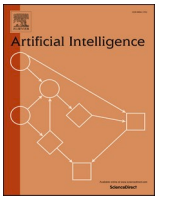
Artificial Intelligence 342 (2025) 104309

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



(Re)Conceptualizing trustworthy AI: A foundation for change

Christopher D. Wirz [a,b,*], Julie L. Demuth [a,b], Ann Bostrom [b,c], Mariana G. Cains [a,b], Imme Ebert-Uphoff [b,d], David John Gagne II [a,b], Andrea Schumacher [a,b], Amy McGovern [b,e], Deianna Madlambayan [b,c]

[a] NSF National Center for Atmospheric Research, Boulder, CO, USA
[b] NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES), USA
[c] University of Washington, Seattle, WA, USA
[d] Colorado State University, Fort Collins, CO, USA
[e] The University of Oklahoma, Norman, OK, USA

ARTICLE INFO

Keywords:
Trust
Artificial intelligence
Machine learning (ML)
Interdisciplinary
Trustworthy AI

ABSTRACT

Developers and academics have grown increasingly interested in developing "trustworthy" artificial intelligence (AI). However, this aim is difficult to achieve in practice, especially given trust and trustworthiness are complex, multifaceted concepts that cannot be completely guaranteed nor built entirely into an AI system. We have drawn on the breadth of trust-related literature across multiple disciplines and fields to synthesize knowledge pertaining to interpersonal trust, trust in automation, and risk and trust. Based on this review we have (re)conceptualized trustworthiness in practice as being both (a) perceptual, meaning that a user assesses whether, when, and to what extent AI model output is trustworthy, even if it has been developed in adherence to AI trustworthiness standards, and (b) context-dependent, meaning that a user's perceived trustworthiness and use of an AI model can vary based on the specifics of their situation (e.g., time-pressures for decision-making, high-stakes decisions). We provide our reconceptualization to nuance how trustworthiness is thought about, studied, and evaluated by the AI community in ways that are more aligned with past theoretical research.

1. Introduction

As the advancement of artificial intelligence (AI) has accelerated, the notions of trustworthy AI and trust in AI have garnered a great deal of attention from researchers [1,41,45,64,98]. These notions are not only important in academic circles but have also captured the attention and efforts of policy and regulatory groups around the world. For example, various national and international groups have published guidelines and recommendations that emphasize the importance of trustworthy AI, including the U.S. executive [23,24,72] and legislative [27,28] branches, the U.S. National Institute of Standards and Technology [70], the Organisation for Economic Co-operation and Development [71], the High-level Expert Group on Artificial Intelligence [34], and the European Parliament [22].

The growing attention to AI comes from a recognition that AI poses many major societal risks. These risks are exacerbated by the rapid development of and adoption of AI in high-stakes industries, such as healthcare (e.g., [47]), military operations (e.g., [82]), weather forecasting (e.g., [6]), criminal justice (e.g., [59]), and human resources (e.g., [57]). Furthermore, the rapid emergence of

* Corresponding author at: NSF National Center for Atmospheric Research, USA and University of Illinois Urbana-Champaign, USA
E-mail addresses: cdwirz@ucar.edu, wirz@illinois.edu (C.D. Wirz).

https://doi.org/10.1016/j.artint.2025.104309
Received 21 March 2024; Received in revised form 30 January 2025; Accepted 18 February 2025
Available online 22 February 2025

generative AI techniques is raising public exposure to risks of AI to an entirely new level, such as by providing capabilities for text generation with ChatGPT [32]. With rapid technology advances come many new possibilities for beneficial outcomes, but also for risks to society that may arise from both malicious and well-intentioned efforts. To this end, governments and research communities are increasingly motivated to find ways to develop AI that is trustworthy, as well as increasing trust in AI in some cases and sowing some distrust in other cases. These motivations, which we will nuance throughout this paper, appear repeatedly in research and policy publications [102], many of which focus on specific requirements (or principles, prerequisites, standards, etc.) for developing AI. Such efforts are important for guiding development practices and are often referred to as resulting in "trustworthy AI."

Trust and trustworthiness are concepts that have rich theoretical and empirical foundations in a range of social science literatures [11,35,51,53,60,75,80,90,96]. Although this knowledge has been leveraged to some degree by the AI research and development communities (e.g., [5,9,97]), many trustworthy AI efforts would benefit from a more robust grounding in these foundational trust literatures [26]. Current research and policy efforts related to the trustworthiness of AI are built on a wide range of definitions, understandings, theoretical traditions, disciplinary views, and assumptions of what trust and trustworthiness are and are not. This represents an opportunity to recenter and further inform discussions about AI trustworthiness and trust, which are complex and value-laden concepts, with insights from the extensive and robust social science literatures (e.g., [2,69]). These literatures have a great deal to offer in how AI researchers and developers understand and define, or conceptualize, what trustworthy AI means and how it is put into practice, or operationalized.

Although the social science literature is wide-ranging, a central tenet is that trust and trustworthiness are *relational* concepts (e.g., [25,55,78]). This means they require both a *trustor*, the one doing the trusting, and a *trustee*, the one being trusted (a person, object, model, etc.). In the AI domain, the predominant focus is on AI as the trustee [39,101], which is a logical starting point given the overarching goals of fostering desired levels of trust in AI and making AI worthy of people's trust. However, the relational nature of trust and trustworthiness requires the potential trustors also be considered—arguably as much as AI as the trustee. The evaluation of a trustee's trustworthiness and trust in that trustee must come from a trustor. Both parties are needed. Yet, the heavy focus on AI has resulted in an under-consideration, or in some cases seemingly ignoring, of the people whose trust is being sought.

To help facilitate a deeper consideration of the trustor, our paper lays out a strategy for the AI community to integrate foundational research on trust into its efforts in a few ways. First, we provide a review and synthesis of key literature pertaining to interpersonal trust, trust in automation, and risk and trust with an eye toward what they can contribute to thinking more systematically and holistically about trustworthy AI. We synthesize theoretical and practical social science in a way that is meant to be accessible to those unfamiliar with trust research and theory. Our aim is to connect these literatures and some of their lessons with the goals and efforts of the trustworthy AI community.

Based on this review, we then offer a (re)conceptualization of AI trustworthiness that elevates the trustor's perspective as an equal and necessary (but to date underrepresented) dimension in efforts to develop and evaluate trustworthy AI. We use the term reconceptualization here to convey that, rather than providing an entirely new way of understanding trustworthy AI, we are integrating strengths and important insights from a range of conceptualizations, relating both to trustworthy AI and trustworthiness and trust in general, into one overarching conceptualization for AI researchers and developers.

Specifically, we advocate reconceptualizing AI trustworthiness from the trustors' perspective as both (a) perceptual, meaning that a given trustor (user, developer, policymaker, community member, etc.) assesses whether, when, and to what extent AI is trustworthy, **even if it has been developed in adherence to AI trustworthiness standards or policies**, and (b) context-dependent, meaning that

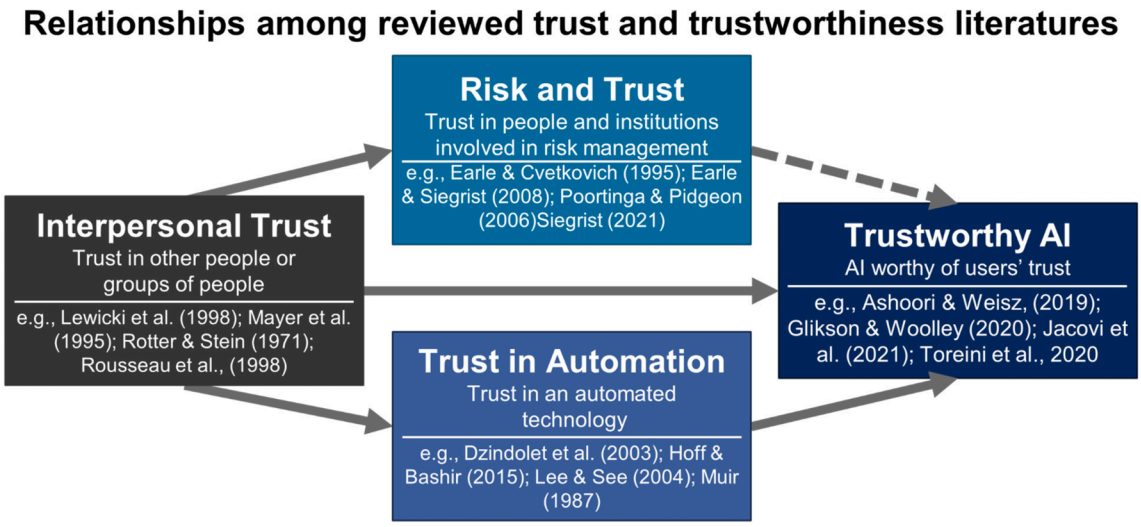


Fig. 1. Conceptual representation of the relationship among the literatures reviewed in this paper. Arrows reference one literature influencing another. The dashed arrows from risk and trust to trustworthy AI indicates the potential for a contribution that has not yet been actualized. Each literature has an overview of its broad focus and some example citations.

a trustor's perceived trustworthiness and trust in an AI model can vary based on their specific use of the model and related situational aspects (e.g., time-pressures for decision-making, high-stakes decisions, prior experience with the model).

