparency you can trust: Transparency requirements for artificial intelligence between legal

norms and contextual concerns. Big Data Soc 6(1). https://doi.org/10.1177/2053951719860542
Feng J, Lazar J, Preece J (2004a) Empathy and online interpersonal trust: a fragile relationship. Behav Inf Technol 23(2):97–106. https://doi.org/10.1080/01449290310001659240
Feng J, Lazar J, Preece J (2004b) Empathy and online interpersonal trust: a fragile relationship. Behav Inf Technol 23(2):97–106. https://doi.org/10.1080/01449290310001659240
Ferrario A, Loi M (n.d.) The meaning of "Explainability Fosters Trust in AI" https://ssrn.com/abstract=3916396
Fooladi Mahani M, Jiang L, Wang Y (2020) A Bayesian trust inference model for human-multi-robot teams. Int J Soc Robot. https://doi.org/10.1007/s12369-020-00705-1
Gabriel I (2020) Artificial intelligence, values, and alignment. Minds Mach 30(3):411–437. https://doi.org/10.1007/s11023-020-09539-2
Galán JJ, Carrasco RA, LaTorre A (2022) Military applications of machine learning: a bibliometric perspective. Mathematics 10(9):1397
Gebhard P, Aylett R, Higashinaka R, Jokinen K, Tanaka H, Yoshino K (2021) Modeling trust and empathy for socially interactive robots. pp. 21–60. https://www.researchgate.net/publication/355177408_Modeling_Trust_and_Empathy_for_Socially_Interactive_Robots
Ghassemi M, Pushkarna M, Wexler J, Johnson J, Varghese P (2018) ClinicalVis: supporting clinical task-focused design evaluation. http://arxiv.org/abs/1810.05798
Gille F, Jobin A, Ienca M, Gille F, Jobin A (2020) What we talk about when we talk about trust: theory of trust for AI in healthcare. ETH Library. https://doi.org/10.3929/ethz-b-000430039
Gille F, Smith S, Mays N (2015) Why public trust in health care systems matters and deserves greater research attention. J Health Serv Res Policy 20(1):62–64
Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: an overview of interpretability of machine learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). pp. 80–89
Gite S, Khatavkar H, Kotecha K, Srivastava S, Maheshwari P, Pandey N (2021) Explainable stock prices prediction from financial news articles using sentiment analysis. PeerJ Comput Sci 7:e340
Glomsrud JA, Ødegårdstuen A, Clair ALS, Smogeli Ø (2019) Trustworthy versus explainable AI in autonomous vessels. Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC). pp. 37–47
Gretton C (2018) Trust and transparency in machine learning-based clinical decision support. In: Zhou J, Chen F (eds.). Human and machine learning: visible, explainable, trustworthy and transparent. pp. 279–292. Springer International Publishing
Guckert M, Gumpfer N, Hannig J, Keller T, Urquhart N (2021) A conceptual framework for establishing trust in real world intelligent systems. Cogn Syst Res 68:143–155. https://doi.org/10.1016/j.cogsys.2021.04.001
Guo J, Liu A, Ota K, Dong M, Deng X, Xiong NN (2022) ITCN: an intelligent trust collaboration network system in IoT. IEEE Trans Netw Sci Eng 9(1):203–218. https://doi.org/10.1109/TNSE.2021.3057881
Guo J, Ma J, Li X, Zhang J, Zhang T (2017) An attribute-based trust negotiation protocol for D2D communication in smart city balancing trust and privacy. J Inf Sci Eng 33(4):1007–1023. https://doi.org/10.6688/JISE.2017.33.4.10
Gupta K, Hajika R, Pai YS, Duenser A, Lochner M, Billinghurst M (2019) In AI we trust: investigating the relationship between biosignals, trust and cognitive load in VR. 1–10. https://doi.org/10.1145/3359996.3364276
Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Mind Mach 30(1):99–120. https://doi.org/10.1007/s11023-020-09517-8
Hale MT, Setter T, Fregene K (2019) Trust-driven privacy in human-robot interactions. 2019 American Control Conference (ACC). pp. 5234–5239. https://doi.org/10.23919/ACC.2019.8815004
Han Y, Perry JL (2020) Conceptual bases of employee accountability: a psychological approach. Perspect Public Manag Gov 3(4):288–304. https://doi.org/10.1093/ppmgov/gvz030
Hatherley JJ (2020) Limits of trust in medical AI. J Med Ethics 46(7):478–481. https://doi.org/10.1136/medethics-2019-105935
Hauer C (2021) Chapter 23—Should we trust robots? The ethical perspective. In: Nam CS, Lyons JB (eds.) Trust in human-robot interaction. pp. 531–551. Academic Press. https://www.sciencedirect.com/science/article/pii/B978012819472000023X
Hawley K (2014) Trust, distrust and commitment. In: Source: Noûs. vol. 48. Wiley, Issue 1
Henderson J, Sharma S, Gee A, Alexiev V, Draper S, Marin C, Hinojosa Y, Draper C, Perng M, Aguirre L, Li M, Rouhani S, Consul S, Michalski S, Prasad A, Chutani M, Kumar A, Alam S, Kandarpa P, … Ghosh, J (2020) Certifai: a toolkit for building trust in AI systems. https://www.cognitivescale.com/certifai/
Hernandez CS, Ayo S, Panagiotakopoulos D (2021) An explainable artificial intelligence (xAI) framework for improving trust in automated ATM tools. 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), 1–10. IEEE
Hind M, Mehta S, Mojsilovic A, Nair R, Ramamurthy KN, Olteanu A, Varshney KR (2018) Increasing trust in AI services through supplier's declarations of conformity. Preprint at arXiv 18, 2813–2869
Hobson, Z, Yesberg JA, Bradford B, Jackson J (2021) Artificial fairness? Trust in algorithmic police decision-making. J ExpCriminol 1–25. https://doi.org/10.1007/s11292-021-09484-9
Hoffman ML (2000) Empathy and moral development: implications for caring and justice. Cambridge University Press. https://www.cambridge.org/core/books/empathy-and-moral-development/0888510CFC9324935DCDFF7609E491FA
Hoffman R, Mueller S, Klein G, Litman J (2021) Measuring trust in the XAI context. PsyArXiv. https://psyarxiv.com/e3kv9/
