DOI: 10.1111/risa.14245

ORIGINAL ARTICLE

Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences

Ann Bostrom[1] | **Julie L. Demuth**[2] | **Christopher D. Wirz**[2] | **Mariana G. Cains**[2] | **Andrea Schumacher**[2] | **Deianna Madlambayan**[1] | **Akansha Singh Bansal**[3] | **Angela Bearth**[4] | **Randy Chase**[5] | **Katherine M. Crosman**[6] | **Imme Ebert-Uphoff**[3] | **David John Gagne II**[7] | **Seth Guikema**[8] | **Robert Hoffman**[9] | **Branden B. Johnson**[10] | **Christina Kumler-Bonfanti**[11] | **John D. Lee**[12] | **Anna Lowe**[13] | **Amy McGovern**[5,14] | **Vanessa Przybylo**[15] | **Jacob T. Radford**[3] | **Emilie Roth**[16] | **Carly Sutter**[15] | **Philippe Tissot**[17] | **Paul Roebber**[18] | **Jebb Q. Stewart**[19] | **Miranda White**[17] | **John K. Williams**[20]

[1]Evans School of Public Policy & Governance, University of Washington, Seattle, Washington, USA

[2]Mesoscale & Microscale Meteorology Lab, National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA

[3]Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado, USA

[4]Consumer Behavior, Institute for Environmental Decisions, ETH Zürich, Zürich, Switzerland

[5]School of Meteorology, University of Oklahoma, Norman, Oklahoma, USA

[6]Department of Marine Technology, Faculty of Engineering, Norwegian University of Science and Technology, Trondheim, Norway

[7]Computational & Information Systems Lab, National Center for Atmospheric Research, Boulder, Colorado, USA

[8]Industrial & Operations Engineering, University of Michigan, Ann Arbor, Michigan, USA

[9]Institute for Human & Machine Cognition, Pensacola, Florida, USA

[10]Decision Research, Eugene, Oregon, USA

[11]Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado, USA

[12]Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, Wisconsin, USA

[13]Marine, Earth and Atmospheric Sciences, North Carolina State University, Raleigh, North Carolina, USA

[14]School of Computer Science, University of Oklahoma, Norman, Oklahoma, USA

[15]Department of Atmospheric and Environmental Sciences, University at Albany, State University of New York, Albany, New York, USA

[16]Roth Cognitive Engineering, Brookline, Massachusetts, USA

[17]Conrad Blucher Institute for Surveying and Science, Texas A&M University-Corpus Christi, Corpus Christi, Texas, USA

[18]School of Freshwater Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

[19]Global Systems Laboratory, Oceanic and Atmospheric Research, National Oceanic and Atmospheric Administration, Boulder, Colorado, USA

[20]The Weather Company, an IBM Business, Andover, Massachusetts, USA

Correspondence
Ann Bostrom, Evans School of Public Policy & Governance, University of Washington, Seattle, WA, USA.
Email: abostrom@uw.edu

Abstract

Demands to manage the risks of artificial intelligence (AI) are growing. These demands and the government standards arising from them both call for trustworthy AI. In response, we adopt a convergent approach to review, evaluate, and synthesize research

Julie L. Demuth, Mesoscale & Microscale Meteorology Lab, National Center for Atmospheric Research, Boulder, CO, USA.
Email: jdemuth@ucar.edu

Funding information
National Science Foundation, Grant/Award Number: ICER-2019758

on the trust and trustworthiness of AI in the environmental sciences and propose a research agenda. Evidential and conceptual histories of research on trust and trustworthiness reveal persisting ambiguities and measurement shortcomings related to inconsistent attention to the contextual and social dependencies and dynamics of trust. Potentially underappreciated in the development of trustworthy AI for environmental sciences is the importance of engaging AI users and other stakeholders, which human–AI teaming perspectives on AI development similarly underscore. Co-development strategies may also help reconcile efforts to develop performance-based trustworthiness standards with dynamic and contextual notions of trust. We illustrate the importance of these themes with applied examples and show how insights from research on trust and the communication of risk and uncertainty can help advance the understanding of trust and trustworthiness of AI in the environmental sciences.

KEYWORDS
artificial intelligence (AI), environmental science, risk communication, trust, trustworthiness

1 | INTRODUCTION

In 2019, the US National Science and Technology Council issued the National Artificial Intelligence Research and Development Strategic Plan (NSTC, 2019), updated from 2016. The plan called for long-term investments in fundamental artificial intelligence (AI)[1] research, with a particular focus on developing trustworthy AI systems. Although the plan did not define *trustworthy*, it mentioned factors relevant to system design (e.g., reliability), ethics (e.g., fairness), output (e.g., explainability, transparency), and performance (e.g., accuracy). Accordingly, "Trustworthy AI" was one of six high-priority themes identified in the National Science Foundation (NSF) call for National Artificial Intelligence Research Institutes (NSF, 2019). The NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES) responded to this call and was selected for funding.

To develop and better understand trustworthy AI, AI2ES engages expertise that integrates multiple disciplines, sectors, industries, and environmental applications, to converge across them. A major effort of the Institute is conducting convergent research to identify factors that (a) influence trust in and trustworthiness of AI for environmental professionals who are applying AI to inform their decisions, (b) determine how AI trustworthiness influences risk perception and use of AI for environmental sciences, and (c) inform the design of trustworthy AI methods for improved environmental decision-making. AI2ES focuses on machine learning (ML) applications, which are a subset of AI. As part of this effort, in August 2022, AI2ES hosted a virtual Workshop on Trust and Trustworthy AI, to review and evaluate the state of conceptualization, measurement, and modeling of trust and trustworthiness of AI systems for the environmental sciences, focusing initially—but not exclusively—on forecasting challenges in weather and coastal oceanography, and on four themes: evidential and conceptual history, measurement, context, and risk and uncertainty.[2] As reflected in this resulting synthesis, perspectives converged on some topics but diverged on others.

A focus of this synthesis is the consideration of how methodological and theoretical advances and challenges identified in risk communication and judgment and decision-making (e.g., Balog-Way et al., 2020; Fischhoff & Broomell, 2020) might augment how trust and trustworthiness are discussed and researched for AI. Although trust processes are central in risk communication and in judgment and decision-making under uncertainty (e.g., Earle, 2010; Padilla et al., 2021; Siegrist, 2021; Slovic, 1993, 2000), to date, such research appears to have only tangentially informed recent national and international efforts to: (a) develop trustworthy AI/ML standards and ethics (Broniatowski, 2021; EC,2020; EC HLEG, 2019; Schwartz et al., 2022; Tabassi, 2023), (b) increase awareness of the importance of trustworthy standards and ethics (Ammanath, 2022; Varshney, 2022), and (c) develop research agendas (National Academies of Sciences, Engineering, and Medicine [NASEM], 2022a).

Two questions underpinned many workshop participants' perspectives across these themes and thus served as focal points for the discussions: *(a) How do people develop their trustworthiness assessments or perceptions?* and *(b) How do AI developers design systems that are (more) worthy of trust?* Importantly, these questions tended to be posed separately, revealing differences in perspectives. Yet, there is also value in posing these questions jointly, allowing them to iteratively inform each other.

In the next section, we introduce the state of research on trust in and trustworthiness of AI, explore the roots and

[1] We adopt this working definition of AI from AI2ES: "Artificial Intelligence is the science and engineering of building machines that perform tasks normally associated with human intelligence."

[2] The workshop brought together 30 social, psychological, geophysical, and computational scientists with research and operational interests in trust, AI, and human–AI teaming (as defined in NASEM, 2022a). All participants authored or co-authored short thought pieces on trust and trustworthiness, written in advance of the workshop through the lenses of the four themes, or wrote summaries of workshop discussions (see Supporting Information). All workshop participants were invited to co-author this synthesis and most agreed.

evolution of trust research, and describe prevailing models of trust and its antecedents. Section 3 dives into measurement, considering first the conceptual ambiguities that have exacerbated measurement challenges, and then the diverse contextual dependencies that complicate measurement. Section 4 explores the implications for trust of differences in values, needs, and uses for AI. Section 5 characterizes the contested notion of calibrating trust. Section 6 highlights how meaningful communication of risk and uncertainty underlies notions of explainable AI (XAI) and interpretable AI. Sections 3 through 6 each close with research recommendations. The concluding sections of the paper summarize our research agenda for trust in and trustworthiness of AI in the environmental sciences. To guide the reader, we summarize those research needs and recommendations here:

1. There is a need for better ways of measuring trust in AI as a dynamic, contingent process;
2. to accomplish this will require a better understanding of which contingencies and contextual factors matter;
3. efforts to develop standards and "calibrate" trust deserve close scrutiny and empirical evaluation;
4. these efforts can benefit from what has been learned in risk and uncertainty communication studies of trust.

The paper closes with reflections on how to build a more inclusive, interdisciplinary research community to pursue this research agenda.