Our reconceptualization lays the foundation for more effective and coherent research, development, implementation, and regulation of AI, especially as such efforts work to keep up with the rapid evolution of AI advancements. We offer ideas for how to elevate the perspectives and needs of the trustor in practice when developing and operationalizing trustworthy AI. In the end, development and policy efforts that focus both on AI and its potential trustors are more likely to lead to AI that is deemed trustworthy, trusted, and used.

2. Theoretical and empirical foundations for trust and trustworthiness

Trust and trustworthiness have long been studied across many disciplines. For example, we see discussions of trust and trustworthiness in the context of communication during the 1950s (e.g., [29,40,58,66]) and more research specifically on trust in the 1960s (e.g., [11–13,79]). In the decades since, trust and trustworthiness have received attention from many fields, such as organizational research (e.g., [53,60]), automation (e.g., [35,51,67]), philosophy (e.g., [21,43,99]), and risk (e.g., [19,20,76,89]). Several reviews have documented the expansiveness of this literature and the many definitions that have been introduced within it (e.g., [2,69]).

Across this rich theoretical and empirical space, there are three broad lines of research that are especially relevant for understanding trustworthiness in the context of AI: (1) interpersonal trust, (2) trust in automation, and (3) risk and trust (Fig 1). Each of these lines of research highlights how trust and trustworthiness are both relational in different ways. One such way is the literatures have focused on different trustor-trustee relationships – interpersonal trust focuses on trust between and among people, trust in automation focuses on how experts and non-experts trust different technologies, and risk and trust focuses how individuals trust people and institutions involved in risk management, often of specific hazards. In this section we briefly highlight these distinct, yet related, lines of research to provide the context for our reconceptualization of trustworthiness in the context of AI. In each section, we highlight and provide a brief background on some of the seminal papers and ideas from each field, then provide commentary and pointed questions highlighting how prior findings are and are not transferable to diverse AI contexts. The goal of Section 2 is to provide those interested in trust and trustworthiness with a broad, high-level overview of these literatures so readers are familiar with their similarities and differences, as well as how they relate to work on trustworthy AI. This foundation then paves the way for our reconceptualization in Section 3, which draws on literature from across these three areas and focuses more specifically on the perceptual and context-dependent nature of trust and trustworthiness.

2.1. Interpersonal trust

Interpersonal trust, especially interpersonal trust within organizations[1], has been the focal point for arguably the bulk of the foundational research on trust and trustworthiness. From a relational perspective, this body of literature focuses on contexts of trust and trustworthiness in which *both the trustees and trustors are people or groups of people.* Some of the most cited and prominent conceptualizations and definitions come out of this broad area of research (e.g., [60]), several of which are often cited in papers about trustworthy AI. Much of this research focuses primarily on trust, as opposed to trustworthiness, and broadly agrees that trust is relational in nature. However, there is less agreement on exactly what this relationship looks like and which factors and related concepts are required and/or important for it. For example, several of the foundational trust researchers approached trust largely through the lens of 'cooperation' and were trying to understand how multiple parties negotiated differences to reach an agreeable goal or outcome (e.g., [12,58]). This approach requires the involved parties to have some sort of mutual dependency on one another, or what Rousseau et al. [81] termed 'interdependence.' Research that follows these traditions and ideas tends to conceptualize trust as being more collaborative and relational among people who need to work together and/or rely on one another to achieve a goal. Although goals and intended outcomes are central to interpersonal trust research, other conceptualizations focus less on coordination and more on the perceptions of the trustor. The broad consensus in this line of work is that trust is based on a trustor's expectation of favorable, or at least acceptable, outcomes from the actions or conduct of another party (e.g., [53,79,93]). Many of these definitions are tailored specifically to trust in other people and depend on several factors, such as whether or not the trustor perceives the trustee to be altruistic [61] or share the trustor's values [93].

One major theme across all these conceptualizations, whether explicitly or implicitly mentioned, is the idea that trust is inherently connected to risk, or the chance of some negative or less than ideal outcome. We see the focus on risk from early trust work (e.g., [12]) to its instrumental role in the widely cited Mayer et al. [60] definition of trust, which emphasizes the willingness to assume risk. This line of work argues that this willingness of the trustor to assume risk, or willingness to be vulnerable, is essential for trust because, without the uncertainty inherent in risk, there would be no need for trust [54,81]. In essence, if one *knows* with total certainty that someone is going to act a certain way or do a certain thing, one does not need to *trust* them to do it. However, this is rarely the case for interpersonal trust—the behavior of others is often uncertain, and thus one assumes varying degrees of risk when trusting others. Ironically, despite the intrinsic element of risk, this literature has not drawn on the field of risk and trust (Section 2.3). Together these high-level points represent the broad understanding of trust from the interpersonal perspective and provide a baseline for the main points that we think are important for digging more into how to conceptualize the trustworthiness of AI.

[1] Note some of what we have broadly labeled "interpersonal" trust also examines trust in "organizations" as well.

A helpful way to understand the perspective of this interpersonal trust literature and how it influences other lines of research and thinking is through the way it approaches trustworthiness. Namely, there are three main characteristics important for assessing trustworthiness in the interpersonal context: *ability, benevolence*, and *integrity* [60]. These characteristics broadly reflect the priorities and focus of this literature, which is useful for understanding the problems these researchers are interested in and how this literature differs from others we will review. For example, *ability* is relatively universal across trust literatures in that who or what is being trusted needs to be able to follow through on what they are being trusted for or evaluated on. However, *benevolence* (the idea that the trustee wants to do good by the trustor) and *integrity* (the adherence to moral principles the trustee deems acceptable) demonstrate how interpersonal trust research focuses heavily on human dimensions that, by most definitions, require thought or consciousness on the part of the trustee [60]. The focus on trust and trustworthiness as being tied to concepts typically reserved for human actors has raised debates about whether or not we should be using the terms in the context of AI (e.g., [26]).

2.2. Trust in automation

The trust-in-automation literature developed largely as a way to better understand human-machine interactions in contexts where the machine is making decisions or has some level of control in situations where that task was originally performed by a human [51, 74]. In this domain, the *trustees are types of automated systems or technologies, and the trustors are people or groups of people.* Relatedly, trust in technology has also been examined in other literatures (e.g., [49,65]). For example, a more recent line of work has emerged in communication research focusing on trust in information systems, many of which use AI algorithms to tailor information to users (e.g., [77,86]). This work has focused on how users' perceptions relate to trust in related actors, like the organization who produced the system (e.g., [96]), and concepts like fairness, accountability, and transparency (e.g., [85]). In this section we focus on trust in automation but note the trust in technology literature also fits within this general portion of the review.

The foundational goals of the trust in automation literature are remarkably similar to those of many researchers and developers concerned with trustworthy AI (e.g., [84]). For example, Muir [[67], p. 527] posed the question "How can we design decision aids which decision makers will appropriately trust and use?" Trust-in-automation scholars set out on a similar path as many of those who have studied trust in AI, which was to begin by reviewing the interpersonal trust literature for insights that could be transferred to the human-machine context (Fig. 1).

However, there are differences in how the concepts of automation and AI have been used in relationship with one another, although they are generally considered distinct concepts (e.g., [83]). The similarity between the two depends on which particular definitions people use, but generally automation refers to processes (mechanical or software) that are to complete tasks that were traditionally done by humans (e.g., [68]). Based on this general understanding, not all automation relies on AI but some AI applications qualify as automation. For example, the first telephone answering machines automated the ability to collect simple voice messages but did not rely on AI, while chatbots on websites use AI to automate the answering of questions so humans do not have to answer them via phone calls.

Despite their differences, the broad similarities between AI and automation mean there are key insights from the trust-in-automation work that we can leverage to better understand and conceptualize trustworthiness in the context of AI. One such point is the extent to which interpersonal conceptualizations of trust apply to non-human entities. As Lee and See [51] point out in their seminal paper, the interpersonal trust conceptualizations require, either explicitly or implicitly, intentionality and relationship symmetry that do not exist in the context of trust in automation, at least not directly. With this point in mind, Lee and See [[51], p. 51] developed the following definition of trust that is popular in the subsequent trust-in-automation literature: "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." This definition builds on the importance of risk and vulnerability, discussed in the previous section, but pivots away from intentionality and relational symmetry to focus on achieving a goal of the trustee. This is an important distinction that we see in discussions of human-AI trust through lenses like "human-AI teaming" [69]. The focus on helping achieve one's goals resembles past conceptualizations that emphasized how trust facilitates coordination (e.g., [12,58]) more than it does conceptualizations that focus on the role trust plays in human-to-human relationships (e.g., [60,81]). The emphasis of the trust-in-automation literature on goals or outcomes shapes how trust and trustworthiness are conceptualized and discussed in this space.

There are three overarching factors from this literature that are generally discussed as contributing to the trustworthiness of automation: *performance, process*, and *purpose* [51,52]. Although we focus on these three factors as being broadly representative of the field, we note several scholars have developed their own factors and theories in this area that we encourage those interested to review as well [16,35,42]. Nonetheless, *performance* broadly refers to what the automation is doing and how well it can actually do it, which mirrors the *ability* dimension of Mayer et al.'s interpersonal trust framework. *Process* addresses how the automation is doing what it is doing and its underlying characteristics, which somewhat mirrors the *integrity* dimensions in the work by Mayer et al. [60]. *Purpose* is meant to reflect the *benevolence* dimension of the interpersonal framework for trust, and refers to the extent to which there is value congruence between the trustor and the automation [51,52]. Importantly, while these three areas are undoubtedly key factors for a trustor when assessing the trustworthiness of some sort of automation, the categorization privileges the trustee by focusing on characteristics of the automation and the trustors' fidelity to them. Focusing heavily on the trustee runs the risk of characterizing trustworthiness as an objective characteristic of the trustee that is agnostic of the trustors. As we will discuss later, such views of trustworthiness can be insufficient.