Hoffman RR, Klein G, Mueller ST (2018) Explaining explanation for "Explainable AI.". Proc Hum Factors Ergonom Soc Annu Meet 62(1):197–201
Holzinger A, Biemann C, Pattichis CSKell DB (2017) What do we need to build explainable AI systems for the medical domain? https://arxiv.org/abs/1712.09923
Hong L, Jiaming T, Yan S (2009) Entropy-based trust management for data collection in wireless sensor networks. Proceedings—5th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2009. https://doi.org/10.1109/WICOM.2009.5302823
Hui CY, McKinstry B, Fulton O, Buchner M, Pinnock H (2021) Patients' and clinicians' perceived trust in internet-of-things systems to support asthma self-management: qualitative interview study. JMIR MHealth UHealth 9(7):e24127
Hurlburt G (2017a) How much to trust artificial intelligence? IT Professional 19(4):7–11. https://doi.org/10.1109/MITP.2017.3051326
Hurlburt G (2017b) How much to trust artificial intelligence? IT Professional 19(4):7–11
Ickes W (1993) Empathic accuracy. J Personal 61(4):587–610. https://doi.org/10.1111/j.1467-6494.1993.tb00783.x
Ingrams A, Kaufmann W, Jacobs D (2021) In AI we trust? Citizen perceptions of AI in government decision making. Policy Internet, 1–20. https://doi.org/10.1002/poi3.276
Ingrand F (2019) Recent trends in formal validation and verification of autonomous robots software. 2019 Third IEEE International Conference on Robotic Computing (IRC), 321–328. https://doi.org/10.1109/IRC.2019.00059
Itani S, Rossignol M, Lecron F, Fortemps P (2019) Towards interpretable machine learning models for diagnosis aid: a case study on attention deficit/hyperactivity disorder. PLoS One 14(4):e0215720
Jackson S, Panteli N (2021) A multi-level analysis of mistrust/trust formation in algorithmic grading. International Federation for Information Processing, 12896 LNCS, 737–743. https://doi.org/10.1007/978-3-030-85447-8_61
Jacobs M, He J, Pradier MF (2021, May 6) Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. Conference on Human Factors in Computing Systems—Proceedings. https://doi.org/10.1145/3411764.3445385
Jacovi A, Marasović A, Miller T, Goldberg Y (2021a) Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. FAccT 2021—Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 624–635. https://doi.org/10.1145/3442188.3445923
Jacovi A, Marasović A, Miller T, Goldberg Y (2021b) Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI. FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, 624–635. https://doi.org/10.1145/3442188.3445923
Jiang H, Kim B, Guan MY, Gupta M (2018) To Trust Or Not to Trust A Classifier. 32 Conference on Neural Information Processing Systems, 1–25. http://arxiv.org/abs/1805.11783
Jobin A, Ienca M, Vayena E (2019a) Artificial intelligence: the global landscape of ethics guidelines
Jobin A, Ienca M, Vayena E (2019b) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399. https://doi.org/10.1038/s42256-019-0088-2
Johnson J (2020) Artificial intelligence, drone swarming and escalation risks in future warfare. RUSI J 165(2):26–36. https://doi.org/10.1080/03071847.2020.1752026
Kamath U, Liu J (2021) Explainable artificial intelligence: an introduction to interpretable machine learning. Springer
Kaplan AD, Kessler TT, Christopher Brill J, Hancock PA (2021) Trust in artificial intelligence: meta-analytic findings. In: Human Factors: The Journal of the Human Factors and Ergonomics Society 65(2). https://doi.org/10.1177/00187208211013
Kaur D, Uslu S, Durresi A (2021) Requirements for trustworthy artificial intelligence—a review. In: Barolli L, Li KF, Enokido T, Takizawa M (eds) pp. 105–115. Springer International Publishing
Kellmeyer P, Mueller O, Feingold-Polak R, Levy-Tzedek S (2018) Social robots in rehabilitation: a question of trust. Sci Robot 3(21):eaat1587. https://doi.org/10.1126/scirobotics.aat1587

Kerasidou A (2020) Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. Bull World Health Organ 98(4):245–250. https://doi.org/10.2471/BLT.19.237198
Kerasidou A (2021a) Ethics of artificial intelligence in global health: explainability, algorithmic bias and trust. J Oral Biol Craniofacial Res 11(4):612–614. https://doi.org/10.1016/j.jobcr.2021.09.004
Kerasidou A (2021b) Ethics of artificial intelligence in global health: Explainability, algorithmic bias and trust. J Oral Biol Craniofacial Res 11(4):612–614. https://doi.org/10.1016/j.jobcr.2021.09.004
Kessler T, Stowers K, Brill, JC, Hancock PA (2017) Comparisons of human-human trust with other forms of human-technology trust. Proceedings of the Human Factors and Ergonomics Society, 2017-October, pp. 1303–1307. https://doi.org/10.1177/1541931213601808
Khalid HM, Shiung LW, Nooralishahi P, Rasool Z, Helander MG, Kiong LC, Ai-Vyrn C (2016) Exploring psycho-physiological correlates to trust: implications for human-robot-human interaction. Proceedings of the Human Factors and Ergonomics Society, pp. 696–700. https://doi.org/10.1177/1541931213601160
Khavas ZR, Ahmadzadeh SR, Robinette P (2020) Modeling trust in human-robot interaction: a survey. In: Wagner AR, Feil-Seifer D, Haring KS, Rossi S, Williams T, He H, Sam Ge S (eds). pp. 529–541. Springer International Publishing
Kim DKD, Kim S (2021a) What if you have a humanoid AI robot doctor?: An investigation of public trust in South Korea. J Commun Healthcare. https://doi.org/10.1080/17538068.2021.1994825
Kim DKD, Kim S (2021b) What if you have a humanoid AI robot doctor?: an investigation of public trust in South Korea. J Commun Healthcare, 1–10. https://doi.org/10.1080/17538068.2021.1994825