2 | STATUS AND FOCUS OF RESEARCH ON TRUST IN AND TRUSTWORTHINESS OF AI

AI has been applied to weather forecasting for decades (e.g., Glahn & Dallavelle, 2000; Glahn & Im, 2011; Malone, 1955). For example, AI has been used to forecast the probability of tornadoes, hail, damaging wind, and hurricane intensity (Billet et al., 1997; Demuth et al., 2006; Gagne et al., 2017; Kitzmiller et al., 1995; McGovern et al., 2014; Williams, 2014; cf. McGovern et al., 2019). Yet end users often criticize AI for being a "black box" because of the perceived inability to understand how AI makes its predictions, an interpretability problem that is broader than environmental sciences (McGovern et al., 2019). The rapid advance of AI techniques and applications has, if anything, increased interest in the interpretability problem (e.g., Future of Life Institute, 2023). AI systems often perform remarkably well but are limited in terms of their ability to provide meaningful explanations for their conclusions and decisions (Lipton, 2018; Rudin, 2019) and are prone to biases and brittleness that can negatively impact perceived trustworthiness (McGovern, Ebert-Uphoff, et al., 2022).

2.1 | Trust of AI in human–AI teaming

Addressing these and other challenges, the National Academies produced a consensus report on human–AI teaming that summarized much of the emerging literature on trust in and trustworthiness of AI (NASEM, 2022a). The report builds from the evidence that effective human–human teaming depends on having *common ground*—a common understanding of the current situation and thus the current goals and priorities—as well as a shared understanding of the roles, capabilities, and limitations of multiple team members (e.g., Klein et al., 2005). This conceptualization of the human–AI relationship uses the term partnership to engage human–human teaming literature and describe how the human negotiates with the AI. The same need for establishing common ground is fundamental to the decision of whether to trust an intelligent machine agent (i.e., AI *partner*) in a particular situation, wherein *user–expert* dialogue (i.e., information exchanges between a human user and an expert system, such as an AI model) is best viewed as a negotiation process (Pollack et al., 1982). Consequently, there is a need to support the cognitive work required to give the human partner confidence that the AI partner is solving the right problem. This requires (a) a *shared representation* of the current state of the world and current goals and priorities, as well as constraints on a satisfactory solution; and (b) an ability for the human partner to contribute to the AI partner's representation of the current state of the world, goals, and priorities (Roth et al., 2018). Enabling this negotiation process requires controls that allow the human partner to add to or correct the AI partner's representation of the situation and to communicate any additional constraints that need to be respected as well as any changes in goals and priorities.

Much AI design research has focused on developing visualizations and controls that allow the human partner on the scene and the AI partner to come to a common representation of the situation being confronted and of the relevant goals, constraints, and priorities to generate a solution that aligns with the human partner's objectives (NASEM, 2022a).

2.2 | Models and antecedents of trust

Building on the above, this section provides a brief but structured overview of research on modeling trust and its antecedents, given the importance of understanding what drives trusting. The section closes with cautions emerging from empirical evidence on the context- and application-specificity of such findings.

An important class of relevant models of trust in AI lists dozens of causal factors that are believed to have a direct causal influence on trust in automation (e.g., Muir, 1987; Schaefer et al., 2016), including cultural differences, operator predispositions, personality, and knowledge about the automation, and various contextual factors. A second class of models includes process models designed to predict or estimate trust judgments (cf. Seong & Bisantz, 2002; see Lee & See, 2004). Hoffman (2017) presented over 40 diverse flavors of trust in automation and machines, illustrating that trust is legion (see also Dorton, 2022). There are also varieties of mistrusting (not trusting a machine decision) and negative trusting (i.e., distrusting—believing the machine will actually

do something bad). Trust is emergent and evolving—not a state or static—and is a property of the interactions between humans and AI. This dynamism of trust in AI is especially salient given the rapid evolution and increasing ubiquity of AI. It follows that the delivery of an AI system (i.e., model) is not a sufficient endpoint of the AI production cycle but rather the beginning of a maintenance and improvement system to maintain trustworthiness and manage trust processes with end users. The recent in-depth review of human–AI teaming research by NASEM (2022a) and systematic review of antecedents of trust in otherware (intelligent and interactive technological systems; Saßmannshausen et al., 2023) reach similar conclusions.

An alternative approach to conceptualizing and modeling trust in automation is to assume that humans anthropomorphize computers (even AI that is embedded and invisible to the user) such that models of human–human trust apply directly. In this vein, anthropomorphism has recently been suggested to be an influential factor in trust in AI systems (Kaplan et al., 2021). The Computers as Social Actors (CASA) experimental paradigm developed at Stanford University has shown that many "social scripts" associated with human–human interaction extend to human–computer interaction (Reeves & Nass, 1996), leading to the conclusion that humans treat computers as social actors. CASA findings and similar work can be interpreted as evidence that trust, a feature of human social interaction, is used in our interactions with machines and automated systems (Lee & See, 2004; Moon, 2000; Morkes et al., 1999; Nass et al., 1994). Lee and See (2004) take this perspective and note that trust in automation is likely a largely affective process.

The extent to which a computer is perceived as social may also influence the extent of the social response (Morkes et al., 1999). The role that trust plays may, however, vary with the characteristics of the AI, including humanlike appearance (de Visser et al., 2012, 2016, 2017), which may slow declines in trust stemming from decreased AI reliability, and humanlike communication style, which is associated with greater perceived trustworthiness (Jensen et al., 2020), up to a point (Kim et al., 2019). In general, characteristics of a technological system that elicit perceptions of humanlike or social qualities appear to influence users' willingness to be vulnerable to (i.e., to trust) that system. While social responses to computers have been found to generalize across levels of computer expertise (Reeves & Nass, 1996), whether and how these findings might extend to professional users, such as weather forecasters, is still unclear.

Three models from risk analysis and management research offer promising additional insights for modeling trust in AI. First, tests of the intuitive detection theorists (IDT) model (White & Eiser, 2006; White & Johnson, 2010) building on signal detection theory found evidence for three types of judgments that predict assessments of risk managers' trustworthiness: perceived competence (the ability to accurately distinguish danger from safety), perceived care (the propensity to act when danger or safety is uncertain), and perceived openness and honesty. Second, the associationist model from Poortinga and Pidgeon (2005, 2006) characterizes trust as the outcome of a general positive or negative attitude toward a technology (e.g., genetically modified food). Third, the trust, confidence, and cooperation (TCC) model of Earle and Siegrist (2006, 2008; Earle et al., 2010) posits that trust is driven by judged sharing of salient values with the target, as those unfamiliar with the target lack other cues, whereas confidence is driven by the target's past performance among those familiar with the target. The TCC model posits that trust influences judgments of both past performance and confidence. The TCC model also admits other exogenous factors, including general (interpersonal) trust. Empirically, the association between trust and confidence is high (Earle & Siegrist, 2006; Johnson & Rickard, 2023), leading to questions about whether trust and confidence can be distinguished, conceptually or practically (see Section 3.1 for further discussion of this).

As can be inferred from these three risk models of trust, antecedents of trust probed in risk and social sciences commonly describe two or three dimensions, primarily the trustor's perceptions of the trustee's competence or ability, honesty, and caring or benevolent attitudes toward those affected by, or vulnerable to, the trustee's decisions (e.g., Johnson, 1999; Siegrist, 2021). These factors tend to appear whether researchers are using measures of general trust in people (e.g., see Poortinga & Pidgeon, 2003) or of trust in specific individuals, groups, or institutions (e.g., see Cvetkovich & Nakayachi, 2007). Approaches to labeling, empirically measuring, and clustering these concepts vary across disciplines and studies, which may obstruct a common understanding of trust antecedents (Johnson, 1999). The emerging field of trustworthy AI has built on conceptualizations and findings from the rich kinds of literature on interpersonal trust (Lewicki et al., 1998; Mayer et al., 1995; Rousseau et al., 1998) and trust in automation (Hoff & Bashir, 2015; Muir, 1987), although some suggest that the analogy to interpersonal trust is a misleading anthropomorphism (Wirz et al., 2023), and others that trust in technologies is better characterized as confidence (Siegrist, 2021).

Because computers are tools programmed by humans, our trust in AI systems may stem in part from our perceptions of the developers of those systems (e.g., Hoff & Bashir, 2015). This could imply that users take a logical approach to their interactions with technology, forming beliefs about technological systems that are based on evaluations of the credibility and skill of the people who created the system (i.e., the developers' past performance and authority). Or it could be a result of our needing to interact with trusted experts in order to achieve a satisfying mental model of how an AI system works (Hoffman et al., 2021). Alternatively, social motivations such as liking may drive interpersonal trust between users and developers; people are more apt to engage in a behavior, such as using a model, if others who share similar values engage in or support that behavior or if the behavior benefits someone we like (Contractor & DeChurch, 2014).

A recurrent theme across many studies is that trust and trustworthiness are related to contextual factors and do not operate the same way across differing AI applications (e.g., Ashoori & Weisz, 2019; Chiou & Lee, 2021; Glikson &

Woolley, 2020; Lewis & Marsh, 2022). Within the context of trust in automation (an umbrella term including artificially intelligent agents), Chiou & Lee's (2021) relational framing of trust emphasizes AI responsivity—the ability to respond to sudden environmental changes. Similarly, Lewis & Marsh's (2022) conceptual model for trust in AI emphasizes the individual and subjective dimensions of trust and trustworthiness. Others have identified the decision context (Ashoori & Weisz, 2019) and task characteristics (Glikson & Woolley, 2020; Hoffman et al., 2018) associated with AI as important points for understanding the trustworthiness of AI. Adding nuance to "AI," Glikson & Woolley (2020) broke down their examination into more specific subcategories of AI (*robotic*, *virtual*, and *embedded*; embedded AI is not visible to the user); even this high-level distinction identified differences among the different application types in findings regarding trust and trustworthiness. Hence, it is paramount to engage a context- and application-dependent conceptualization of trustworthy AI, and of the usability and usefulness of AI, with the latter two related, but not equivalent, to trustworthiness.