Although some scholars have focused primarily on conceptualizing trustworthiness as being related to objective criteria of the automation, we want to highlight a few important perspectives from this literature that have engaged a more nuanced understanding. First, Hoff and Bashir [35] present a robust conceptual framework for trust in automation that engages the performance and purpose

dimensions outlined in the preceding paragraph but that supplements this with factors like situational context and the users' dispositions. Second, Chiou & Lee [7], p. 7 have presented a thoughtful relational framing of trust that emphasizes responsivity, "the degree to which the automation effectively adapts to the person and situation," and the complexities of different relationships within human-automation teams. This framework also advocates moving away from focusing on "calibrating trust," or aligning trustees' perceptions with those of different types of experts, to facilitating processes that enable trusting. We feel these perspectives are invaluable for more effectively conceptualizing trustworthiness, and both have been influential in our reconceptualization of trustworthy AI. But first, in the next section we highlight research that focuses on interpersonal trust but does so in the context of understanding how this trust relates to perceptions of risks, such as of technologies like AI, which has not yet been deeply engaged in this space and offers additional merits to the reconceptualization.

2.3. Risk and trust

The concept of trust has also attracted a great deal of attention from scholars who study risk perceptions and other attitudes and behaviors related to risk and risk management. However, the risk perception literature approaches trust in a way that somewhat blends the foci of the interpersonal and automation literature outlined above by focusing on how trust in people, organizations, or institutions relates to attitudes and behaviors connected to a specific risk (a technology, disease outbreak, natural hazard, etc.), the management of that risk, and information shared about that risk. For example, a risk scholar might study community members' trust in regulators to manage the risks AI poses to their job security, whereas a trust in automation scholar might study how employees trusted a newly deployed AI-powered tool they were expected to use on the job. Although this is a subtle distinction, we feel it is an area that, in some ways, can augment and offer more direct contributions to the emerging trustworthy AI literature than the interpersonal and automation work can on their own.

There are several commonalities between how the risk literature we are highlighting in this subsection conceptualizes trust and how the literature on interpersonal trust reviewed above does. For example, in the risk literature, we still see the trustor's expectations [17], the trustee's intentions and abilities [18], as well as shared values between the trustor and trustee [91] all being discussed as key factors for trust. This line of work has engaged the concepts of cooperation and confidence, which are both deeply intertwined with trust throughout all three areas of literature we review but are especially prominent in the risk literature. The relationships among trust, confidence, and cooperation have attracted a great deal of theoretical focus from risk researchers to the extent that several theoretical models have been developed, evaluated, and debated for many years (e.g., [19,89]). One difference from the other literatures is the way risk scholars have focused heavily on disentangling the conceptualization and measurement of each of these terms, as well as how they relate to one another.

An interesting theme in some of the trust and risk literature is that trust serves as an heuristic (i.e., a mental shortcut) and may be substituted for an analytical approach to assessing the acceptability of a risk, especially in the presence of stress, or absence of specific knowledge or expertise (e.g., [20,89,90,92]). Generally, these studies focus on publics' trust in risk management organizations, professionals, or individuals who have authority related to the risk, and how information from them can influence judgments of the acceptability of technologies and hazards. This is distinct from how trust has been discussed in the context of AI – in the risk literature *the trustees are people related to the technology* whereas in the trustworthy AI literature *the trustee is the technology itself*. The latter perspective mirrors that of the trust in automation literature and the former represents a perspective somewhere in between the interpersonal and automation literatures, which has not been fully engaged in the trustworthy AI literature (for exceptions see e.g., [3,8]).

In the risk and trust frameworks, cooperation is largely defined and applied in similar terms as in the interpersonal trust section (e.g., [20]), but this line of work places a strong emphasis on confidence and has dedicated a great deal of effort to thinking about how the concept relates to and is distinct from trust. However, confidence itself is a complicated and messy topic with seemingly more open theoretical questions than definitive answers [33]. Some risk research generally defines confidence as an expectation or belief about things that will happen in the future, which may be based on experience or evidence of competence (e.g., [18,91]). The authors argue that trust is distinct from confidence because "trust involves risk and vulnerability, but confidence does not; trust is based on social relations, whereas confidence is based on familiarity; the objects of trust are person-like entities, but one can have confidence in just about anything" ([91], p. 706).

There are several ideas embedded in this statement that have potential implications for AI. For example, when we discuss trustworthiness and trust in AI, what risks or vulnerabilities are we inferring or implying that potential users are assuming? Is our goal really increasing or decreasing potential users' trust in AI or do we mean confidence? Are there times we would hope users do *not* trust AI, or even distrust AI, and, if so, how do we convey that effectively? Can a user confidently not trust or distrust AI? Who or what is the 'target' for trust (or confidence) in AI - the models, developers, or both? While digging into these important, specific questions is beyond the scope of this paper, we argue these and related questions that have emerged from our detailed review of past conceptualizations of trust demonstrate the need for a more reflexive and nuanced conceptualization of trustworthiness for AI.

Overall, the trust and risk literature provides some of the key foundations for our conceptualization of trustworthiness in the context of AI. First, the field of risk largely embraces that concepts like risk and trust are context dependent [89] and entail subjective judgements (e.g., [20,31,73]). In other words, the way we perceive risks depends on our values, experiences, and the specific situations we are in. Using genetically modified organisms (GMOs) as an example, a geneticist's risk perceptions might focus on whether or not modified genes will transfer to wild genomes whereas a parent may be concerned about potential health effects from their child consuming foods made with GMO ingredients. For the geneticists, some genetic transfer into the wild may be acceptable, but the parent likely would not find any negative effects for their child to be acceptable. Here we see the importance of (1) the individual's

perceptions and (2) the context in which they are perceiving the risks. These two ideas are the foundation of this reconceptualization of trustworthiness, which we discuss in more detail in the next section.

3. Our reconceptualization of trustworthy AI

The aim of our reconceptualization is to provide a stronger, more consistent foundation for trustworthy AI development moving forward by leveraging important insights from a range of social science literature and elevating the importance of focusing on potential trustors of AI. In this section, we build upon our literature review above to reconceptualize AI trustworthiness from the trustors' perspective, which is necessary for the relational concepts of trust and trustworthiness but is often under-emphasized. We specifically reconceptualize the trustworthiness of AI as both (a) perceptual, meaning that a given trustor assesses whether, when, and to what extent AI is trustworthy and (b) context-dependent, meaning that a trustor's perceived trustworthiness and use of an AI model can vary based on the specifics of their situation.

For our reconceptualization, both *developers* and *users* of AI models are considered as important types of trustors. Note, we define users broadly to include direct and indirect use of AI, but we also acknowledge that those who are affected by others' use of AI can be important potential trustors. Some considerations may be more relevant for how developers assess the trustworthiness of and trust in their own work as they build a model, whereas other considerations may be more important for users of the consequent AI system. However, as we discuss in the following sections, the considerations of these two broad groups are not mutually exclusive.

3.1. *The trustworthiness of AI is perceptual – the result of a subjective evaluation*

In some of the trustworthy AI literature, trustworthiness of an AI model is treated as an inherent, objective characteristic (Fig. 2B). For example, although the Jacovi et al. [41] framework acknowledges there is a perceptual element of the human-AI trust relationship (which they call 'anticipation'), they present this as different from the trustworthiness of a given model. They define AI model trustworthiness as whether a model is capable of maintaining explicit contracts, which seems to be treated as objectively determined. They present the role of perception as the human trusting the AI model to uphold the contract, which they then classify as either 'warranted' or 'unwarranted' if 'caused' by the seemingly objectively determined AI model trustworthiness.

Conceptualizing trustworthiness as purely objective implies that it can be entirely engineered or designed a priori, regardless of context, user, or evaluator. Moreover, deeming developers' criteria for assessing model trustworthiness as objective fails to recognize the subjectivity in choices about what metrics to use, how, and why they (ought to) matter as well as other human biases in model development [62]. These can lead to top-down and potentially normative views of trustworthiness, whereas the theoretical and empirical work on trust and trustworthiness outlined in the previous sections suggests the need for engaging with bottom-up and trustor-centered perspectives.

We therefore offer that the trustworthiness of AI for users should be conceptualized in a way that centers the trustor and is the result of their subjective evaluation – i.e., is 'perceptual' (Fig. 2A). In other words, AI model trustworthiness is in the eye of the beholder - a person assesses trustworthiness for themselves. As Fig. 2 demonstrates, our reconceptualization distinguishes trustworthiness from model traits (B) and positions it as the result of a subjective evaluation made by an individual (A). Although the user's evaluation may depend to some degree on the AI model characteristics as created by the developers, these two are not necessarily equivalent.