Kim J (2021) When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations. Psychol Market, Wiley. https://onlinelibrary.wiley.com/doi/abs/10.1002/mar.21498
Kim J, Giroux M, Lee JC (2021) When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations. Psychol Mark 38(7):1140–1155. https://doi.org/10.1002/mar.21498
Knowles B, Richards JT (2021) The sanction of authority: Promoting public trust in AI. FAccT 2021—Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 262–271. https://doi.org/10.1145/3442188.3445890
Kok BC, Soh H (2020) Trust in robots: challenges and opportunities. Curr Robot Rep. 1(4):297–309. https://doi.org/10.1007/s43154-020-00029-y
Kumar B, Singh AV, Agarwal P (2021) Trust in banking management system using firebase in Python using AI. 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2021. https://doi.org/10.1109/ICRITO51393.2021.9596273
Langer EJ, Blank A, Chanowitz B (1978) The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. J Personal Soc Psychol 36(6):635
Lazányi K (2019) Generation Z and Y–are they different, when it comes to trust in robots? 2019 IEEE 23rd International Conference on Intelligent Engineering Systems (INES), IEEE. 191–194
Lee JD, See KA (2004) Trust in technology: designing for appropriate reliance. Hum Factors 46(1):50–80
Lee MK, Rich K (2021a) Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. CHI '21: CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3411764.3445570
Lee MK, Rich K (2021b) Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. 1–14. https://doi.org/10.1145/3411764.3445570
Lee O-KD, Ayyagari R, Nasirian F, Ahmadian M (2021) Role of interaction quality and trust in use of AI-based voice-assistant systems. J Syst Inf Technol 23(2):154–170
Lee SS (2021a) Philosophical evaluation of the conceptualisation of trust in the NHS' Code of Conduct for artificial intelligence-driven technology. J Medical Ethics. https://doi.org/10.1136/medethics-2020-106905
Lee SS (2021b) Philosophical evaluation of the conceptualisation of trust in the NHS' Code of Conduct for artificial intelligence-driven technology. J Medical Ethics. https://doi.org/10.1136/medethics-2020-106905
Leonard PG (2018a) Social licence and digital trust in data-driven applications and AI: a problem statement and possible solutions. Cult Anthropol EJ. https://doi.org/10.1016/j.jmb.2018.05.044
Leonard PG (2018b) Social licence and digital trust in data-driven applications and AI: a problem statement and possible solutions. SSRN
Leta Jones M, Kaufman E, Edenberg E (2018) AI and the ethics of automating consent. IEEE Security Priv 16(3):64–72
Li C, Guo W, Sun SC, Al-Rubaye S, Tsourdos A (2020) Trustworthy deep learning in 6G-enabled mass autonomy: from concept to quality-of-trust key performance indicators. IEEE Vehicular Technol Mag 15(4):112–121. https://doi.org/10.1109/MVT.2020.3017181
Li J, Chen X, Hovy E, Jurafsky D (2016) Visualizing and understanding neural models in NLP. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016—Proceedings of the Conference, 681–691. https://doi.org/10.18653/v1/n16-1082
Li J, Wu J, Li J, Bashir AK, Piran MdJ, Anjum A (2021) Blockchain-based trust edge knowledge inference of multi-robot systems for collaborative tasks. IEEE Commun Mag 59(7):94–100. https://doi.org/10.1109/MCOM.001.2000419
Li J, Zhou Y, Yao J, Liu X (2021a) An empirical investigation of trust in AI in a Chinese petrochemical enterprise based on institutional theory. Sci Rep 11(1). https://doi.org/10.1038/s41598-021-92904-7
Li J, Zhou Y, Yao J, Liu X (2021b). An empirical investigation of trust in AI in a Chinese petrochemical enterprise based on institutional theory. Sci Rep 11(1). https://doi.org/10.1038/s41598-021-92904-7
Li J, Zhou Y, Yao J, Liu X (2021c) An empirical investigation of trust in AI in a Chinese petrochemical enterprise based on institutional theory. Sci Rep.11(1):13564. https://doi.org/10.1038/s41598-021-92904-7
Lipton ZC (2019) The mythos of model interpretability. Preprint at arXiv https://arxiv.org/abs/1606.03490
Lockey S, Gillespie N, Holm D, Someh IA (2021) A review of trust in artificial intelligence: challenges, vulnerabilities and future directions. https://hicss.hawaii.edu/
Lu Z, Qu G, Liu Z (2019) A survey on recent advances in vehicular network security, trust, and privacy. IEEE Trans Intell Trans Syst 20(2):760–776. https://doi.org/10.1109/TITS.2018.2818888
Lui A, Lamb GW (2018) Artificial intelligence and augmented intelligence collaboration: regaining trust and confidence in the financial sector. Inf Commun Technol Law 27(3):267–283
Lyons JB, Clark MA, Wagner AR, Schuelke MJ (2017) Certifiable trust in autonomous systems: making the intractable tangible. AI Mag 38(3):37–49. https://doi.org/10.1609/aimag.v38i3.2717
Madsen A, Reddy S, Chandar S (2021) Post-hoc Interpretability for Neural NLP: a survey. https://arxiv.org/abs/2108.04840
Mahbooba B, Timilsina M, Sahal R, Serrano M (2021) Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. Complexity, 2021. https://doi.org/10.1155/2021/6634811
Masis S (2021) Interpretable machine learning with Python: learn to build interpretable high-performance models with hands-on real-world examples. Packt Publishing Ltd
Maurtua I et al. (2017) Human–robot collaboration in industrial applications: safety, interaction and trust. Int J Adv Robot Syst 1–10
Mayer RC, Davis JH, David Schoorman F (1995) An integrative model of organizational trust (vol. 20, issue 3). https://www.jstor.org/stable/258792?seq=1&cid=pdf-
Mccarthy J, Hayes PJ (1969) Some philosophical problems from the standpoint of artificial intelligence. http://www-formal.stanford.edu/jmc/