3 | MEASUREMENT

Trust measurement has been studied for decades, with studies of self-report measures, behavioral measures, and trust games dating back to the mid-20th century (Bauer & Freitag, 2017) and including recent, diverse measures of trust in automation and AI (Hoffman et al, 2018; Roth et al., 2022). Yet trust measurement continues to be complicated by ongoing discussions about the degree to which trust is subjective and distinct from concepts such as confidence, credibility, and reliance, likely reflecting the importance of context and application. The need for better measures is widely recognized (NASEM 2022a, p. 54; Vereschak et al., 2021)

3.1 | Measuring what? Trust, confidence, and related concepts

The TCC model of trust makes a key distinction between *confidence* as a function of reliability and competence and *trust* as strongly influenced by perceived goal alignment and motivations, although some of its authors have called for better explanations of and distinctions between trust and confidence (Siegrist, 2010). In the TCC model, trust is characterized as "social and relational" and defined as "the willingness, in the expectation of beneficial outcomes, to make oneself vulnerable to another based on a judgment of similarity of intentions or values," whereas confidence is deemed "instrumental and calculative" and defined as "the belief, based on experience or evidence (e.g., past performance), that certain future events will occur as expected" (Earle, 2009, p. 786). Similarly, in the human factors literature, the distinction has been made between an agent's benevolence (alignment of its goals with the user's goals) and its competence (ability to perform a task; Lee & See, 2004).

Both trust and confidence are dynamic and situational. For instance, in the context of AI and weather, a forecaster can trust a type of AI guidance as often skillful and therefore useful in one weather situation but still have low confidence that the guidance is skillful for a given prediction in some other weather situation. Ensembles of weather predictions—which can situationally generate a distribution of outcomes that are more or less sharp and therefore with narrower or wider confidence intervals—support this idea of confidence being situational (see also Henderson et al., 2023).

Consistent with the distinction that confidence is a function of reliability and competence, and in contrast to other views (cf. NASEM, 2022a), Siegrist (2021) concludes that "trust is not an appropriate term to describe the reliance on modern technology" (p. 484). Siegrist further identifies a lack of causal research as inhibiting progress in understanding the relationship between trust and risk perception, and the distinction between trust and confidence, while acknowledging that these are hard to measure (Earle & Siegrist, 2006; Johnson, 1999; Johnson & Rickard, 2023). Causal research using validated measures might help distinguish judgments based on past performance from those based on salient value similarity and provide additional insights into how to measure these concepts. Judgments made on past performance include *confidence* in the TCC model and *discrimination ability* and *response bias* in the IDT model. Judgments based on salient value similarity include *trust* in the TCC model and *benevolence* in communication and human factors models.

3.2 | Measuring in what context? Contextual dependencies

By implicitly overlooking contextual dependencies and dynamics that are important in determining and measuring trust, some trustworthy AI efforts shift the onus of trustworthiness away from system designers alone to share it with operators:

> To achieve trust, AI system designers need to create accurate, reliable systems with informative, user-friendly interfaces, while the operators must take the time for adequate training to understand system operation and limits of performance. Complex systems that are widely trusted by users, such as manual controls for vehicles, tend to be transparent (the system operates in a manner that is visible to the user), credible (the system's outputs are accepted by the user), auditable (the system can be evaluated), reliable (the system acts as the user intended), and recoverable (the user can recover control when desired). (NSTC, 2019, p 25).

Contextual dependency must be a consideration in any analysis of complex cognitive work systems. The multiple contextual dependencies of trust measurements may include:

(1) attributes of the model user, such as expertise and biases; (2) attributes of the modeler or others involved in the AI algorithm production chain; (3) the complexity and specifics of the problem being modeled and of the model itself, including societal implications of the model application in a given situation; (4) time constraints and pressures for decision-making; and (5) what complementary and redundant information is simultaneously available for the problem at hand.

For instance, in the course of their work, forecasters predicting severe convective weather (e.g., tornadoes, large hail, strong winds) are likely to consult multiple sources of numerical weather prediction (NWP) model guidance, including several predictions of the same output, which is often referred to as a "poor man's ensemble" (Arribas et al., 2005; Novak et al., 2008), probabilistic guidance, and traditional ensembles. They also consult multiple parameters predicted by NWP guidance —including individual ingredients for severe convective weather (e.g., convective available potential energy, shear) and proxies or storm "surrogates" (e.g., updraft helicity tracks, see Sobash et al., 2016)—and their corresponding magnitudes at multiple, successive forecast valid times (e.g., 00,06,12,18 UTC). Furthermore, forecasters face the constraints and challenges of systematic forecast processes that are shaped by the operational setting, including that new NWP guidance and observational data (e.g., from radar, satellite, surface stations) are constantly updating and made available to them, on timescales of minutes to hours. They also face time pressures to assess the information they have, to determine hazardous weather risks, and to issue either an event-driven product, like a watch or warning, or one of dozens of other products (e.g., a freeform forecast discussion) by their issuance time as mandated by federal policy (NWS, 2022b).

Weather forecasters' operational risk assessment and forecast decision-making environments are complex, dynamic, and extremely information-rich. Thus, when trust in new AI guidance is developed or evaluated in operational settings, interactions among these contextual factors may limit researchers' ability to draw conclusions about the AI that generalize to other applications or settings. These contextual factors also illustrate the numerous ways in which laboratory experiments or testbeds may fall short. Together, this suggests that systematic attention to contextual attributes and interactions will be required to understand what observed behavior may generalize for which AI models.

Box 1. Trustworthy for whom? AI/ML model development for environmental applications

In any given decision-making context, different groups and individuals will interact with the subsequent information value chain, including with the use of AI therein. Take the example of zooplankton data collected and analyzed with the use of AI. If the driving decision-making question is how to sustainably develop and manage a zooplankton fishery, the question (the initial stage of the information value chain) is likely to be determined by public decision-makers and those who influence them (legislators and/or fisheries agency personnel, fisheries industry groups). Once the question is defined, much of the work from metrics/measures through knowledge production is done by technical and scientific experts (engineers, biologists, and the like). From information to decision-making, we will see further involvement of fisheries managers; in many systems, fisheries stakeholders (lay experts) are also likely to be involved at the decision-making stage. Outcomes will be experienced by all, but most immediately by fishers themselves—both fishers of zooplankton and other operators affected by zooplankton fishery development. Outcomes may also be salient to the general public, particularly if there is media attention that springs from novelty, controversy, management failures, or some other source. Each of the above groups (and individuals within each group) will have different perceptions of the trustworthiness of the technologies (including AI) used in the information value chain, as well as different perceptions of the trustworthiness of other groups and individuals involved therein (Siegrist, 2021).

In a contrasting example from wildland firefighting, a fire behavior analyst observes environmental cues, discusses the interpretation of the fire weather (AI) model output with the incident meteorologist, determines what the fire activity will be, for example, whether the wind will shift, and communicates this to a supervisor and through to affected firefighting crews. The expertise and experience of the incident meteorologist contribute to their reliance on and trust in the fire weather model output and in turn to the meteorologist's interactions with the fire behavior analyst and their trust and actions. In this example the ultimate outcomes may be salient to firefighters, local communities, and the broader public and affect their trust in AI.

3.3 | Research recommendations: Measuring trust and trustworthiness

Better definitions that distinguish trust and trustworthiness from related concepts, such as confidence and reliability, are needed to inform measurement. Additional research is needed to develop a fuller theoretical understanding of the relationship between trust and trustworthiness and their antecedents (including contextual factors), moderators, and outcomes. These challenges of measuring trust are amplified by the dynamics of trusting. Better characterization of these dynamics is needed for AI. What processes might influence trust through time as AI guidance and real-world contexts for that

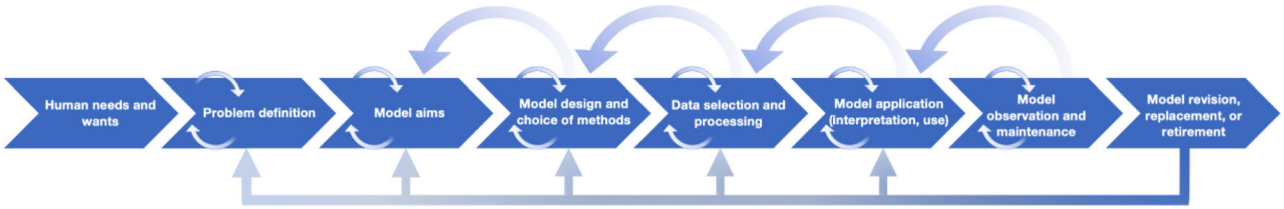

FIGURE 1 The iterative, value-driven artificial intelligence/machine learning (ML) model production cycle. Adapted from the Information Value Chain (Abbasi et al., 2016) and the ML life cycle (Custis, 2022).

guidance evolve? The dynamics of trust are rarely the primary focus of research; there is a need for studies measuring how trust can influence post hoc evaluations of performance and is in turn influenced by performance over time, taking into account time pressures on AI use (Box 1). Intertwined with this is the need for research on how diverse decision-making strategies affect trust judgments. When and how does analytic thinking influence trust in AI and AI use decisions? What are the roles of heuristics in decision-making under uncertainty with AI, including new heuristics that may be specific to AI? Longitudinal process-tracing studies would be beneficial for understanding how trusting and related cognitive and affective processes evolve over the (co-)development and use of specific AI models or model types by professionals with similar expertise charged with different AI-related decision tasks, as would multimethod, multidisciplinary efforts to validate measures of trust, reliance, and related concepts.