This idea builds on theoretical work from across the interpersonal, trust-in-automation, and risk and trust literatures. For example, classic work on interpersonal trust specifically positions the trustor in control of decision making and evaluations related to trust [46, 100]. Relatedly, Mayer et al. [60] argued that an individual's "perception and interpretation of the context of the relationship will affect both the need for trust and the evaluation of trustworthiness" (page 727). We see similar arguments from the trust-in-automation literature for the need to: "explicitly recognize that trust is based on the perceived qualities of another and is therefore subject to all the vagaries of individual interpretation. Thus, the perceived properties, which support an expectation may be quite independent of a

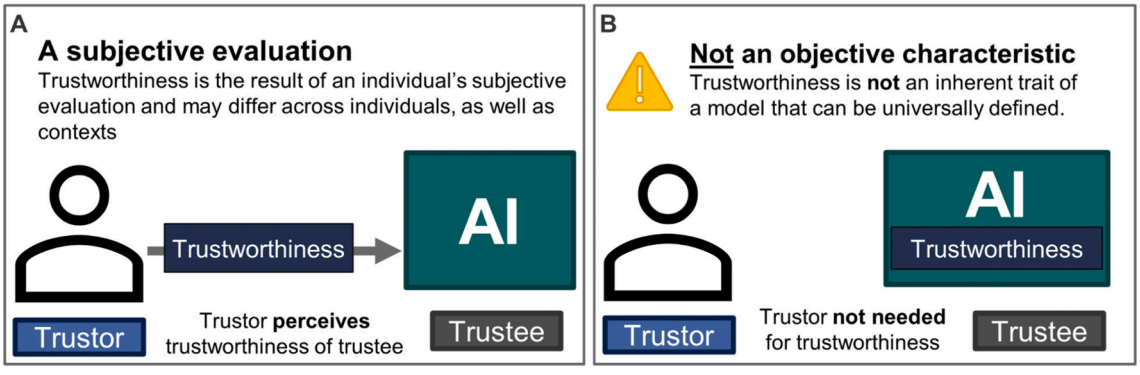


Fig. 2. Conceptual representation to illustrate that AI trustworthiness is perceptual (the result of a subjective evaluation), not an objective characteristic.

referent's actual properties, which are referred to as the referent's trustworthiness" ([67], p. 531). Our reconceptualization here also builds on theoretical work from the field of risk, as discussed earlier. Risk perceptions are characterized with both objective and subjective dimensions to capture both the technical and value-ladened components, respectively [31]. The subjective dimensions are often extremely important for understanding individuals' behaviors about a risk [94,95]. We offer that this also applies to AI, whereby use of AI depends on users' subjective assessments of its trustworthiness.

The problem with not appreciating the inherent subjectivity and perceptuality of trustworthiness in the context of AI is that it facilitates focusing well-meaning energy on efforts that are not necessarily going to yield the intended results. For example, efforts to develop benchmarks or standards for AI to pass to be considered "trustworthy" often treat trustworthiness as a trait that can be universally defined and standardized across all potential evaluators. Such validation efforts are a necessary foundation, but they are not sufficient for accomplishing the goal of developing AI that is actually deemed trustworthy by intended users whose uses, values, and perceptions –which may or may not overlap with developers'–are valid and must be considered. Appreciating this subjectivity and perceptuality and working with intended trustors or users to better understand their perspectives will yield better results in developing trust in the AI.

Some work emphasizes that AI models with sufficiently high performance or verification metrics are trustworthy and will (or should) be trusted whereas models with lower performance will (or ought) not be. Although performance likely has implications for potential trustors like researchers', developers', and users' perceived trustworthiness [3,4], *equating* trustworthiness with performance is problematic, particularly for users. This is because it fails to consider the users' perceptions of how the initial goal of the AI was framed, of the data collection and processing, of the decisions made by the research team in the development phase, or of how the model output is shared and communicated with the user. Any or all of these factors–or other factors not listed here–may impact the perceived trustworthiness of the AI. Furthermore, models that perform well may be biased, unethical, and/or untrustworthy for many different reasons [64]. Some scholars have argued similar points, namely that we should take a broader approach to understanding trustworthy AI [48,98], while others have gone further and demonstrated the importance of predictors beyond performance when examining trust in AI [44]. Given the complexity and multidimensional nature of AI trustworthiness, we assert that although model performance may be part of a user's evaluation, the trustworthiness of AI involves more than just performance.

In the end, efforts to characterize and develop trustworthiness need to explicitly involve both trustors of the AI and the AI as the trustee. Conversely, we should reject efforts that only consider AI and aim to prescribe any level of trustworthiness to AI without also considering a trustor or group of trustors. This is essential because although there are likely factors some groups will agree are important for trustworthiness, there are no attributes or characteristics of AI that make it universally worthy of trust. The relational nature of trust and trustworthiness mean that any given model can simultaneously be deemed trustworthy by some trustors and untrustworthy by others. For example, an AI model to predict coastal sea level rise in a future climate may be deemed trustworthy by a group of climate scientists because it follows current best practices and meets accepted performance thresholds for the field. However, a city emergency manager and lobbyist might both not see the same model as trustworthy for different reasons. The emergency manager may find the resolution of the model is too coarse to help them make their key decisions and the lobbyist may be worried about the developers' views influencing how the model was developed. All three of these perceptions of trustworthiness can exist at the same time and none of them change the underlying model's trustworthiness because trustworthiness exists relationally between the trustors and the trustee. The main implication here is that development and management efforts must understand and work with perceptions of potential trustors to develop AI that is perceived to be trustworthy.

The idea that the trustworthiness of AI is not something that can be fully engineered or guaranteed a priori seems to be gaining some traction in the last few years. Lewis and Marsh [56] argue that decisions related to trust depend on many factors and we should work to provide evidence for these decisions rather than attempting to force or coerce others to conform with our own trust decisions. They refer to this as facilitating trust empowerment rather than trust enforcement [14,15]. Similarly, Ashoori & Weisz [1] provide empirical evidence that beautifully demonstrates how individual differences in perceptions of AI make it impossible to prescribe or guarantee trust in AI systems, despite past research demonstrating the ability to predict that trust will increase under some circumstances.

3.2. The trustworthiness of AI is context-dependent

The fundamental importance of context is well-represented across the many different lines of trust and trustworthiness research. One important example of how central context has been for conceptual work on trust and trustworthiness comes from the Mayer et al. [[60], pp. 726–7] paper that is often cited in trustworthy AI work: "the specific consequences of trust will be determined by contextual factors such as the stakes involved, the balance of power in the relationship, the perception of the level of risk, and the alternatives available to the trustor." Although the Mayer et al. definition of trust is widely cited in the trustworthy AI literature, this additional recognition of context, much less an emphasis of its importance, is not often cited and thus seemingly is not widely considered. Furthermore, Mayer et al. caution that their model was developed for organizational relationships and that "its propositions may not generalize to relationships in other contexts" (page 730).

We see this strong emphasis on the context-dependency of trust and the direct caution of generalizing conceptual work across contexts as a clear indication of how crucial context is for trust. Similar ideas are expressed in the trust-in-automation literature. For example, Lee and See [[51], p. 56] argue "Trust does not develop in a vacuum but, instead, evolves in a complex individual, cultural, and organizational context. [...] Understanding trust requires a careful consideration of the individual, organizational, and cultural context." Relatedly, Hoff and Bashir [[35], p. 413] conceptualize situational trust as dependent on "the specific context of an interaction." Others have demonstrated the importance of context for trust-in-automation and how it relates to the dynamic nature of trust

[36,37]. We see this emphasis on the importance of context for trust-in-automation carried through to more recent conceptualizations, like that of Chiou & Lee [7]. Taken together, this literature demonstrates how central context is, and should be, for understanding the trustworthiness of a considered AI system.

A good guiding question for understanding the context-dependent nature of AI trustworthiness is "worthy of trust by whom, for what, and when?" Who exactly are we expecting to evaluate trustworthiness, are we expecting these potential users to trust the accuracy of information a model is providing in one particular instance, or are we asking users to always trust the decisions made by the model? Even these simple examples demonstrate how generic and acontextual the term 'trustworthy AI' can be and how context is essential for making meaning of it. The interpersonal trust literature demonstrated the multiplexity of trust - you can trust some people in some situations but not trust the same person in other situations or for other tasks [53].

In Fig. 3, we demonstrate some of the different forms context can take in the AI space. We specifically highlight the *development, use,* and *policy* contexts as three domains that are interacting, mutually influential processes that influence perceptions of trustworthiness. The *development context* covers things like the creation of fundamental AI algorithms or systems (e.g., image classifiers) and the application of AI to a specific problem (e.g., to predict severe weather). These development processes involve many of the technical components around how the AI was designed, built, trained, tested, and so forth. In addition, there are key factors related to the *use context,* which covers the purpose or goal of the model, how the model is operationalized and communicated, what the model is actually doing, whether the AI is intended to make decisions for a user or to provide information that will guide the user's decision making, how the AI model relates to tools or systems the user has worked with in the past, and what the potential implications are of using the AI. Efforts that aim to develop 'trustworthy AI' that only focus on the development context run the risk of not being broadly and fully successful. Such efforts should also consider and integrate considerations of the use context, to the extent possible, knowing that doing so will likely increase users' perceptions of the AI's trustworthiness but still not guarantee it. Lastly, the *policy context* adds another layer because it focuses on the governance of dimensions of both the development and use contexts. The policy context can deal with how to set recommendations, leading practices, standards, benchmarks, and regulations for how to develop AI, but also for how and when AI should and should not be used.