McDonald AD, Alambeigi H, Engström J, Markkula G, Vogelpohl T, Dunne J, Yuma N (2019) Toward computational simulations of Behavior during Automated Driving Takeovers: a review of the empirical and modeling literatures. In: Human factors. vol. 61, issue 4. SAGE Publications Inc. pp. 642–688
McLeod C (2020) Trust. In: Stanford Encyclopedia of Philosophy. Metaphysics Research Lab
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv (CSUR) 54(6):1–35
Mehri V, Kurt Tutschku (2017) Flexible privacy and high trust in the next generation internet: The use case of a cloud-based marketplace for AI. SNCNW-Swedish National Computer Networking Workshop
Mehrotra S, Jonker CM, Tielman ML (2021a) More similar values, more trust?—the effect of value similarity on trust in human-agent interaction. 777–783. https://doi.org/10.1145/3461702.3462576
Mehrotra S, Jonker CM, Tielman ML (2021b) More SImilar Values, More Trust?—The effect of value similarity on trust in human-agent interaction. AIES 2021—Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 777–783. https://doi.org/10.1145/3461702.3462576
Mesbah N, Tauchert C, Olt CM, Buxmann P (2019) Promoting trust in AI-based expert systems. Promote AI-based Expert Systems. Twenty-fifth Americas Conference on Information Systems, Cancun
Meske C, Bunde E (2020a) Transparency and trust in human-AI-interaction: the role of model-agnostic explanations in computer vision-based decision support. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12217 LNCS, pp. 54–69. https://doi.org/10.1007/978-3-030-50334-5_4
Meske C, Bunde E (2020b) Transparency and trust in human-AI-interaction: the role of model-agnostic explanations in computer vision-based decision support. International Conference on Human-Computer Interaction, ACM, pp 54–69

Meyer JM (2021) Investigating the determinants and conditions of trust in AI by physicians. CIS 2021 TREOs. 27. https://aisel.aisnet.org/treos_icis2021/27
Millar J, Barron B, Hori K (2018) Accountability in AI: promoting greater societal trust. G7 Multistakeholder Conference on Artificial Intelligence, CIFAR, pp 1–15
Miller T, Howe P, Sonenberg L (2017) Explainable AI: beware of inmates running the asylum or: how i learnt to stop worrying and love the social and behavioural sciences. http://arxiv.org/abs/1712.00547
Misztal B (2013) Trust in modern societies: the search for the bases of social order. John Wiley & Sons
Mizanoor Rahman SM, Wang Y, Walker ID, Mears L, Pak R, Remy S (2016) Trust-based compliant robot-human handovers of payloads in collaborative assembly in flexible manufacturing. 2016 IEEE International Conference on Automation Science and Engineering (CASE), IEEE. pp. 355–360
Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M (2013) Playing atari with deep reinforcement learning. https://arxiv.org/abs/1312.5602
Molnar C (2020) Interpretable machine learning. Lulu. com
Molnar C, Casalicchio G, Bischl B (2019) Quantifying model complexity via functional decomposition for better post-hoc interpretability. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 193–204
Molnar C, Casalicchio G, Bischl B (2020) Interpretable machine learning–a brief history, state-of-the-art and challenges. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, pp 417–431
Montemayor C, Halpern J, Fairweather A (2021) In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. AI Soc. https://doi.org/10.1007/s00146-021-01230-z
Moravčík M, Schmid M, Burch N, Lisý V, Morrill D, Bard N, Davis T, Waugh K, Johanson M, Bowling M (2017) Deepstack: expert-level artificial intelligence in heads-up no-limit poker. Science 356(6337):508–513
Muddamsetty SM, Jahromi MNS, Moeslund TB (2021) Expert level evaluations for explainable AI (XAI) methods in the medical domain. International Conference on Pattern Recognition, ACM, pp 35–46
Muller T, Liu Y, Mauw S, Zhang J (2014) On robustness of trust systems. In: Zhou J, Gal-Oz N, Zhang J, Gudes E (eds). Springer. pp. 44–60
Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M (2020) Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 368:1–12
Nalepa GJ, Araszkiewicz M, Nowaczyk S, Bobek S (2019) Building Trust to AI Systems Through Explainability: Technical and Legal Perspectives. Proceedings of the 2nd Explainable AI in Law Workshop, CEUR
Nandi A, Pal AK (2022) Machine learning interpretability taxonomy. In: Interpreting machine learning models. Springer. pp. 35–44
Noor P (2020) Can we trust AI not to further embed racial bias and prejudice? BMJ m363. https://doi.org/10.1136/bmj.m363
Nourani M, Kabir S, Mohseni S, Ragan ED (2019) The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. Proc AAAI Conf Hum Comput Crowdsourcing 7:97–105. https://ojs.aaai.org/index.php/HCOMP/article/view/5284
Oelke D, Keim DA, Endert A, Keim D, Chau P, Reports D (n.d.) Report from Dagstuhl Seminar 20382 Creative Commons BY 3.0 Unported license Interactive Visualization for Fostering Trust in AI. Dagstuhl Reports, 10, 37–42. https://doi.org/10.4230/DagRep.10.4.37
Oh S, Kim JH, Choi S-W, Lee HJ, Hong J, Kwon SH (2019) Physician confidence in artificial intelligence: an online mobile survey. J Med Internet Res 21(3):e12422. https://doi.org/10.2196/12422
Okamura K, Yamada S (2020a) Adaptive trust calibration for human-AI collaboration. PLoS ONE 15(2). https://doi.org/10.1371/journal.pone.0229132
Okay FY, Yıldırım M, Özdemir S (n.d.) Interpretable machine learning: a case study of healthcare. 2021 International Symposium on Networks, Computers and Communications (ISNCC), IEEE, pp 1–6