4 | THE CONTEXTUAL DEPENDENCIES OF DIVERSE PERSPECTIVES ON TRUST IN AI

Risk communication research shows there is likely to be a variety of stakeholders with different motivations and mental models that will impact their trust in a new technology (Fischhoff et al., 1984; NRC, 1996). To design effective risk communication about a technology or risk requires engaging with and understanding the needs of diverse stakeholders (e.g., Fischhoff, 1995; Morgan et al., 2002 ; Prior & Paton, 2008; Stanton & Jensen, 2021). Such stakeholders include the technology developers themselves, the professional users of the technology, and the larger set of vested or interested parties that may include funders, regulators, politicians, as well as other users and affected groups. Further, stakeholders' needs and perspectives are shaped by individual, institutional, and cultural social events and conditions (Renn, 2008a, 2008b). Respecting this heterogeneity of perspectives is an important tenet of stakeholder engagement and co-design. Co-design goes a step beyond human- or user-centered design (HCD or UCD) to include those for whom an AI model is being developed in an ongoing participatory design process (Bjerknes & Bratteteig, 1995; Carroll, 1996). Accordingly, co-design and engagement across the entire AI lifecycle (Figure 1) can be seen as fundamental for developing trustworthy AI (see also Hoffman et al., 2010).

The perspectives of both developers and diverse users matter with regard to developing trustworthy AI. Involving both in the design process has the benefit of increasing the robustness and usability of the resulting AI (Singer et al., 2022). The broader benefits of co-design include those identified for UCD, such as more efficient, effective, and safe products (Norman, 1988). However, there are outstanding questions about the capacity and resources needed to develop, scale, and sustain co-design processes. A particular concern is how broadly the net should be cast in co-design processes with regard to potential AI model users. In their examination of UCD (e.g., with one-time consultation as one end of a continuum and participatory co-design processes as the other), Abras et al. (2004) highlight that concomitant costs and benefits of such processes differ across this continuum.[3]

Reflecting on these ideas of support for UCD but concern about the degree of doing it, we note that users of AI need to be involved in the development process, in order to co-develop and co-design guidance and systems that are useful, usable, and used by them, but this idea poses challenges in practice. Who needs to be involved, to what extent, and during which parts of the AI model development process? There are challenges with having many different groups be heard, including developers, decision-makers, and communities who may be affected. For many users, AI is only one type of information for which their co-produced involvement would be beneficial; this poses a threat of participant fatigue. There are related ethical issues about asking for developers' and users' time: When and to what degree is it worthwhile for a user to offer their input? Additionally, there are intertwined research design and measurement limitations and challenges with asking users to provide input and evaluate AI information outside of their naturalistic environments. This approach forces all parties involved to balance the benefits of co-development with the increased demands it places on already constrained time and resources in increasingly fast-paced and competitive research environments.

The diversity of perspectives on trustworthy AI can be simplified to illustrate the co-design challenge by thinking about users (or audiences) as divided into "deep divers" and "surface skimmers." Deep divers are those who have at least some AI expertise, are deeply interested in the modeling, and

[3] In 2005, Norman cautioned that UCD can go too far, resulting in overly complicated designs tailored for specific users, and called for technologies to be activity-centered rather than user-centered per se (Norman, 2005).

are readily considered by developers and others (e.g., those governing AI) as intended users/audiences. In contrast, "surface skimmers," soon to become the majority of AI users, are those who would like to use the AI model and could potentially benefit from its application but are not deeply invested and have little to no AI expertise. While it may be feasible to identify and recruit deep divers for co-design processes, surface skimmers are by definition not easy to identify reliably and may lack motivation to participate in co-design processes. Even this categorization is simplified, however, and may be context-dependent since a given user may, for example, be a "deep diver" in one use context and a "surface skimmer" in another.

Further complicating co-design of AI models, trust can be influenced at any point in the iterative, value-driven AI model production cycle (Figure 1). Some steps of this process, such as data selection, may influence trust in AI models more than they influence trust in other (non-AI) environmental modeling efforts.

4.1 | Research recommendations: Modeling the role of humans in AI trust and trustworthiness

The research summarized in this section points to a need for better ways to effectively share knowledge among user groups and research teams while balancing the needs and demands of each group. It's important to distinguish participatory processes like co-design and co-production from user-oriented or UCD and production; the former involves a greater degree of interaction between developers and users, whereas the latter may involve only periodic interactions between developers and users, making it less costly and less prone to over-complication but also potentially less informative. Models of human trust processes are needed that can predict and inform such trade-offs in addition to informing the goals of AI models. Analogizing with Fischhoff's (1995) stages of risk communication development, questions requiring further research in AI development processes (Figure 1) include: Who are the users and audiences that developers should consider, what are their needs, and what effects do differing approaches to engaging users have on trust and trustworthiness of AI? A systematic review of where and to what extent AI models have been co-produced and the perceived benefits and costs of such co-production processes would be one way of advancing this research.

5 | STANDARDS AND THE SUBJECTIVITY OF TRUST

Although trust can be described as warranted or unwarranted (Jacovi et al., 2021), whether trust can actually be calibrated to an external standard that defines trustworthiness is debated. Many discussions of trustworthy AI are firmly grounded in the notion that trustworthiness can or should be characterized by standards or benchmarks. In contrast, trust can also be described as inherently emotional and subjective (Section 5.2), for which reason it can be argued that its appropriateness cannot and should not be judged solely by AI performance standards (Wirz et al., 2023). Emphasis on the importance of decision-makers' goals and contextual determinants of trust and trustworthiness has increased, as is evident from recent considerations of "contracts" (Jacovi et al., 2021), "off-label" use (i.e., outside the boundaries of the situations the model was designed for, Schwartz et al., 2022 p. 17), and model cards (e.g., Mitchell et al., 2019) to signal the contextual determinants of model performance and what constitutes appropriate use. These latter efforts attempt to address the question of how to know and communicate what is "within label" and what is "off-label." To date, there has been little explicit effort to reconcile these ways of thinking (i.e., contracts, on/off-label use, model cards) about how to signal what constitutes appropriate use of an AI model, although risk management and standard-setting efforts do reference them (e.g., the NIST Risk Management Framework mentions reviewing uses of AI systems for "off-label" purposes [MAP1.5]; Schwartz et al., 2022).

5.1 | What does it mean to calibrate trust?

National discussions of AI strategy explicitly ground the notion of trustworthiness in performance standards. The NSTC 2019 update notes that "assessing, promoting, and assuring all aspects of AI trustworthiness requires measuring and evaluating AI technology performance through benchmarks and standards. Beyond being safe, secure, reliable, resilient, explainable, and transparent, trustworthy AI must preserve privacy while detecting and avoiding inappropriate bias" (NSTC 2019, p. 33), thereby emphasizing performance, privacy, and lack of bias as bases for trustworthiness.

It could be argued that the point of standards is to provide ratings of how well AI systems perform on given criteria and make that information available to potential users. Accordingly, minimum thresholds for acceptability that apply across all criteria or observers need not be set or applied. It appears that governmental and professional organizations are doing just that, however, as they have created lists of criteria to assess and measure the trustworthiness of AI (e.g., Tabassi et al, 2023; EC AI HLEG, 2019, 2022; White House OSTP; and Exec. Order No. 13960, 2020).

If trustworthiness standards apply, measurements can be made with respect to those standards. In apparent direct contradiction to those who argue trust is inherently subjective and its appropriateness should not be determined by reference to standards, others have called for assessing whether trust in a model is well-calibrated (warranted) according to standards, or is biased high (over-trust, unwarranted) or low (under-trust, unwarranted). In this view, trust is termed calibrated to the extent that it matches trustworthiness (Lee & See, 2004). Both under- and over-trust can undermine the potential for AI to support and enhance decision-making. Fur-

ther, from a measurement perspective, the wider the range of trustworthiness perceived as similar, the poorer the resolution of trust, and the less clear it will be whether it is well-calibrated according to some standard.

From the perspective of calibrating trust to standards, both performance and purpose can be used to describe the bases for trustworthy AI (Lee & Moray, 1992), contingent on context. *Performance* contributes to trustworthiness only to the extent the AI can contribute usefully in a specific context consistent with the decision-maker's goals. *Purpose* refers to why the AI was developed, corresponding to the designer's intent (Lee & See, 2004). Using AI as intended is more likely to lead to the trustworthiness of a system than if it is used for an "off-label" purpose. This can be compared to the notion of a contract in Jacovi et al. (2021), and similarly implies a need for metadata (information about the model) indicating when AI model use might be "off-label," making the model less trustworthy.