These three contexts are interrelated and influence one another, especially given the interdependence and iterative nature of the AI development, its intended uses, and the policies used to govern it. These interdependencies and interactions are represented in Fig. 3. One example is the lack of standards for distinguishing between hard and soft constraints for models. The lack of standards or requirements (policy context) for developing such mechanisms (development context) leads to problems when models that are not constrained to providing real citations or resources or conserving mass are then relied on for decision making (use context). Furthermore, not all AI models are equal or the same, so these contextual considerations will also vary based on the specific model and how it was developed. The importance of context is also seen throughout the trust and trustworthiness literature reviewed in Section 2.

The automation literature illustrates how the specific uses or applications of a technology matter for how people perceive it and interact with it. We see this largely with how the user's role is defined in relation to the technology. Parasuraman and Riley [74] discussed how a user's control of automation has many levels along a continuum, rather than there being a sharp dichotomy between user-controlled and automation-controlled systems. We can then pair this with the importance of decisional freedom, the level of autonomy a user has over how automation is used within their operational context, which Hoff and Bashir [35] argued may lead to higher levels of trust. Together, we see the importance of the decision and use contexts for assessing trustworthiness. When we apply these ideas to AI, there are similar dimensions that may be relevant for understanding how users perceive the trustworthiness of AI, which we have begun to outline in Fig. 3. While empirical investigation is needed to explore this area more, there are likely differences

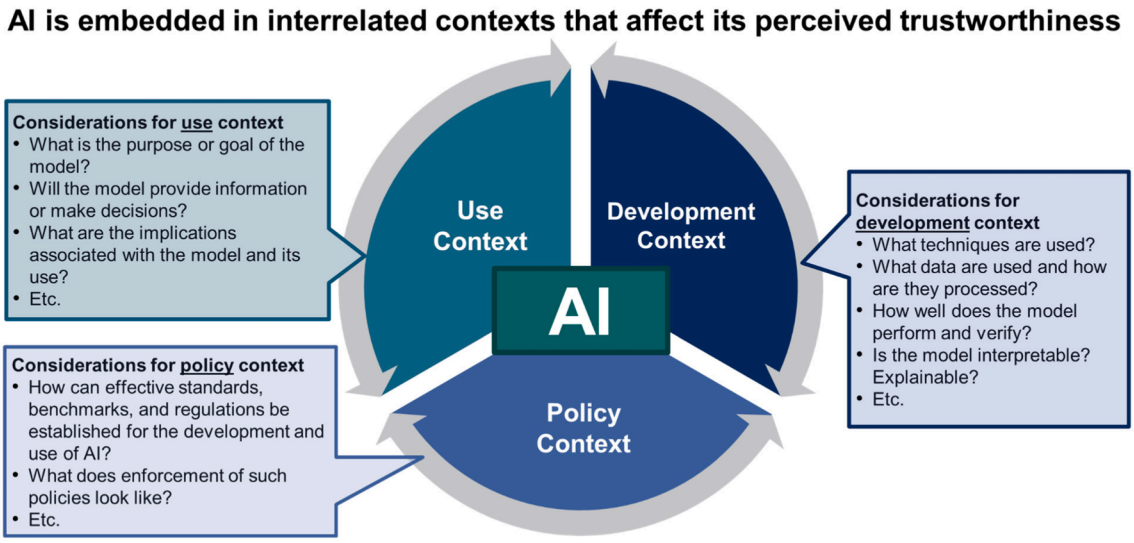


Fig. 3. Conceptual representation to illustrate that AI trustworthiness depends on use, development, and policy contexts, which are interrelated.

in how trustworthiness is perceived in these different contexts.

Relatedly, Ashoori and Weisz [1] found that decision stakes were a key factor in the perceived acceptability of using AI, with AI being more acceptable for lower stakes decisions than higher stakes ones. These findings are consistent with those from the automation literature, which have found task complexity [74] and the risks associated with the task [35] are important for use and reliance decisions. Overall, the users' decision context and how the AI integrates into that context are likely crucial factors for how users perceive the trustworthiness of AI.

Putting these points together into an example, you might perceive a colleague to be trustworthy enough to lead the development of an AI model for a project, while simultaneously perceiving the same colleague undeserving of your trust for writing their sections of the annual report on time for that same project. In this example, the trustworthiness of your colleague depends on if you are assessing it for model development or report writing; there are different answers and relevant factors for the assessment, despite being about the same person for the same project. Such context-dependence of trustworthiness also is true for AI. As Lewis and Marsh [[56], p. 34] put it, "Asking if you trust AI is akin to asking if you trust the human race: perhaps a nice headline but in any event, meaningless. And, even if we answer 'no', what does that actually mean, given the situation we are already in?" This statement is well-supported by empirical and conceptual work from the trust-in-automation literature, as outlined above.

Appreciating the context-dependent nature of trustworthiness, especially when conducting empirical research, is essential to better understand whether, when, and why observations related to trustworthiness generalize from context to context and application to application. For example, the trustworthiness of a recommendation algorithm predicting which journal articles you may be interested in reading is very different from that of a model designed to guide forecasters' predictions of a tornado path, because there are substantially different implications of using such algorithms. This example highlights how adding nuance to the applications, contexts, and decision-making surrounding specific applications can strengthen efforts to develop AI that is trustworthy, trusted, and used.

Fortunately, there are a few examples of empirical and theoretical work that point to the importance of context and specific applications in examinations of trustworthy AI. First, Glikson & Woolley, [30] have added nuance to 'AI' by breaking down their examination into more specific subcategories of AI (robotic, virtual, and embedded). Even this high-level distinction demonstrated differences with respect to trust and trustworthiness among the different application types. Second, the decision context [1] and task characteristics [30,38] associated with AI have also gained attention from researchers as important points for understanding the trustworthiness of AI. Such work further underscores the importance of engaging a context- and application-dependent conceptualization of trustworthy AI.

4. Conclusion

We seem to be at a pivotal point in the advancement and proliferation of AI, which has prompted a wide-spread motivation to develop AI that can be trusted and to develop people's trust in AI. As we have demonstrated, trust and trustworthiness are complex, relational concepts that cannot be completely guaranteed nor built entirely into an AI system. This is because AI trustworthiness must also consider the trustor and is both (a) perceptual, meaning that a user assesses whether, when, and to what extent AI model output is trustworthy, even if it has been developed in adherence to AI trustworthiness standards or policies, and (b) context-dependent, meaning that a user's perceived trustworthiness and use of an AI model can vary based on the specifics of their situation.

For research and policy efforts to meet their goals of developing trust in AI and AI worthy of trust, they must appreciate these points. Then, researchers and developers will need to work with and learn from potential trustors to develop AI that they both perceive to be trustworthy in their respective contexts. This also comes with the appreciation that researchers, developers, policymakers, users, and the public all have their own perceptions within their own contexts, which may or may not overlap with one another but are

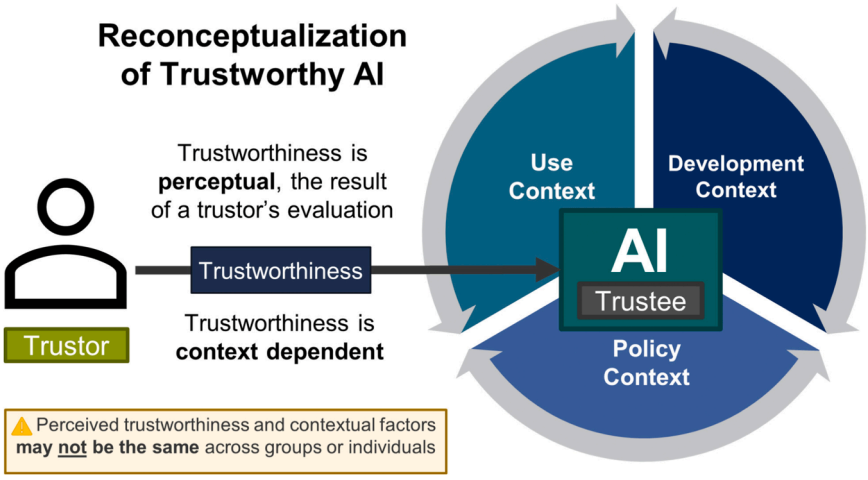


Fig. 4. Conceptual representation to illustrate that trustworthiness is perceptual (the result of a subjective evaluation) and context-dependent.

nonetheless valid. Through this process our reconceptualization will increase the likelihood that efforts to develop trustworthy AI will not just be academic exercises but will be truly successful.

We summarize our reconceptualization of trustworthy AI into one conceptual diagram (Fig. 4). The figure highlights the importance of context and individuals' perceptions for understanding and communicating about the trustworthiness of AI. We also note that each AI model or system will likely have many users and affected communities who will each have their own perceptions and contexts that may differ from others'. Individuals and groups may perceive the same model very differently based on the factors that are important to them, their backgrounds, and their values. Furthermore, the same person may also view the same model differently based on the context in which it is being applied.

Our reconceptualization carries practical implications for how to approach trustworthy AI moving forward, especially with respect to policy making and governance. We highlight three opportunity areas: (1) developing governance frameworks that focus on regulating processes, rather than performance, (2) a stronger focus on user engagement and co-development, and (3) increased support for interdisciplinary and convergence research.