Okamura K, Yamada S (2020b) Adaptive trust calibration for human-AI collaboration. PLoS ONE 15(2):e0229132. https://doi.org/10.1371/journal.pone.0229132
Olivero N, Lunt P (2004) Privacy versus willingness to disclose in e-commerce exchanges: The effect of risk awareness on the relative role of trust and control. J Econ Psychol 25(2):243–262
Ostherr K, Borodina S, Bracken RC, Lotterman C, Storer E, Williams B (2017) Trust and privacy in the context of user-generated health data. Big Data Soc 4(1):205395171770467. https://doi.org/10.1177/2053951717704673
Pan Q, Wu J, Li J, Yang W, Guan Z (2020) Blockchain and AI empowered trust-information-centric network for beyond 5G. IEEE Netw 34(6):38–45
Pan Z, Yang C-N, Sheng VS, Xiong N, Meng W (2019) Machine learning for wireless multimedia data security. In: Security and Communication Networks. vol. 2019. Hindawi
Papenmeier A, Englebienne G, Seifert C (2019) How model accuracy and explanation fidelity influence user trust. http://arxiv.org/abs/1907.12652
Pawar U, O'Shea D, Rea S, O'Reilly R (2020) Incorporating explainable artificial intelligence (XAI) to aid the understanding of machine learning in the healthcare domain. AICS. pp. 169–180
Peake G, Wang J (2018) Explanation mining: post hoc interpretability of latent factor models for recommendation systems. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, pp 2060–2069
Peukert C, Kloker S (2020) Trustworthy AI: how ethicswashing undermines consumer trust. In: WI2020 Zentrale Tracks. GITO Verlag. pp. 1100–1115. https://doi.org/10.30844/wi_2020_j11-peukert
Pickering B (2021) Trust, but verify: informed consent, AI technologies, and public health emergencies. Future Internet 13(5):132. https://doi.org/10.3390/fi13050132
Pieters W (2011a) Explanation and trust: What to tell the user in security and AI. Ethics Inf Technol 13(1):53–64. https://doi.org/10.1007/s10676-010-9253-3
Pieters W (2011b) Explanation and trust: what to tell the user in security and AI. Ethics Inf Technol 13(1):53–64. https://doi.org/10.1007/s10676-010-9253-3
Pieters W (2011c) Explanation and trust: what to tell the user in security and AI. Ethics Inf Technol 13(1):53–64. https://doi.org/10.1007/s10676-010-9253-3
Pintelas E, Livieris IE, Pintelas P (2020) A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. Algorithms 13(1):17
Pitardi V, Marriott HR (2021a) Alexa, she's not human but… unveiling the drivers of consumers' trust in voice-based artificial intelligence. Psychol Mark 38(4):626–642. https://doi.org/10.1002/mar.21457
Pitardi V, Marriott HR (2021b) Alexa, she's not human but… unveiling the drivers of consumers' trust in voice-based artificial intelligence. Psychol Mark 38(4):626–642. https://doi.org/10.1002/mar.21457
Prasad M (2019) Social choice and the value alignment problem. In: Yampolskiy RV (ed.). Artificial Intelligence Safety and Security. CRC Press. pp. 291–314
Qayyum A, Usama M, Qadir J, Al-Fuqaha A (2020) Securing connected & autonomous vehicles: challenges posed by adversarial machine learning and the way forward. IEEE Commun Surv Tutor 22(2):998–1026
Reuben J (2018) Towards a differential privacy theory for edge-labeled directed graphs. SICHERHEIT, Gesellschaft Für Informatik
Richards NM, Hartzog W (2015) Taking trust seriously in privacy law. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2655719
Robinson SC (2020) Trust, transparency, and openness: how inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). Technol Soc 63. https://doi.org/10.1016/j.techsoc.2020.101421
Robotics Australia Group (2022) A robotics roadmap for Australia. Www.Roboausnet.Com.Au
Roessingh JJ, Toubman A, van Oijen J, Poppinga G, Hou M, Luotsinen L (2017) Machine learning techniques for autonomous agents in military simulations—Multum in Parvo. 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE. pp. 3445–3450
Roski J, Maier EJ, Vigilante K, Kane EA, Matheny ME (2021) Enhancing trust in AI through industry self-governance. J Am Med Inf Assoc 28(7):1582–1590
Ross K (2020, June 13) Data deception: how data provenance failure undermines trust in AI analytics. Datarwe, 395(10240). https://doi.org/10.1016/S0140-6736(20)31290-3
Roszel M, Norvill R, Hilger J, State R (2021) Know your model (KYM): increasing trust in AI and machine learning. http://arxiv.org/abs/2106.11036
Roth-Berghofer TR, Cassens J (2005) Mapping goals and kinds of explanations to the knowledge containers of case-based reasoning systems. In: Muñoz-Ávila H, Ricci F (eds.). Case-based reasoning research and development. Springer Berlin Heidelberg. pp. 451–464
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. Springer Science and Business Media LLC
Ryan M (2020a) In AI we trust: ethics, artificial intelligence, and reliability. Sci Eng Ethics 26(5):2749–2767. https://doi.org/10.1007/s11948-020-00228-y
Ryan M (2020b) In AI we trust: ethics, artificial intelligence, and reliability. Sci Eng Ethics 26(5):2749–2767. https://doi.org/10.1007/s11948-020-00228-y
Ryan PA (2017) Smart contract relations in e-commerce: legal implications of exchanges conducted on the blockchain. Technol Innov Manag Rev 7(10):14–21
Sakai T, Nagai T (2022) Explainable autonomous robots: a survey and perspective. Adv Robot 36(5–6):219–238
Salem M, Dautenhahn K (2015) Evaluating trust and safety in HRI: practical issues and ethical challenges. http://uhra.herts.ac.uk/handle/2299/16336
Sarpatwar K, Ganapavarapu VS, Shanmugam K, Rahman A, Vaculin R (2019) Blockchain enabled AI marketplace: the price you pay for trust. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). https://openaccess.thecvf.com/content_CVPRW_2019/html/BCMCVAI/Sarpatwar_Blockchain_Enabled_AI_Marketplace_The_Price_You_Pay_for_Trust_CVPRW_2019_paper.html