5.2 | On the insufficiency of the idea of calibrating trust

Because trust involves an attitudinal dimension with an affective component, trustworthiness based on standards and criteria only partially determines trust. People filter trust-relevant information through a lens of emotion: people not only think about trust, they also feel it (Fine & Holyfield, 1996). Feelings can cause trust to spread across time and system components in a way that changes in context would not predict (Loewenstein et al., 2001; Schwarz & Clore, 2003; Wormwood et al., 2019). As noted previously, trusting is dynamic, and depends on the context of the person, AI, situation, and temporal sequence of interactions (Chiou & Lee, 2021), with each interaction between the person and AI shaped by previous interactions (Chiou & Lee, 2021; Klein et al., 2006).

Further, trustworthiness is also social. Developers and users may change with time, with concomitant changes in trust. AI in the environmental sciences is usually applied in a larger social context. For example, in severe weather prediction, multiple people may be interacting with the AI, other model guidance, and one another, across varying decision contexts. Organizations also shape trust in and the trustworthiness of AI, in that they have purposes and shape social networks, which in turn shape interactions between people and AI systems.

5.3 | Research recommendations: Goal alignment, calibration, and standard setting

Reconciling notions of trust calibration and trust as subjective is a work in progress. Proposed models of the calibration of trust and perceived needs to calibrate goals for and trust in AI models treat calibration as feasible and useful. From the developer's perspective, finding better ways to align the expectations and use of AI within the bounds of the system's competencies, or what the model can do well, remains a challenge. From the regulator's perspective, calls for risk management have been interpreted as a demand for standards. From the user perspective, standards may enhance the possibility of trust and appropriate use of AI, but do not necessarily determine trust; rather, a gap between perceived trust and trustworthiness as defined by standards might be diagnostic of opportunities for interventions. Some opportunities may reflect conditions that are easier to change (e.g., standards that omit important factors in trust) than others (e.g., the gap is due to contextual variation).

Differences among outcomes and actors that might be identified as requiring calibration carry diverging implications for research moving forward in this space. For example, even if levels of trust are calibrated between developers and users of a specific AI model, that calibration process may not necessarily lead to increased use of the model within its competencies. Relatedly, the alignment of goals among users and with an AI model, or with the model's developer(s), may not necessarily foster trust in the model. Although calibration and/or alignment might have been pursued in individual AI studies, these nuances highlight the need to more systematically examine calibration and alignment in ways that clearly identify the specific concepts of interest, as well as the actors engaged in calibration and alignment of AI models for the environmental sciences, and weather forecasting in particular.

One potential way to facilitate calibration might be through the use of checklists, as has been done in the risk domain (e.g., Haynes et al., 2015). The general idea is to use checklists to identify important dimensions of the outcome to be calibrated (e.g., trust, confidence, expectations, reliance, use) in order to help assess how closely two groups are (or are not) aligned. Checklists can be flexible and adaptive—the specific items could be highly subjective or more objective depending on needs. However, if checklists were used, they would require some level of co-development as discussed in the previous section.

For those looking for guidance on how to design trustworthy AI that goes beyond performance standards like accuracy and reliability, a complementary approach is increasing the responsiveness of the AI (Chiou & Lee, 2021). This has been suggested as a promising path to increasing the joint performance of the person and AI. The multiple pathways for the person and AI to interact, both directly and through the situation (through observations, actions, influences, and signs), provide methods to better align the goals of the AI and the person applying it. These interactions go well beyond older conceptions of AI that relegated people to entering the initial information and acting on the resulting AI decisions (Roth et al, 1987). In this view, the person would be enhancing the trustworthiness of the AI, and the AI might enhance the trustworthiness of the person. Research on interactivity could facilitate both goal alignment as articulated by the human–AI teaming notion and explainability as interactivity and interrogation. Evaluating these psychological mechanisms as well as how trust changes over time in actual work contexts along

with the outcomes of trust, trustworthiness, and use would be important.

In risk research, it is valuable to characterize risks and set some standards, such as exposure standards, but also to recognize that some subpopulations may be more vulnerable. For example, children can react adversely to some exposures (e.g., drug doses) that might not harm adults. Analogously, AI model performance that informs decisions with few or no consequences in some contexts may risk causing unacceptable harm in other contexts. Further, it is well known that risk preferences and decisions regarding technology are a function of factors beyond the probabilities and sizes of adverse consequences; they depend also, for example, on perceptions of how familiar or dreaded the technology is (a.k.a. psychometric factors; Fischhoff et al., 1984; Slovic, 2000). Thus, for risk management, it is imperative to understand decision context-specific risk perceptions and attitudes of decision-makers and those who influence or are affected by their decisions. Analogously, it will be important to understand whether trust and trustworthiness of AI have psychometric properties that are not captured by standard AI performance metrics.

6 | TRUST, RISK, AND SCIENTIFIC UNCERTAINTY

6.1 | Communicating risk and scientific uncertainty for understanding and trust

Both explainability and interpretability of ML models have been lauded as essential for trustworthy AI and listed as criteria for trustworthy AI systems (e.g., Tabassi, 2023). Rudin et al. (2022) have, for example, offered five principles for creating a predictive AI model that is not a black box. In their terms, interpretability is achieved to the extent that the AI obeys a domain-specific set of constraints that allow it to be more easily understood by humans (Principle 1). Rudin et al. emphasize that interpretable models can create or enable trust or distrust but do not necessarily do so; rather, they permit a decision of trust than create trust itself (Principle 2). The fifth of their principles is that for high-stakes decisions, interpretable models are preferable to "explained" black box models.

As illustrated by the first of Rudin et al.'s principles, definitions of interpretability and explainability point to several ways in which research on judgment and decision-making—and risk communication research in particular—can provide insights into trust in AI. Trust in AI, and the extent to which AI is deemed trustworthy, is contingent on communications processes and products in AI, such as model or XAI outputs, or interfaces for imposing constraints on AI models; the visual presence of AI tends to increase trust in AI (Gilkson & Woolley, 2020). Many studies have called for or investigated explanations and XAI (McGovern, Bostrom, et al., 2022) as an approach to increasing trust (e.g., Hoffman et al., 2018; Lockey et al., 2021; Miller, 2019; Mueller et al., 2019; Tulio et al., 2007), while such explanations have often relied on visualizations (McGovern et al., 2019).

Risk communication research (e.g., MacEachren et al., 2012; Padilla et al., 2018, 2023; Spiegelhalter, 2017; van der Bles et al., 2019, 2020) shows, however, that the effects of visualizations and other communications of uncertainty on trust are likely to depend on the quality and context of communication (see also Box 2). Thus, there is a need for meaningful and useful information rather than just more information.

6.2 | Research recommendations: The role of communicating risk and uncertainty in AI trust

In the last few decades, more systematic studies of science and risk communication have begun to chart a course toward a cumulative understanding of risk communication as a dynamic system (e.g., NASEM, 2017). In this system, behavior is driven by human social motives: the need for accuracy—which can be influenced by social proof (what others are doing, whether scientific evidence supports the behavior) and authority (including scientific authority)—and the need for social belonging and affiliation. Social interactions are based on liking (acting similarly to those one likes), reciprocity, consistency, and accountability (Contractor & DeChurch, 2014). Contractor and DeChurch demonstrate empirically how risk communication can capitalize on this by pairing sources of peer influence (based on social networks) with social interactions (based on social motives), in a structured influence process (SIP). Translating and testing the SIP framework and findings from other research on communicating risk and uncertainty for embedded AI contexts such as environmental forecasting is not trivial but might be fruitful.

Box 2. Trust processes in weather applications of AI models

AI models developed for predicting power outages from extreme weather events such as hurricanes have been implemented and used in-house by major utilities spanning the United States, in some cases, being run with every weather forecast update to support decisions on power restoration crew allocation (Guikema et al., 2014; Kabir et al., 2019; Li et al., 2010; Quiring et al., 2014; Singhee et al., 2016). The outputs of these models have also been used in real-time by the Federal Emergency Management Agency (FEMA) and the Department of Defense (DoD) to support hurricane preparation decision-making. The models themselves are probabilistic supervised ML models, with some being complicated Bayesian model averaging ensembles. The challenge of balancing consistency and accuracy through iterative model update rounds,

and the challenge of conveying the (sometimes very large) uncertainty in the model predictions without losing trust in the model, have made the importance of model stability and clear communication of uncertainty and model limitations obvious. These experiences have demonstrated to model developers that gaining trust requires being honest about what the model can and cannot do and what the uncertainties are in the model predictions. This has to be done in a way that model users without expertise in AI can truly understand and internalize.

The Weather Company (TWC) is one of the most trusted brands in the United States (Morning Consult, https://morningconsult.com/most-trusted-brands-2022/). Although TWC may be associated by some with certain TV personalities, trust in a brand seems different than trust in a person, or even a system. Many TWC users presumably do not know that TWC uses a combination of automated AI and humans over the loop in creating its forecasts or that it has been independently evaluated as the world's most accurate forecaster (ForecastWatch, 2021). Rather, their trust is more plausibly a consequence of the longevity, ubiquity, depth of content, and always-on nature of TWC forecast products—and perhaps its reputation for scientific integrity in its editorial pieces. Familiarity and experience with a brand appear to breed trust over time; "being there" for the ordinary days seems to lend credence to more extreme, impactful, or actionable forecasts and messages. AI does most of the routine work of making the TWC forecast, but the fact that the humans are there to provide guard rails and improvements to the AI, plus communicate unusual weather stories via alerts, insights and videos, and in phone calls and reports to commercial customers, seems to give users more trust in TWC forecasts. TWC communicates uncertainty for some variables, as with the probability of precipitation, and generally takes care to not use overly definitive language in forecast summaries. Although there is some evidence that people trust a forecast more when the provider supplies information about its uncertainty, doing so in a broad, quantitative yet easily digestible manner remains an unsolved challenge.