First, our reconceptualization of trustworthy AI supports a need for more process-focused, rather than performance-focused, policies and governance frameworks. The general approach to governing the trustworthiness of AI has emphasized establishing different principles AI must adhere to or criteria it must meet in order to be deemed trustworthy. Such approaches mirror performance policies, in which the policies dictate specific outcomes that must be met but do not enforce *how* these outcomes must be met. Although effective for some regulatory scenarios, performance-focused approaches "depend on the ability of government agencies to specify, measure, and monitor performance" [[10] p. 562]. Specifying, much less measuring and monitoring, a comprehensive and generalizable performance threshold for the trustworthiness of AI is not feasible given it is a relational concept that depends on the potential trustor and the context in which it is being applied, and attempts to do so can conflate trustworthiness for other dimensions of AI (e.g., [50]). For policies and governance frameworks specifically aiming to address trust and trustworthiness, the focus should instead be on processes and practices that embrace the dynamic, perceptual, and context-dependent nature of trust. In other words, instead of attempting to define universal, timeless standards for trustworthiness, efforts should focus on identifying the needs, perceptions, and contexts of specific trustors or groups of trustors and fostering practices that work to facilitate trust among them. One such way to do this is through a stronger emphasis on the second and third opportunity areas we offer, discussed next.

Second, our reconceptualization demonstrates the need to elevate the perspective and importance of the trustor in efforts to develop trustworthy AI. Although the development characteristics and performance of AI as the trustee are incredibly important, efforts focusing on them alone are not likely to create AI systems that are actually widely trusted. Both larger-scale efforts (e.g., standards, regulations) and smaller-scale efforts (e.g., developer-user interactions, transitioning research models to operations) will likely fall short of yielding the desired outcomes if they do not focus on the intended trustors by incorporating the understanding that trustworthiness is in the eye of the beholder and depends on many contextual factors, which are dynamic and change over time. This notion is aligned with existing movements in the field, specifically the notion of human-centered AI systems (e.g., [87,88]). However, there is still not nearly enough of an investment in and focus on meaningfully engaging with a wide range of potential trustors.

Individuals, regardless of their expertise, assess the trustworthiness of an AI system for themselves and weigh a variety of different factors when doing so. As all scientists do, developers make subjective decisions, both consciously and subconsciously, when developing and designing models – from the data used and how it is processed to the model architecture and training approach. These decisions stem from how trustworthy they perceive the different pieces to be (e.g., datasets, packages, platforms), as well as by what they assume others will perceive to be trustworthy when they evaluate the resulting AI system as a whole. So although developers likely emphasize different points, they too are assessing trustworthiness in perceptual and context dependent ways - just as users do. The goals and perceptions of both developers and users are valid, but they may not align, for good reason. Experts in this space can learn from the field of risk analysis that users have valid perceptions that are value- and experience-laden and context-dependent, which may not align with experts' technical assessments. Moreover, aligning or 'calibrating' these perspectives is not necessarily the goal because these concepts imply one or both groups need to change. Instead, it is important to understand, appreciate, and work with both sets of perspectives. Rather than trying to change users, the community must understand the relevant context and perceptions for trustworthiness by working with users throughout the development process and beyond to better meet their needs.

Trustworthy AI policy and governance frameworks need to integrate users' needs and perspectives in substantial ways; they cannot simply be mentioned as important or left as a suggestion. This can be done in several ways, such as co-development, various forms of engagement, or research focused on advancing our understanding of users. Investments must be made in conducting rigorous empirical work with both developers and users to understand their needs, goals, perceptions, and contexts. Such work is crucial for developing more effective and generalizable solutions and leading practices. Time and resource investments can be more fully realized if they are coupled and co-developed with user-centered perspectives. This will facilitate co-developing trustworthy AI through relationship building and engaging the users early and often in the process - from problem identification throughout the AI's life cycle [3].

Lastly, our reconceptualization points to the need to increase support for interdisciplinary and convergence research relating to AI development and governance. The community needs to invest time and resources in fostering research teams that are able to address the increasingly complex challenges that AI creates and may help address by meaningfully integrating expertise and experience across computer science, social science, and relevant domain sciences [3,63]. Part of this investment needs to be focused on providing the infrastructure and time for these teams to form, learn from, and appreciate one another, and then identify the pressing needs and questions that transcend each of their individual domains.

In the end, such interdisciplinary research and user engagement will not produce magic, cure-all solutions but will instead help us understand the essential complexities and nuances that become hidden by disciplinary boundaries and unchecked assumptions. Only by deeply integrating across a range of disciplines, perspectives, and experiences can these hidden areas be revealed and the most

pressing questions be actualized.

CRediT authorship contribution statement

Christopher D. Wirz: Writing – review & editing, Writing – original draft, Visualization, Project administration, Investigation, Conceptualization. **Julie L. Demuth:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Conceptualization. **Ann Bostrom:** Writing – review & editing, Funding acquisition. **Mariana G. Cains:** Writing – review & editing, Visualization. **Imme Ebert-Uphoff:** Writing – review & editing, Funding acquisition. **David John Gagne:** Writing – review & editing, Funding acquisition. **Andrea Schumacher:** Writing – review & editing. **Amy McGovern:** Writing – review & editing, Funding acquisition. **Deianna Madlambayan:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to acknowledge the support and insights from all of our colleagues from the NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES). This material is based upon work supported by the National Science Foundation [grant number RISE-2019758]. This material is based upon work supported by the NSF National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977.

Data availability

No data was used for the research described in the article.

References

[1] Ashoori M., & Weisz J.D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/1912.02675.
[2] Bauer P.C. (2019). Clearing the jungle: conceptualizing trust and trustworthiness. 10.2139/ssrn.2325989.
[3] , 2nd A. Bostrom, J.L. Demuth, C.D. Wirz, M.G. Cains, A. Schumacher, D. Madlambayan, A.S. Bansal, A. Bearth, R. Chase, K.M. Crosman, I. Ebert-Uphoff, D. J. Gagne, S. Guikema, R. Hoffman, B.B. Johnson, C. Kumler-Bonfanti, J.D. Lee, A. Lowe, A. McGovern, J.K. Williams, Trust and trustworthy artificial intelligence: a research agenda for AI in the environmental sciences, Risk Anal. Off. Publ. Soc. Risk Anal. (2024), https://doi.org/10.1111/risa.14245.
[4] M.G. Cains, C.D. Wirz, J.L. Demuth, A. Bostrom, D.J. Gagne, A. McGovern, D. Madlambayan, Exploring NWS forecasters' assessment of AI guidance trustworthiness, Weather Forecast. 39 (8) (2024) 1219–1241.
[5] A. Campagner, R. Angius, F. Cabitza, A question of trust: old and new metrics for the reliable assessment of trustworthy AI, in: Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies. 16th International Conference on Health Informatics, Lisbon, Portugal, 2023, https://doi.org/10.5220/0011679600003414.
[6] Chase R.J., Harrison D.R., Burke A., Lackmann G.M., & McGovern A. (2022). A machine learning tutorial for operational meteorology, part I: traditional machine learning. In arXiv [physics.ao-ph]. arXiv. http://arxiv.org/abs/2204.07492.
[7] E.K. Chiou, J.D. Lee, Trusting automation: designing for responsivity and resilience, Hum. Factors (2021) 187208211009995, https://doi.org/10.1177/00187208211009995.
[8] H. Cho, R. Hooi, Risk perceptions and trust mechanisms related to everyday AI. Research Handbook on Artificial Intelligence and Communication, Edward Elgar Publishing, 2023, pp. 163–175, https://doi.org/10.4337/9781803920306.00019.
[9] H. Choung, P. David, A. Ross, Trust and Ethics in AI, AI & Society, 2022, https://doi.org/10.1007/s00146-022-01473-4.
[10] C. Coglianese, The limits of performance-based regulation, Univ. Mich. J. Law Reform 50 (2016) 525.
[11] M. Deutsch, Trust and suspicion, J. Confl. Resolut. 2 (4) (1958) 265–279, https://doi.org/10.1177/002200275800200401.
[12] M. Deutsch, The effect of motivational orientation upon trust and suspicion, Hum. Relat. 13 (2) (1960) 123–139, https://doi.org/10.1177/001872676001300202.
[13] Deutsch M. (1962). Cooperation and trust: some theoretical notes. Nebr Symp Motiv, 330, 275–320. https://psycnet.apa.org/fulltext/1964-01869-002.pdf.
[14] N. Dwyer, Traces of Digital trust: an Interactive Design Perspective, [Phd, Victoria University], 2011. https://vuir.vu.edu.au/17663/.
[15] N. Dwyer, A. Basu, S. Marsh, Reflections on measuring the trust empowerment potential of a digital environment, Trust Manag. VII (2013) 127–135, https://doi.org/10.1007/978-3-642-38323-6_9.
[16] M.T. Dzindolet, L.G. Pierce, H.P. Beck, L.A. Dawe, The perceived utility of human and automated aids in a visual detection task, Hum. Factors 44 (1) (2002) 79–94, https://doi.org/10.1518/0018720024494856.
[17] T.C. Earle, Thinking aloud about trust: a protocol analysis of trust in risk management, Risk Anal. Off. Publ. Soc. Risk Anal. 24 (1) (2004) 169–183, https://doi.org/10.1111/j.0272-4332.2004.00420.x.
[18] T.C. Earle, Trust in risk management: a model-based review of empirical research, Risk Anal. Off. Publ. Soc. Risk Anal. 30 (4) (2010) 541–574, https://doi.org/10.1111/j.1539-6924.2010.01398.x.
[19] T.C. Earle, G. Cvetkovich, Social Trust: Toward a Cosmopolitan Society, Greenwood Publishing Group, 1995. https://play.google.com/store/books/details?id=z1khILCNxiwC.
[20] T. Earle, M. Siegrist, Trust, confidence and cooperation model: a framework for understanding the relation between trust and risk perception, Int. J. Glob. Environ. Issues 8 (1–2) (2008) 17–29, https://doi.org/10.1504/IJGENVI.2008.017257.
[21] C.Z. Elgin, Trustworthiness, Philos. Pap. 37 (3) (2008) 371–387, https://doi.org/10.1080/05568640809485227.
[22] E. Parliament. (2024, March 13). Artificial Intelligence Act: mEPs adopt landmark law. European Parliament News. https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law.
[23] Executive Order 13859. (2020). Executive Order on promoting the use of trustworthy artificial intelligence in the Federal government. https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/.