Saßmannshausen T, Burggräf P, Wagner J, Hassenzahl M, Heupel T, Steinberg F (2021) Trust in artificial intelligence within production management–an exploration of antecedents. Ergonomics 64(10):1333–1350. https://doi.org/10.1080/00140139.2021.1909755
Scharowski N, Brühlmann F (2020) Transparency and trust in AI: measuring the effect of human-friendly AI explanations on objective and subjective trust
Schlicker N, Langer M (2021) Towards warranted trust: a model on the relation between actual and perceived system trustworthiness. ACM International Conference Proceeding Series. pp. 325–329. https://doi.org/10.1145/3473856.3474018
Schmidt P, Biessmann F (2019) Quantifying interpretability and trust in machine learning systems. https://arxiv.org/abs/1901.08558
Schmidt P, Biessmann F, Teubner T (2020a) Transparency and trust in artificial intelligence systems. J Decis Syst 29(4):260–278. https://doi.org/10.1080/12460125.2020.1819094
Schmidt P, Biessmann F, Teubner T (2020b) Transparency and trust in artificial intelligence systems. J Decis Syst 29(4):260–278. https://doi.org/10.1080/12460125.2020.1819094
Schwarting W, Alonso-Mora J, Rus D (2018) Planning and decision-making for autonomous vehicles. Annual Review of Control. Robot Autonomous Syst 1:187–210
Sengupta PP, Chandrashekhar YS (2021) Building trust in AI: opportunities and challenges for cardiac imaging. JACC Cardiovasc Imaging 14(2):520–522. https://doi.org/10.1016/j.jcmg.2021.01.002
Shaban-Nejad A, Michalowski M, Brownstein JS, Buckeridge DL(2021) Guest editorial explainable AI: towards fairness, accountability, transparency and trust in healthcare IEEE J Biomed Health Inform 25(7):2374–2375. https://doi.org/10.1109/JBHI.2021.3088832
Shaban-Nejad A, Michalowski M, Brownstein JS, Buckeridge DL (2021b) Guest editorial explainable AI: towards fairness, accountability, transparency and trust in healthcare. IEEE J Biomed Health Inform 25(7):2374–2375. https://doi.org/10.1109/JBHI.2021.3088832
Shafiei SB, Hussein AA, Muldoon SF, Guru KA (2018) Functional brain states measure mentor-trainee trust during robot-assisted surgery. Sci Rep 8(1):3667. https://doi.org/10.1038/s41598-018-22025-1
Shailaja K, Seetharamulu B, Jabbar MA (2018) Machine learning in healthcare: a review. 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, pp 910–914
Sharan NN, Romano DM (2020) The effects of personality and locus of control on trust in humans versus artificial intelligence. Heliyon 6(8). https://doi.org/10.1016/j.heliyon.2020.e04572
Shi S, Gong Y, Gursoy D (2021) Antecedents of trust and adoption intention toward artificially intelligent recommendation systems in travel planning: a heuristic–systematic Model. J Travel Res 60(8):1714–1734. https://doi.org/10.1177/0047287520966395
Siau K (2018) Building trust in artificial intelligence, machine learning, and robotics supply chain management view project. www.cutter.com
Silva W, Fernandes K, Cardoso JS (2019) How to produce complementary explanations using an ensemble model. 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, pp 1–8
Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529(7587):484–489
Skopik F, Schall D, Dustdar S (2009) Start Trusting Strangers? Bootstrapping and Prediction of Trust. In: Vossen G, Long DD, Yu JX (eds.). Proceedings of the 10th International Conference on Web Information Systems. Springer-Verlag. pp. 275–289 http://www.citeulike.org
Sligar AP (2020) Machine learning-based radar perception for autonomous vehicles using full physics simulation. IEEE Access 8:51470–51476
Smith-Renner A, Fan R, Birchfield M, Wu T, Boyd-Graber J, Weld DS, Findlater L (2020, April 21) No explainability without accountability: an empirical study of explanations and feedback in interactive ML. Conference on Human Factors in Computing Systems— Proceedings. https://doi.org/10.1145/3313831.3376624
Song Y, Luximon Y (2020) Trust in AI agent: A systematic review of facial anthropomorphic trustworthiness for social robot design. Sensors 20(18):1–21). https://doi.org/10.3390/s20185087
Song Z, Tang Z, Liu H, Guo D, Cai J, Zhou Z (2021) A clinical-radiomics nomogram may provide a personalized 90-day functional outcome assessment for spontaneous intracerebral hemorrhage. Eur Radiol 31(7):4949–4959. https://doi.org/10.1007/s00330-021-07828-7
Sperrle F, El-Assady M, Guo G, Chau DH, Endert A, Keim D (2020) Should we trust (X)AI? Design dimensions for structured experimental evaluations. http://arxiv.org/abs/2009.06433
Spiegelhalter D (2020) Should we trust algorithms? Harv Data Sci Rev 2(1):1–12. https://doi.org/10.1162/99608f92.cb91a35a
Spreitzer GM (1995) Psychological empowerment in the workplace: dimensions, measurement, and validation. Acad Manag J 38(5):1442–1465. https://doi.org/10.2307/256865
Srinivasan AV (2019) Developing a model for improving trust in artificial intelligence. Technology, Policy and Management, Technology, Policy and Management, TU Delft
Srinivasan R, San Miguel González B (2022) The role of empathy for artificial intelligence accountability. J Responsible Technol 9:100021. https://doi.org/10.1016/j.jrt.2021.100021
Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L (2020) Interpretability of machine learning-based prediction models in healthcare. Wiley Interdiscip Rev Data Min Knowl Discov 10(5):e1379
Sullivan Y, Bourmont M, Dunaway M (2022) Appraisals of harms and injustice trigger an eerie feeling that decreases trust in artificial intelligence systems. Ann Oper Res 308(1):525–548