Communicating uncertainty can be seen as important for scientific, ethical, and practical purposes. Recent research marks progress in our understanding of methods and effects of communicating uncertainty (e.g., Gulacsik et al., 2022; Padilla et al., 2018; Spiegelhalter, 2017; van der Bles et al., 2019, 2020; see also NASEM, 2018). But for AI models, for which uncertainties can be difficult to obtain, this is relatively uncharted territory. A necessary first step will be to assess whether findings to date about the communication of uncertainty hold in the context of communicating uncertainties of AI model outputs.

7 | ADVANCING RESEARCH ON TRUST IN AI

Summarizing, the workshop yielded recommendations pertaining to four intersectional research foci that are important to advance our understanding of the trust in and trustworthiness of AI (Table 1). Note that context is a key consideration within each research focus.

These four research foci echo prior calls for attention to context and dynamics in trust research (e.g., Rousseau et al., 1998). They also align with recommendations from NASEM (2022a) but with greater emphasis on understanding the goals and roles of AI users and stakeholders in dynamic, context-specific trust processes, and on how interactions with AI and communications about model performance and uncertainties (e.g., XAI) might influence these.

The desire for a set of best practices that will engender trust in AI is reflected in the many current efforts to develop standards for trustworthy AI. Developers want to know how to design and validate systems that are more worthy of trust by professionals or lay end users who are applying those models in their decision-making but also by potentially affected communities. Ideally, developers would like to be able to anticipate what will increase perceptions of trustworthiness. AI modelers as well as some of the others attending the workshop came into the workshop with the supposition that the trustworthiness of AI models in the environmental sciences could be equated with high performance, reliability, reliance, and confidence. In light of this together with the evidence cited earlier that these are not the only determinants of trustworthiness and are contingent on, for example, contextual factors, there is a need for additional research mapping out how notions of performance, reliability, reliance, and confidence correspond to, predict, or are distinct from trust and trustworthiness.

Further, there is a need to move beyond "developing trust" as the goal; developing toward a trust ideal can be a trap (Hoffman et al., 2013) since trust is dynamic—an activity that occurs over time—rather than a state. Similar to the asymmetry of trust found in other risk contexts (Poortinga & Pidgeon, 2004; Slovic, 1993), the limited research that exists on trust in embedded AI finds a clear asymmetry in trust processes: errors (i.e., low reliability) of embedded AI reduce trust sharply, while restoring such lost trust is slow and difficult (Glikson & Woolley, 2020). Trust discussions include numerous related objectives and outcomes of interest—such as reliance, confidence, and use—that trust alone may not predict. Additional research is needed to determine how trust research agendas might be expanded to measure a broader utility of AI.

Such efforts should build on earlier related efforts (NASEM, 2018, 2021, 2022a, 2022b; NOAA SAB, 2021) but will require deeper interdisciplinary interactions (Peek & Guikema, 2021), including additional workshops and interdisciplinary sessions at AI conferences, and conferences focusing on risk, judgment and decision-making, and communications, as well as on weather, climate, and other

TABLE 1 Research recommendations to advance understanding of trust in and trustworthiness of artificial intelligence (AI).

Research focus	Information needs	Research questions
User-oriented development and co-development	How to effectively exchange knowledge between user groups and research teams while balancing their respective needs and demands, considering resource (time and funds) limitations How to ensure user-relevant AI models and systems that yield actionable output	Which users and audiences, and their needs, should developers consider across different decision contexts? When should developers start interacting with users? Practically, how often should developers and users interact during the development process? This will depend on the AI model or system being developed, but specific guidance would be useful What effects do differing ways to engage users (e.g., user-oriented development, participatory processes) have on users' trust in AI?
Understanding and measuring trust and trustworthiness	Fuller theoretical understanding of the relationship between trust and trustworthiness, and among their antecedents, moderators, and outcomes Better measurements of how trust affects evaluations of prior performance and is in turn influenced by performance over time Better measurements of the dynamics of trust in AI and the contextual dependencies of trust and trustworthiness	How do diverse decision-making strategies affect trust judgments? When and how does analytic thinking influence trust in AI and AI use decisions? What are the roles of heuristics in decision-making with AI, including new heuristics that may be specific to AI? How does trust relate to confidence, and how do these concepts influence decision-making? What factors influence users' decisions to trust, rely on, or use AI across different contexts?
Goal alignment, calibration, and standard setting	Better understanding of the strengths and limitations of trustworthiness standards for AI	How do standards influence trust? Can deficits in trust signpost how to improve standards for AI trustworthiness? Do trust and trustworthiness of AI have psychometric properties not captured by standard AI performance metrics? How can we better align the goals of users and developers, as well as of others influenced by the AI, in order to better align the AI with users?
Integrating risk and uncertainty communication research with research on trust in AI	Causal understanding of how social relationships influence trust in contexts where embedded AI is used Frameworks for developing and testing strategies for communicating uncertainties of AI model outputs	Do findings about communication of uncertainty and its effects on trust in information hold up in the context of communicating uncertainties of AI model outputs? What AI-specific types or sources of uncertainty could be communicated, and how do these affect trustworthiness and trust?

environmental sciences. This research agenda also requires revisiting mechanisms for co-production, co-design, and co-development of AI models of environmental processes, including systematic research conducted by social and behavioral scientists and engineers in HCD, but also new ways of involving users and eliciting their feedback. Possibilities include research-to-operations testbeds and similar mechanisms, visitor exchange programs between developers and users, and liaison programs, such as the satellite liaisons program (Goodman et al., 2012; Satellite Liaisons, 2013). These may require creating infrastructure for knowledge transfer and cross-training across disciplines and domains throughout all career stages and for involving actors throughout all phases of the value-driven AI model production cycle.

8 | CONCLUSION AND SUMMARY

In early 2023, the public release of ChatGPT4 amplified public debates about AI risks, with calls for immediate, precautionary risk management to make AI systems more "accurate, safe, interpretable, transparent, robust, aligned, trustworthy, and loyal" (Future of Life Institute, 2023). Yet what it means to manage AI risks and make AI systems trustworthy remains far from obvious.

The understanding of trust in and trustworthiness of AI has advanced on several fronts over the last few decades (Hoffman, 2017; NASEM, 2022a; Vereschak et al., 2021). Yet, the accelerating pace of progress on AI makes concerted investments in convergent research on trust in AI urgent, if applications and policies are to be informed by and benefit from insights from the risk and social sciences.

We offer focused research recommendations to advance the field, informed by recent progress in risk communication. These recommendations are grounded in specific considerations of how to develop trustworthy AI for environmental sciences. Our recommendations explicitly address the engagement of both AI developers and those who use and are affected by AI in this research across diverse decision contexts. To advance this research agenda will require engaging a broader set of scientists, regulators, policymakers, and users in research on trustworthy AI, dedicating more of the resources being invested in AI to this convergent research agenda and developing new research methods and infrastructure to enable progress.

ACKNOWLEDGMENTS

We gratefully acknowledge support for this work from the National Science Foundation under Grant Number ICER-201975 and the helpful insights contributed by Theodore Jensen and Phil Davis.

ORCID

Ann Bostrom https://orcid.org/0000-0002-6399-3404
Julie L. Demuth https://orcid.org/0000-0001-8918-9839
Christopher D. Wirz https://orcid.org/0000-0002-8990-5505
Mariana G. Cains https://orcid.org/0000-0002-6729-6729
Deianna Madlambayan https://orcid.org/0009-0006-1991-9481
Angela Bearth https://orcid.org/0000-0003-1270-6468
Katherine M. Crosman https://orcid.org/0000-0003-0731-2501
Imme Ebert-Uphoff https://orcid.org/0000-0001-6470-1947
David John Gagne II https://orcid.org/0000-0002-0469-2740
Seth Guikema https://orcid.org/0000-0001-6024-0303
Amy McGovern https://orcid.org/0000-0001-6675-7119
Jacob T. Radford https://orcid.org/0000-0001-6824-8967
Philippe Tissot https://orcid.org/0000-0002-2954-2378

REFERENCES

Abras, C., Maloney-Krichmar, D., & Preece, J. (2004). User-centered design. In W. Bainbridge (Ed.), *Encyclopedia of human-computer interaction*. Sage Publications.

Abbasi, A., Sarker, S., & Chiang, R. (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, *17*(2), I–XXXII. https://doi.org/10.17705/1jais.00423

Ammanath, B. (2022). *Trustworthy AI: A business guide for navigating trust and ethics in AI*. John Wiley & Sons.

Arribas, A., Robertson, K. B., & Mylne, K. R. (2005). Test of a poor man's ensemble prediction system for short-range probability forecasting. *Monthly Weather Review*, *133*(7), 1825–1839.

Ashoori, M., & Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. arXiv preprint arXiv:1912.02675.

Balog-Way, D., McComas, K., & Besley, J. (2020). The evolving field of risk communication. *Risk Analysis*, *40*(S1), 2240–2262.