[24] Executive Order 14110. (2023). Executive Order on promoting the use of trustworthy artificial intelligence in the Federal government. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[25] M. Frederiksen, Relational trust: outline of a Bourdieusian theory of interpersonal trust, J. Trust Res. 4 (2) (2014) 167–192, https://doi.org/10.1080/21515581.2014.966829.

[26] O. Freiman, Making sense of the conceptual nonsense "trustworthy AI, AI Ethics (2022), https://doi.org/10.1007/s43681-022-00241-w.

[27] GAO, Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities (No. GAO-21-519SP), U.S. Government Accountability Office, 2021. https://www.gao.gov/products/gao-21-519sp.

[28] GAO, Artificial Intelligence: Agencies Have Begun Implementation But Need to Complete Key Requirements, United States Government Accountability Office, 2023. https://www.gao.gov/assets/d24105980.pdf.

[29] K. Giffin, The contribution of studies of source credibility to a theory of interpersonal trust in the communication process, Psychol. Bull. 68 (2) (1967) 104–120, https://doi.org/10.1037/h0024833.

[30] E. Glikson, A.W. Woolley, Human trust in artificial intelligence: review of empirical research, Annals 14 (2) (2020) 627–660, https://doi.org/10.5465/annals.2018.0057.

[31] S.O. Hansson, Risk: objective or subjective, facts or values, J. Risk Res. 13 (2) (2010) 231–238.

[32] W.D. Heaven, The inside story of how ChatGPT was built from the people who made it, MIT Technol. Rev. (2023). https://www.technologyreview.com/2023/03/03/1069311/inside-story-oral-history-how-chatgpt-built-openai/.

[33] J. Henderson, J. Spinney, J.L. Demuth, Conceptualizing confidence: a multi-sited qualitative analysis in a severe weather context, Bull. Am. Meteorol. Soc. (2022), https://doi.org/10.1175/BAMS-D-22-0137.1. , -1(aop).

[34] HLEG, Ethics guidelines for trustworthy AI. High-level Expert Group On Artificial Intelligence, European Commission, 2019.

[35] K.A. Hoff, M. Bashir, Trust in automation: integrating empirical evidence on factors that influence trust, Hum. Factors 57 (3) (2015) 407–434, https://doi.org/10.1177/0018720814547570.

[36] R. Hoffman, A taxonomy of emergent trusting in the Human–Machine relationship, Ed., in: P.J. Smith, R.R. Hoffman (Eds.), Cognitive Systems Engineering: The Future For a Changing World, CRC Press, 2017, pp. 137–163.

[37] R.R. Hoffman, M. Johnson, J.M. Bradshaw, A. Underbrink, Trust in automation, IEEE Intell. Syst. 28 (1) (2013) 84–88, https://doi.org/10.1109/MIS.2013.24.

[38] Hoffman R.R., Mueller S.T., Klein G., & Litman J. (2018). Metrics for explainable AI: challenges and prospects. In arXiv [cs.AI]. arXiv. http://arxiv.org/abs/1812.04608.

[39] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J.D. Ser, W. Samek, I. Jurisica, N. Díaz-Rodríguez, Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Inf. Fusion 79 (2022) 263–278, https://doi.org/10.1016/j.inffus.2021.10.007.

[40] Hovland C.I., Janis I.L., & Kelley H.H. (1953). Communication and persuasion. https://psycnet.apa.org/fulltext/1953-15071-000.pdf.

[41] Jacovi A., Marasović A., Miller T., & Goldberg Y. (2021). Formalizing trust in artificial intelligence: prerequisites, causes and goals of Human trust in AI. *arXiv: 2010.07487* [cs]. http://arxiv.org/abs/2010.07487.

[42] J.Y. Jian, A.M. Bisantz, C.G. Drury, Foundations for an empirically determined scale of trust in automated systems, Int. J. Cogn. Ergon 4 (1) (2000) 53–71.

[43] K. Jones, Trustworthiness, Ethics 123 (1) (2012) 61–85, https://doi.org/10.1086/667838.

[44] A.D. Kaplan, T.T. Kessler, J.C. Brill, P.A. Hancock, Trust in artificial intelligence: meta-analytic findings, Hum. Factors (2021) 187208211013988, https://doi.org/10.1177/00187208211013988.

[45] D. Kaur, S. Uslu, K.J. Rittichier, A. Durresi, Trustworthy artificial Intelligence: a review, ACM Comput. Surv. 55 (2) (2022) 1–38, https://doi.org/10.1145/3491209.

[46] H.W. Kee, R.E. Knox, Conceptual and methodological considerations in the study of trust and suspicion, J. Confl. Resolut. 14 (3) (1970) 357–366, https://doi.org/10.1177/002200277001400307.

[47] M. Kim, H. Sohn, S. Choi, S. Kim, Requirements for trustworthy artificial intelligence and its application in healthcare, Healthc. Inform. Res. 29 (4) (2023) 315–322, https://doi.org/10.4258/hir.2023.29.4.315.

[48] B. Knowles, J.T. Richards, The sanction of authority: promoting public trust in AI, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 262–271, https://doi.org/10.1145/3442188.3445890.

[49] N.K. Lankton, D. Harrison McKnight, J. Tripp, Technology, humanness, and trust: rethinking trust in technology, J. Assoc. Inf. Syst. 16 (10) (2015) 1, https://doi.org/10.17705/1jais.00411.

[50] J. Laux, S. Wachter, B. Mittelstadt, Trustworthy artificial intelligence and the European Union AI act: on the conflation of trustworthiness and acceptability of risk, Regul. Gov. 18 (1) (2024) 3–32.

[51] J.D. Lee, K.A. See, Trust in automation: designing for appropriate reliance, Hum. Factors 46 (1) (2004) 50–80, https://doi.org/10.1518/hfes.46.1.50_30392.

[52] J. Lee, N. Moray, Trust, control strategies and allocation of function in human-machine systems, Ergonomics 35 (10) (1992) 1243–1270, https://doi.org/10.1080/00140139208967392.

[53] R.J. Lewicki, D.J. McAllister, R.J. Bies, Trust and distrust: new relationships and realities, AMRO 23 (3) (1998) 438–458, https://doi.org/10.5465/amr.1998.926620.

[54] J.D. Lewis, A. Weigert, Trust as a social reality, Soc. Forces 63 (4) (1985) 967–985, https://doi.org/10.1093/sf/63.4.967.

[55] J.D. Lewis, A.J. Weigert, The Social Dynamics of Trust: theoretical and empirical research, 1985-2012, Soc. Forces 91 (1) (2012) 25–31, https://doi.org/10.1093/sf/sos116.

[56] P.R. Lewis, S. Marsh, What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence, Cogn. Syst. Res. 72 (2022) 33–49, https://doi.org/10.1016/j.cogsys.2021.11.001.

[57] L. Li, T. Lassiter, J. Oh, M.K. Lee, Algorithmic hiring in practice: recruiter and HR professional's perspectives on AI use in hiring, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 166–176, https://doi.org/10.1145/3461702.3462531.

[58] J.L. Loomis, Communication, the development of trust, and cooperative behavior, Hum. Relat. 12 (4) (1959) 305–315, https://doi.org/10.1177/001872675901200402.

[59] M.A. Malek, Criminal courts' artificial intelligence: the way it reinforces bias and discrimination, AI Ethics 2 (1) (2022) 233–245, https://doi.org/10.1007/s43681-022-00137-9.

[60] R.C. Mayer, J.H. Davis, F.D. Schoorman, An integrative model of organizational trust, Acad. Manag. Rev. 20 (3) (1995) 709, https://doi.org/10.2307/258792.

[61] D.J. McAllister, Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations, Acad. Manag. J. 38 (1) (1995) 24–59, https://doi.org/10.5465/256727.

[62] A. McGovern, A. Bostrom, M. McGraw, R.J. Chase, D.J. Gagne, I. Ebert-Uphoff, K.D. Musgrave, A. Schumacher, Identifying and categorizing bias in AI/ML for earth sciences, Bull. Am. Meteorol. Soc. 105 (3) (2024) E567–E583, https://doi.org/10.1175/BAMS-D-23-0196.1.

[63] A. McGovern, J. Demuth, A. Bostrom, C.D. Wirz, P.E. Tissot, M.G. Cains, K.D. Musgrave, The value of convergence research for developing trustworthy AI for weather, climate, and ocean hazards, npj Nat. Hazards 1 (1) (2024) 13.