Szeli L (2020a) UX in AI: trust in algorithm-based investment decisions. Jr Manag Sci 5(1):1–18
Szeli L (2020b) UX in AI: trust in algorithm-based investment decisions. Jr Manag Sci 5(1):1–18
Taddeo M, McCutcheon T, Floridi L (2019) Trusting artificial intelligence in cybersecurity is a double-edged sword. Nat Mach Intell 1(12):557–560. https://doi.org/10.1038/s42256-019-0109-1
Tallant J (2017) Commitment in cases of trust and distrust. Thought 6(4):261–267
Taylor J, Yudkowsky E, Lavictoire P, Critch A (2016) Alignment for advanced machine learning systems
The racist hijacking of Microsoft's chatbot shows how the internet teems with hate | Paul Mason | The Guardian. (n.d.) Retrieved March 23, 2022, from https://www.theguardian.com/world/2016/mar/29/microsoft-tay-tweets-antisemitic-racism
Thelisson E (2017) Towards trust, transparency and liability in AI/AS systems. IJCAI. pp. 5215–5216
Thiebes S, Lins S, Sunyaev A (2021a) Trustworthy artificial intelligence. Electron Mark 31(2):447–464. https://doi.org/10.1007/s12525-020-00441-4
Thiebes S, Lins S, Sunyaev A (2021b) Trustworthy artificial intelligence. Electron Mark 31(2):447–464. https://doi.org/10.1007/s12525-020-00441-4
Thiebes S, Lins S, Sunyaev A (2021c) Trustworthy artificial intelligence. Electron Mark 31(2):447–464. https://doi.org/10.1007/s12525-020-00441-4
Tomsett R, Braines D, Harborne D, Preece A, Chakraborty S (2018) Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. https://arxiv.org/abs/1806.07552
Tomsett R, Preece A, Braines D, Cerutti F, Chakraborty S, Srivastava M, Pearson G, Kaplan L (2020) Rapid trust calibration through interpretable and uncertainty-aware AI. Patterns 1(4):100049. https://doi.org/10.1016/j.patter.2020.100049
Troshani I, Rao Hill S, Sherman C, Arthur D (2021) Do we trust in AI? Role of anthropomorphism and intelligence. J Comput Inf Syst 61(5):481–491. https://doi.org/10.1080/08874417.2020.1788473
Tschopp M (2019, July 18) Artificial intelligence: is it worth the risk? SCIP. https://www.scip.ch/en/?labs.20190718
Tutul AA, Nirjhar EH, Chaspari T (2021a) Investigating trust in human-machine learning collaboration: a pilot study on estimating public anxiety from speech. ICMI 2021—Proceedings of the 2021 International Conference on Multimodal Interaction. pp. 288–296. https://doi.org/10.1145/3462244.3479926
Tutul AA, Nirjhar EH, Chaspari T (2021b) Investigating trust in human-machine learning collaboration: a pilot study on estimating public anxiety from speech. ICMI 2021—Proceedings of the 2021 International Conference on Multimodal Interaction. pp. 288–296. https://doi.org/10.1145/3462244.3479926
van Dyke TP, Midha V, Nemati H (2007a) The effect of consumer privacy empowerment on trust and privacy concerns in e-commerce. Electron Mark 17(1):68–81. https://doi.org/10.1080/10196780601136997
van Dyke TP, Midha V, Nemati H (2007b) The effect of consumer privacy empowerment on trust and privacy concerns in e-commerce. Electron Mark 17(1):68–81. https://doi.org/10.1080/10196780601136997
Varshney KR (2019) Trustworthy machine learning and artificial intelligence. XRDS: Crossroads, The ACM Magazine for Students 25.3:26–29
Varshney KR (2022) Trustworthy machine learning. files/8693/Varshney-2022-Trustworthy Machine Learning.pdf
Vereschak O, Bailly G, Caramiaux B (2021) On the way to improving experimental protocols to evaluate users' trust in AI-assisted decision making. https://hal.sorbonne-universite.fr/hal-03418712
Villani C (2018) For a meaningful artificial intelligence. A parliamentary mission from 8th September 2017 to 8th March 2018
Vodrahalli K, Gerstenberg T, Zou J (2021) Do humans trust advice more if it comes from AI? An analysis of human-AI interactions. http://arxiv.org/abs/2107.07015
Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, Cumbers S, Jonas A, McAllister KSL, Myles P, others (2018) Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. Preprint at arXiv:1812.10404

von Eschenbach WJ (2021) Transparency and the Black Box Problem: why we do not trust AI. Philos Technol 34(4):1607–1622. https://doi.org/10.1007/s13347-021-00477-0
Wagner AR, Borenstein J, Howard A (2018) Computing ethics overtrust in the robotic age. In: Communications of the ACM (vol. 61, issue 9). Association for Computing Machinery. pp. 22–24
Wagner C (2009) 'The Japanese way of robotics': Interacting 'naturally' with robots as a national character? RO-MAN 2009—The 18th IEEE International Symposium on Robot and Human Interactive Communication. IEEE. pp. 510–515
Wang J, Moulden A (2021) AI trust score: a user-centered approach to building, designing, and measuring the success of intelligent workplace features. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–7). Association for Computing Machinery. pp. 1–7
Wang M, Liu S, Zhu Z (2020) Can you trust AI-assisted network automation? A DRL-based approach to mislead the automation in SD-IPoEONs. https://github.com/lsq93325/Traffic-creation/blob/master/README.md?tdsourcetag=s
Wang N, Pynadath DV, Hill SG (2015) Building trust in a human-robot team with automatically generated explanations. Los Angeles. 12. files/5941/Wang et al. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2015
Wang W (2021a) Trust in AI-based conversational agents in the customer service—a theoretical framework. AMCIS 2021 TREOs. https://aisel.aisnet.org/treos_amcis2021/45
Wang W (2021b) Trust in AI-based conversational agents in the customer service-a theoretical framework