Bauer, P. C., & Freitag, M. (2017). *Measuring Trust*. Oxford Handbooks Online. https://doi.org/10.1093/oxfordhb/9780190274801.013.1

Bearth, A., & Siegrist, M. (2022). The social amplification of risk framework: A normative perspective on trust? *Risk Analysis*, *42*(7), 1381–1392.

Billet, J., DeLisi, M., Smith, B., & Gates, C. (1997). Use of regression techniques to predict hail size and the probability of large hail. *Weather and Forecasting*, *12*, 154–164. https://doi.org/10.1175/1520-0434(1997)012⟨0154:UORTTP⟩2.0.CO;2

Bjerknes, G., & Bratteteig, T. (1995). User participation and democracy: A discussion of Scandinavian research on system development. *Scandinavian Journal of Information Systems*, *7*(1), Article 1.

Broniatowski, D. (2021). *Psychological foundations of explainability and interpretability in artificial intelligence*. (NIST Interagency/Internal Report [NISTIR] 8367).National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8367

Carroll, J. M. (1996). Encountering others: Reciprocal openings in participatory design and user-centered design. *Human–Computer Interaction*, *11*(3), 285–290.

Chiou, E. K., & Lee, J. D. (2021). Trusting automation: Designing for responsivity and resilience. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *65*(1), 001872082110099. https://doi.org/10.1177/00187208211009995

Contractor, N. S., & DeChurch, L. A. (2014). Integrating social networks and human social motives to achieve social influence at scale. *Proceedings of the National Academy of Sciences*, *111*(Suppl4), 13650–13657.

Custis, C. (2022). Partnership on AI presentation, Plenary on Ethical and Responsible AI, 4th NOAA Workshop on Leveraging AI in Environmental Sciences, September 6–9, 2022, virtual.

Cvetkovich, G., & Nakayachi, K. (2007). Trust in a high-concern risk controversy: A comparison of three concepts. *Journal of Risk Research*, *10*(2), 223–237.

de Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012, September). The world is not enough: Trust in cognitive agents. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. *56*, No. 1, pp. 263–267). Sage CA: Los Angeles, CA: Sage Publications.

de Visser, E. J., Monfort, S. S., Goodyear, K., Lu, L., O'Hara, M., Lee, M. R., Parasuraman, R., & Krueger, F. (2017). A little anthropomorphism goes a long way: Effects of oxytocin on trust, compliance, and team performance with automated agents. *Human factors*, *59*(1), 116–133.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349.

Demuth, J. L., DeMaria, M., & Knaff, J. A. (2006). Improvement of Advanced Microwave Sounding Unit tropical cyclone intensity and size estimation algorithms. *Journal of Applied Meteorology and Climatology*, *45*(11), 1573–1581.

Dorton, S. L. (2022). Supradyadic trust in artificial intelligence. *Artificial Intelligence and Social Computing*, *28*, 92–100.

Earle, T. C. (2010). Trust in risk management: A model-based review of empirical research. *Risk Analysis: An International Journal*, *30*(4), 541–574.

Earle, T., Siegrist, M., & Gutscher, H. (2010). Trust, risk perception and the TCC model of cooperation. In M. Siegrist, T. C. Earle, & H. Gutscher (Eds.), *Trust in risk management: Uncertainty and scepticism in the public mind, 1-49*. Earthscan.

Earle, T. C. (2009). Trust, Confidence, and the 2008 Global Financial Crisis. *Risk Analysis*, *29*(6), 785–792. Portico. https://doi.org/10.1111/j.1539–6924.2009.01230.x

Earle, T. C., & Siegrist, M. (2006). Morality information, performance information, and the distinction between trust and confidence. *Journal of Applied Social Psychology*, *36*(2), 383–416.

Earle, T., & Siegrist, M. (2008). Trust, confidence and cooperation model: a framework for understanding the relation between trust and risk perception. *International Journal of Global Environmental Issues*, *8*(1–2), 17–29.

European Commission (EC). (2020). *White paper on Artificial Intelligence—A European approach to excellence and trust* [White paper]. European Commission, Brussels, 19.2.2020, COM(2020) 65 final. https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

European Commission High-Level Expert Group on AI (EC AI HLEG). (2019). *Ethics guidelines for trustworthy artificial intelligence*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission High-Level Expert Group on AI (EC AI HLEG). (2022). The Assessment List for Trustworthy AI (ALTAI) for self-assessment. https://doi.org/10.2759/002360

Exec. Order No. 13960: Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government. (2020). 85 Fed. Reg. 78939.

Fine, G. A., & Holyfield, L. (1996). Secrecy, trust, and dangerous leisure: Generating group cohesion in voluntary organizations. *Social psychology quarterly*, 22–38.

Fischhoff, B., Lichtenstein, S., Derby, S. L., Slovic, P., & Keeney, R. (1984). *Acceptable risk*. Cambridge University Press.

Fischhoff, B., & Broomell, S. B. (2020). Judgment and decision making. *Annual Review of Psychology*, *71*(1), 331–355.

Fischhoff, B. (1995). Risk perception and communication unplugged: twenty years of process 1. *Risk analysis*, *15*(2), 137–145.

ForecastWatch. (2021). *Global and regional weather forecast accuracy overview: 2017-2020*. https://www.forecastwatch.com/AccuracyOverview2017-2020

Future of Life Institute (2023). Pause giant AI experiments: An open letter. Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Gagne, D. J., McGovern, A., Haupt, S., Sobash, R., Williams, J., & Xue, M. (2017). Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Weather Forecasting*, *32*, 1819–1840. https://doi.org/10.1175/WAF-D-17-0010.1

Gallup. (2019). *Wellcome Global Monitor—First wave findings*. https://wellcome.org/sites/default/files/wellcome-global-monitor-2018.pdf

Glahn, H. R., & Dallavalle, J. P. (2000). *MOS-2000. National Weather Service Techniques Development Laboratory, TDL Office Note 00-1*. https://www.mdl.nws.noaa.gov/~qa/pdf_files/TDL_Office_Note_00-1.pdf

Glahn, B., & Im, J. S. (2011). *Algorithms for effect objective analysis of surface weather variables*. https://www.nws.noaa.gov/mdl/pubs/Documents/Papers/GlahnAndIm_ams_seattle_2011.pdf

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Goodman, S. J., Gurka, J., DeMaria, M., Schmit, T. J., Mostek, A., Jedlovec, G., Siewert, C., Feltz, W., Gerth, J., Brummer, R., Miller, S., Reed, B., & Reynolds, R. R. (2012). The GOES-R proving ground: Accelerating user readiness for the next-generation geostationary environmental satellite system. *Bulletin of the American Meteorological Society*, *93*(7), 1029–1040.

Guikema, S. D., Nateghi, R., Quiring, S. M., Staid, A., Reilly, A. C., & Gao, M. (2014). Predicting hurricane power outages to support storm response planning. *IEEE Access*, *2*, 1364–1373.

Gulacsik, G., Joslyn, S. L., Robinson, J., & Qin, C. (2022). Communicating uncertainty information in a dynamic decision environment. *Weather, Climate, and Society*, *14*(4), 1201–1216.

Haynes, A. B., Berry, W. R., & Gawande, A. A. (2015). What do we know about the safe surgery checklist now? *Annals of Surgery*, *261*(5), 829–830.

Henderson, J., Spinney, J., & Demuth, J. L. (2023). Conceptualizing confidence: A multisited qualitative analysis in a severe weather context. *Bulletin of the American Meteorological Society*, *104*(2), E459–E479. https://doi.org/10.1175/BAMS-D-22-0137.1

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hoffman, R. R. (2017). A taxonomy of emergent trusting in the human–machine relationship. In P. J. Smith, & R. R. Hoffman (Eds.), *Cognitive systems engineering: The future for a changing world* (pp. 137–163). Taylor and Francis.

Hoffman, R. R., Deal, S. V., Potter, S., & Roth, E. M. (2010). *The practitioner's cycles, Part 2: Solving envisioned world problems*. IEEE Intelligent Systems, 25(3), 6–11.

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems*, *28*(1), 84–88. https://doi.org/10.1109/MIS.2013.24

Hoffman, R. R., Klein, G., Mueller, S. T., Jalaeian, M., & Tate, C. (2021). *The stakeholder playbook for explaining AI systems* (Technical Report). DARPA Explainable AI Program. https://osf.io/preprints/psyarxiv/9pqez/

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). *Measuring trust in the XAI context* (Technical Report). DARPA Explainable AI Program. https://psyarxiv.com/e3kv9/download?format=pdf

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). *Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI*. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 624-635). https://doi.org/10.1145/3442188.3445923

Jensen, T., Khan, M. M. H., & Albayram, Y. (2020, July). The role of behavioral anthropomorphism in human-automation trust calibration. In International Conference on Human-Computer Interaction (pp. 33–53). Cham: Springer International Publishing.

Johnson, B. B. (1999). Exploring dimensionality in the origin of hazard-related trust. *Journal of Risk Research*, *2*, 325–354.

Johnson, B. B., & Rickard, L. N. (2023). Trust, confidence, familiarity, and support for land-based recirculating aquaculture facilities. *Risk Analysis*, *43*(7), 1339–1355.