[64] A. McGovern, I. Ebert-Uphoff, D.J. Gagne, A. Bostrom, Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science, Environ. Data Sci. 1 (2022), https://doi.org/10.1017/eds.2022.5.

[65] D.H. Mcknight, M. Carter, J.B. Thatcher, P.F. Clay, Trust in a specific technology: an investigation of its components and measures, ACM Trans. Manag. Inf. Syst. 2 (2) (2011) 1–25, https://doi.org/10.1145/1985347.1985353.

[66] G.D. Mellinger, Interpersonal trust as a factor in communication, J. Abnorm. Psychol. 52 (3) (1956) 304–309, https://doi.org/10.1037/h0048100.

[67] B.M. Muir, Trust between humans and machines, and the design of decision aids, Int. J. Man Mach. Stud. 27 (5) (1987) 527–539, https://doi.org/10.1016/S0020-7373(87)80013-5.

[68] M. Muro, R. Maxim, J. Whiton, Automation and Artificial Intelligence: How Machines Are Affecting People and Places, Asian Development Bank Institute, 2019. http://hdl.handle.net/11540/9686.
[69] NASEM, Human-AI Teaming: State of the Art and Research Needs, The National Academies Press, 2022, https://doi.org/10.17226/26355.
[70] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), National Institute of Standards and Technology, 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf?isid=enterprisehub_us&ikw=enterprisehub_us_lead%2Fhow-to-responsibly-use-ai-powered-hr-tools_textlink_https%3A%2F%2Fnvlpubs.nist.gov%2Fnistpubs%2Fai%2FNIST.AI.100-1.pdf.
[71] OECD, Recommendation of the Council on Artificial Intelligence, Organisation for Economic Co-operation and Development (OECD), 2019. No. OECD/LEGAL/0449, https://legalinstruments.oecd.org/api/print?ids=648&lang=en.
[72] OSTP, The Blueprint for an AI Bill of Rights: Making Automated Systems Work For the American People, The Office of Science and Technology Policy, 2022. https://www.whitehouse.gov/ostp/ai-bill-of-rights.
[73] H. Otway, K. Thomas, Reflections on risk perception and policy, Risk Anal. Off. Publ. Soc. Risk Anal. 2 (2) (1982) 69–82.
[74] R. Parasuraman, V. Riley, Humans and automation: use, misuse, disuse, abuse, Hum. Factors 39 (2) (1997) 230–253, https://doi.org/10.1518/001872097778543886.
[75] W. Poortinga, K. Bickerstaff, I. Langford, J. Niewöhner, N. Pidgeon, The British 2001 Foot and Mouth crisis: a comparative study of public risk perceptions, trust and beliefs about government policy in two communities, J. Risk Res. 7 (1) (2004) 73–90, https://doi.org/10.1080/1366987042000151205.
[76] W. Poortinga, N.F. Pidgeon, Prior attitudes, salient value similarity, and dimensionality: toward an integrative model of trust in risk regulation1, J. Appl. Soc. Psychol. 36 (7) (2006) 1674–1700, https://doi.org/10.1111/j.0021-9029.2006.00076.x.
[77] Renner M., Lins S., Söllner M., & Thiebes S. (2021). Achieving trustworthy artificial intelligence: multi-source trust transfer in artificial In-telligence-capable technology. Systems, Austin, USA. https://www.researchgate.net/profile/Maximilian-Renner/publication/354923121_Achieving_Trustworthy_Artificial_Intelligence_Multi-Source_Trust_Transfer_in_Artificial_Intelligence-capable_Technology/links/61545d782b34872782fad1f7/Achieving-Trustworthy-Artificial-Intelligence-Multi-Source-Trust-Transfer-in-Artificial-Intelligence-capable-Technology.pdf.
[78] B.G. Robbins, What is trust? A multidisciplinary review, critique, and synthesis, Soc. Compass 10 (10) (2016) 972–986, https://doi.org/10.1111/soc4.12391.
[79] J.B. Rotter, A new scale for the measurement of interpersonal trust, J. Pers. 35 (4) (1967) 651–665. http://pdf.xuebalib.com:1262/17zuyR9VKGZT.pdf.
[80] J.B. Rotter, D.K. Stein, Public attitudes toward the trustworthiness, competence, and altruism of twenty selected occupations, J. Appl. Soc. Psychol. 1 (4) (1971) 334–343, https://doi.org/10.1111/j.1559-1816.1971.tb00371.x.
[81] D.M. Rousseau, S.B. Sitkin, R.S. Burt, C. Camerer, Introduction to Special Topic Forum: not so different after all: a cross-discipline view of trust, Acad. Manag. Rev. 23 (3) (1998) 393–404. http://www.jstor.org/stable/259285.
[82] J.M. Schraagen, Responsible use of AI in military systems: prospects and challenges, Ergonomics (2023) 1–16, https://doi.org/10.1080/00140139.2023.2278394.
[83] S.S. Shekhar, Artificial intelligence in automation, Artif. Intell. 3085 (06) (2019) 14–17.
[84] T.B Sheridan, Considerations in modeling the Human supervisory controller, IFAC Proc. 8 (1, Part 3) (1975) 223–228, https://doi.org/10.1016/S1474-6670(17)67555-4. Volumes.
[85] D. Shin, User perceptions of algorithmic decisions in the personalized AI system:perceptual evaluation of fairness, accountability, transparency, and explainability, J. Broadcast. Electron. Media 64 (4) (2020) 541–565, https://doi.org/10.1080/08838151.2020.1843357.
[86] D. Shin, B. Zaid, F. Biocca, A. Rasul, In platforms we trust?Unlocking the black-box of news algorithms through interpretable AI, J. Broadcast. Electron. Media 66 (2) (2022) 235–256, https://doi.org/10.1080/08838151.2022.2057984.
[87] B. Shneiderman, Human-centered artificial intelligence: reliable, safe & trustworthy, Int. J. Hum. Comput. Interact. 36 (6) (2020) 495–504, https://doi.org/10.1080/10447318.2020.1741118.
[88] B. Shneiderman, Human-Centered AI, Oxford University Press, 2022.
[89] M. Siegrist, Trust and risk perception: a critical review of the literature, Risk Anal. Off. Publ. Soc. Risk Anal. 41 (3) (2021) 480–490, https://doi.org/10.1111/risa.13325.
[90] M. Siegrist, G. Cvetkovich, Perception of hazards: the role of social trust and knowledge, Risk Anal. Off. Publ. Soc. Risk Anal. 20 (5) (2000) 713–719, https://doi.org/10.1111/0272-4332.205064.
[91] M. Siegrist, T.C. Earle, H. Gutscher, Test of a trust and confidence model in the applied context of electromagnetic field (EMF) risks, Risk Anal. Off. Publ. Soc. Risk Anal. 23 (4) (2003) 705–716, https://doi.org/10.1111/1539-6924.00349.
[92] M. Siegrist, H. Gutscher, T.C. Earle, Perception of risk: the influence of general trust, and general confidence, J. Risk Res. 8 (2) (2005) 145–156, https://doi.org/10.1080/1366987032000105315.
[93] S.B. Sitkin, N.L. Roth, Explaining the limited effectiveness of legalistic “remedies” for trust/distrust, Organ. Sci. 4 (3) (1993) 367–392. http://www.jstor.org/stable/2634950.
[94] P. Slovic, Perception of risk, Science 236 (4799) (1987) 280, https://doi.org/10.1126/science.3563507.
[95] P. Slovic, Perceived risk, trust, and democracy, Risk Anal. Off. Publ. Soc. Risk Anal. 13 (6) (1993) 675–682, https://doi.org/10.1111/j.1539-6924.1993.tb01329.x.
[96] M. Söllner, A. Hoffmann, J.M. Leimeister, Why different trust relationships matter for information systems users, Eur. J. Inf. Syst. 25 (3) (2016) 274–287, https://doi.org/10.1057/ejis.2015.17.
[97] S. Thiebes, S. Lins, A. Sunyaev, Trustworthy artificial intelligence, Electron. Mark. 31 (2) (2021) 447–464, https://doi.org/10.1007/s12525-020-00441-4.
[98] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C.G. Zelaya, A. van Moorsel, The relationship between trust in AI and trustworthy machine learning technologies, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 272–283, https://doi.org/10.1145/3351095.3372834.
[99] S. Wright, Trust and trustworthiness, Philosophia 38 (3) (2010) 615–627, https://doi.org/10.1007/s11406-009-9218-0 (Mendoza).
[100] D.E. Zand, Trust and managerial problem solving, Adm. Sci. Q. 17 (2) (1972) 229–239, https://doi.org/10.2307/2393957.
[101] F. Zerka, V. Urovi, A. Vaidyanathan, S. Barakat, R.T.H. Leijenaar, S. Walsh, H. Gabrani-Juma, B. Miraglio, H.C. Woodruff, M. Dumontier, P. Lambin, Blockchain for privacy preserving and trustworthy distributed machine learning in multicentric medical imaging (C-DistriM), IEEE Access 8 (2020) 183939–183951, https://doi.org/10.1109/ACCESS.2020.3029445.
[102] A. Jobin, M. Ienca, The global landscape of AI ethics guidelines, Nat Mach Intell 1 (2019) 389–399. https://doi.org/10.1038/s42256-019-0088-2.