Wang W, Siau K (2018) Living with artificial intelligence–developing a theory on trust in health Chatbots. Proceedings of the Sixteenth Annual Pre-ICIS Workshop on HCI Research in MIS
Wang Y, Singh MP (n.d.) Trust representation and aggregation in a distributed agent System. AAAI. www.aaai.org
Weitz K, Schiller D, Schlagowski R, Huber T, André E (2019) "Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, ACM, pp 7–9
Weller A (2017) Transparency: motivations and challenges. http://arxiv.org/abs/1708.01870
White paper on artificial intelligence: a European approach to excellence and trust. (2020) European Commission
Wiens J, Shenoy ES (2018) Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis 66(1):149–153
Williams R, Cloete R, Cobbe J, Cotterill C, Edwards P, Markovic M, Naja I, Ryan F, Singh J, Pang W (2022) From transparency to accountability of intelligent systems: moving beyond aspirations. Data Policy, vol 4, issue 7. Cambridge University Press, p e7
Wu D, Huang Y (2021a) Why do you trust Siri?: The factors affecting trustworthiness of intelligent personal assistant. Proc Assoc Inf Sci Technol 58(1):366–379. https://doi.org/10.1002/pra2.464
Wu D, Huang Y (2021b) Why do you trust Siri?: The factors affecting trustworthiness of intelligent personal assistant. Proc Assoc Inf Sci Technol 58(1):366–379. https://doi.org/10.1002/pra2.464
Xu G, Zhao Y, Jiao L, Feng M, Ji Z, Panaousis E, Chen S, Zheng X (2021) TT-SVD: an efficient sparse decision-making model with two-way trust recommendation in the AI-enabled IoT systems. IEEE Internet Things J 8(12):9559–9567. https://doi.org/10.1109/JIOT.2020.3006066
Yan A, Xu D (2021a) AI for depression treatment: addressing the paradox of privacy and trust with empathy, accountability, and explainability. International Conference on Information Systems (ICIS 2021): Building Sustainability and Resilience with IS: a Call for Action, 1937. Association for Information Systems
Yan A, Xu D (2021b) AI for depression treatment: addressing the paradox of privacy and trust with empathy, accountability, and explainability. International Conference on Information Systems (ICIS 2021): Building Sustainability and Resilience with IS: A Call for Action, 1937. Association for Information Systems
Yang L, Zhang Z, Xiong S, Wei L, Ng J, Xu L, Dong R (2018) Explainable text-driven neural network for stock prediction. 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS). IEEE. pp. 441–445
Yoon N, Lee H-K (2021) AI recommendation service acceptance: assessing the effects of perceived empathy and need for cognition. J Theor Appl Electron Commer Res 16(5):1912–1928. https://doi.org/10.3390/jtaer16050107
Zarifis A, Kawalek P, Azadegan A (2021) Evaluating if trust and personal information privacy concerns are barriers to using health insurance that explicitly utilizes AI. J Internet Commer 20(1):66–83. https://doi.org/10.1080/15332861.2020.1832817
Zhang C, Li W, Luo Y, Hu Y (2021) AIT: an AI-enabled trust management system for vehicular networks using blockchain technology. IEEE Internet Things J 8(5):3157–3169. https://doi.org/10.1109/JIOT.2020.3044296
Zhang Y, Liao QV, Bellamy RKE (2020) Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. 295–305. https://doi.org/10.1145/3351095.3372852
Zhao J, Abrahamson K, Anderson JG, Ha S, Widdows R (2013) Trust, empathy, social identity, and contribution of knowledge within patient online communities. Behav Inf Technol 32(10):1041–1048. https://doi.org/10.1080/0144929X.2013.819529
Zhou J, Verma S, Mittal M, Chen F (2021) Understanding relations between perception of fairness and trust in algorithmic decision making. 2021 8th International Conference on Behavioral and Social Computing (BESC). pp. 1–5. https://doi.org/10.1109/BESC53957.2021.9635182
Ziegler C-N, Lausen G (2004) Analyzing COrrelation between Trust and User Similarity in Online Communities. In: Jensen C, Poslad S, Dimitrakos T (eds). Springer. pp. 251–265
Zierau N, Engel C, Söllner M, Leimeister JM (2020) Trust in smart personal assistants: a systematic literature review and development of a research agenda. In: WI2020 Zentrale Tracks. GITO Verlag. pp. 99–114. https://doi.org/10.30844/wi_2020_a7-zierau
Zierau N, Flock K, Janson A, Söllner M, Leimeister JM (2021) The influence of AI-based Chatbots and their design on users' trust and information sharing in online loan applications
Zierau N, Hausch M, Bruhin O, Söllner M (2020) Towards developing trust-supporting design features for AI-based Chatbots in customer service
Zolanvari M, Yang Z, Khan K, Jain R, Meskin N (2021) TRUST XAI: model-agnostic explanations for AI with a case study on IIoT security. IEEE Internet Things J 10(4):2967–2978

Acknowledgements

We are grateful to Jason D'Cruz (SUNY) and Kush Raj Varshney (IBM), with whom, in the course of this research, we discussed earlier versions of this paper and had the benefit of their insightful comments and suggestions, which helped me in revising several sections of the manuscript. We would also like to show our gratitude to Yaser Pouresmail and Mohsen Javaherian for their help and invaluable comments and discussion.

Author contributions

SA, AA: investigation, conceptualization, methodology, data curation, formal analysis, original draft, writing, review, and editing. EM, MK, HA: formal analysis, review, and visualization.

Competing interests

The author(s) declares no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

Informed consent was not required as this study did not involve human participants.

Additional information

Correspondence and requests for materials should be addressed to Saleh Afroogh.