Jones, J. M. (2022, July 5). Confidence in U.S. institutions down; average at new low. *Gallup*. https://news.gallup.com/poll/394283/confidence-institutions-down-average-new-low.aspx

Kabir, E., Guikema, S. D., & Quiring, S. M. (2019). Predicting thunderstorm-induced power outages to support utility restoration. *IEEE Transactions on Power Systems*, *34*(6), 4370–4381.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2021). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, *65*(2), 187208211013988. https://doi.org/10.1177/00187208211013988

Kim, S. Y., Schmitt, B. H., & Thalmann, N. M. (2019). Eliza in the uncanny valley: Anthropomorphizing consumer robots increases their perceived warmth but decreases liking. *Marketing Letters*, *30*, 1–12.

Kitzmiller, D., McGovern, W., & Saffle, R. (1995). The WSR-88D severe weather potential algorithm. *Weather and Forecasting*, *10*, 141–159. https://doi.org/10.1175/1520-0434(1995)010⟨0141:TWSWPA⟩2.0.CO;2

Klein, G., Feltovich, P. J., Bradshaw, J. M., & Woods, D. D. (2005). Common ground and coordination in joint activity. *Organizational Simulation*, *53*, 139–184.

Klein, Moon, & Hoffman (2006). Making Sense of Sensemaking 2: A Macrocognitive Model. *IEEE Intelligent Systems*, *21*(5), 88–92. https://doi.org/10.1109/mis.2006.100

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

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. *Academy of Management Review*, *23*(3), 438–458.

Lewis, P. R., & Marsh, S. (2022). What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research*, *72*, 33–49.

Li, H., Treinish, L. A., & Hosking, J. R. M. (2010). A statistical model for risk management of electric outage forecasts. *IBM Journal of Research and Development*, *54*(3), 8:1–8:11. https://doi.org/10.1147/JRD.2010.2044836

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin and Review*, *127*(2), 267–286.

Lockey, S., Gillespie, N., Holm, D., & Someh, I. A. (2021, January). A review of trust in artificial intelligence: challenges, vulnerabilities and future directions. In Proceedings of the Annual Hawaii International Conference on System Sciences (Vol. *2020*, pp. 5463–5472). Hawaii International Conference on System Sciences.

MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., & Gahegan, M. (2012). Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2496–2505.

Malone, T. (1955). Application of statistical methods in weather prediction. *Proceedings of the National Academy of Sciences USA*, *41*, 806–815. https://doi.org/10.1073/pnas.41.11.806

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review. Academy of Management*, *20*(3), 709–734. https://doi.org/10.2307/258792

McCright, A. M., Dentzman, K., Charters, M., & Dietz, T. (2013). The influence of political ideology on trust in science. *Environmental Research Letters*, *8*, 044029. https://doi.org/10.1088/1748-9326/8/4/044029

McGovern, A., Bostrom, A., Davis, P., Demuth, J. L., Ebert-Uphoff, I., He, R., Hickey, J., Gagne, D. J. II, Snook, N., Stewart, J. Q., Thorncroft, C.,

Tissot, P., & Williams, J. K. (2022). NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES). *Bulletin of the American Meteorological Society*, *103*(7), E1658–E1668. https://doi.org/10.1175/BAMS-D-21-0020.1

McGovern, A., Ebert-Uphoff, I., Gagne, D. J., & Bostrom, A. (2022). Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environmental Data Science*, 1, e6.

McGovern, A., Lagerquist, R., ii Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. https://doi.org/10.1175/bams-d-18-0195.1

McGovern, A., Gagne, D. J., Williams, J. K., Brown, R. A., & Basara, J. B. (2014). Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. *Machine learning*, *95*, 27–50.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, GA (pp. 220–229).

Moon, Y. (2000). Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of consumer research*, *26*(4), 323–339.

Morgan, M. G., Fischhoff, B., Bostrom, A., & Atman, C. J. (2002). *Risk Communication: A Mental Models Approach*. Cambridge Books.

Morkes, J., Kernal, H. K., & Nass, C. (1999). Effects of humor in task-oriented human-computer interaction and computer-mediated communication: A direct test of SRCT theory. *Human-Computer Interaction*, *14*(4), 395–435.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv preprint arXiv:1902.01876.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, *27*(5–6), 527–539.

National Academies of Sciences, Engineering, and Medicine. (2021). *Assessing and improving AI trustworthiness: Current contexts and concerns: Proceedings of a Workshop—In Brief. The National Academies Press*. https://doi.org/10.17226/26208

National Academies of Sciences, Engineering, and Medicine. (2017). *Communicating science effectively: A research agenda*. The National Academies Press. https://doi.org/10.17226/23674

National Academies of Sciences, Engineering, and Medicine. (2022a). *Human-AI teaming: State-of-the-art and research needs*. The National Academies Press. https://doi.org/10.17226/26355

National Academies of Sciences, Engineering, and Medicine. (2018). *Integrating social and behavioral sciences within the weather enterprise*. The National Academies Press. https://doi.org/10.17226/24865

National Academies of Sciences, Engineering, and Medicine. (2022b). *Machine learning and artificial intelligence to advance earth system science: Opportunities and challenges: Proceedings of a Workshop*. The National Academies Press. https://doi.org/10.17226/26566

National Research Council. (1996). *Understanding risk: Informing decisions in a democratic society*. The National Academies Press. https://doi.org/10.17226/5138

National Oceanic and Atmospheric Administration (NOAA) Science Advisory Board. (2021). A report on priorities for weather research (NOAA Science Advisory Board Report). NOAA Science Advisory Board. https://sab.noaa.gov/wp-content/uploads/2021/12/PWR-Report_Final_12-9-21.pdf

National Science and Technology Council Select Committee on Artificial Intelligence (NSTC). (2019). *National Artificial Intelligence Research and Development Strategic Plan: 2019 Update* (PUBID-06-21-2019-001-01). NSTC. https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf

National Science Foundation 20-503. (2019). *National Artificial Intelligence (AI) Research Institutes: Accelerating research, transforming society, and growing the American workforce*. https://www.nsf.gov/pubs/2020/nsf20503/nsf20503.htm

Nass, C., Steuer, J., & Tauber, E. R. (1994, April). Computers are social actors. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 72–78).

Norman, D. A. (2005). Human-centered design considered harmful. *Interactions*, *12*(4), 14–19.

Norman, D. A. (1988). *The psychology of everyday things*. Basic Books.

Novak, D. R., Bright, D. R., & Brennan, M. J. (2008). Operational forecaster uncertainty needs and future roles. *Weather and Forecasting*, *23*(6), 1069–1084.

Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: A cognitive framework across disciplines. *Cognitive Research: Principles and Implications*, *3*(1), 29.

Padilla, L. M., Powell, M., Kay, M., & Hullman, J. (2021). Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology*, *11*, 579267.

Padilla, L., Kay, M., & Hullman, J. (2023). Uncertainty Visualization. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J.L. Teugels (Eds.), *Wiley StatsRef: Statistics Reference Online*.

Peek, L., & Guikema, S. (2021). Interdisciplinary theory, methods, and approaches for hazards and disaster research: An introduction to the special issue. *Risk Analysis*, *41*(7), 1047–1058.

Peek, L., & Guikema, S. (2021). Interdisciplinary theory, methods, and approaches for hazards and disaster research: An introduction to the special issue. *Risk Analysis*, *41*(7), 1047–1058.

Pollack, M., & Hirschberg, J., Weber, B. (1982). User participation in the reasoning processes of expert systems. *Proceedings of the AAAI-82*, Pittsburgh, PA (pp. 358–361).

Poortinga, W., & Pidgeon, N. F. (2003). Exploring the dimensionality of trust in risk regulation. *Risk Analysis: An International Journal*, *23*(5), 961–972.

Poortinga, W., & Pidgeon, N. F. (2004). Trust, the asymmetry principle, and the role of prior beliefs. *Risk Analysis: An International Journal*, *24*(6), 1475–1486.

Poortinga, W., & Pidgeon, N. F. (2005). Trust in risk regulation: Cause or consequence of the acceptability of GM food? *Risk Analysis: An International Journal*, *25*(1), 199–209.

Poortinga, W., & Pidgeon, N. F. (2006). Prior attitudes, salient value similarity, and dimensionality: Toward an integrative model of trust in risk regulation 1. *Journal of Applied Social Psychology*, *36*(7), 1674–1700.

Prior, T. D., & Paton, D. (2006). Understanding the context: The value of community engagement in bushfire risk communication and education. Observations following the East Coast Tasmania bushfires of December 2006. *Australiasian Journal of Disaster and Trauma Studies*, *2*(1), 1–14.

Quiring, S. M., Schumacher, A. B., & Guikema, S. D. (2014). Incorporating hurricane forecast uncertainty into a decision-support application for power outage modeling. *Bulletin of the American Meteorological Society*, *95*(1), 47–58.

Renn, O. (2008a). Concepts of risk: An interdisciplinary review–Part 1: Disciplinary risk concepts. *GAIA-Ecological Perspectives for Science and Society*, *17*(1), 50–66.

Renn, O. (2008b). Concepts of risk: An interdisciplinary review–Part 2: Integrative approaches. *GAIA-Ecological Perspectives for Science and Society*, *17*(2), 196–204.

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places. Center for the Study of Language and Information*. Cambridge University Press.

Roth, E. M., Bennett, K. B., & Woods, D. D. (1987). Human interaction with an "intelligent" machine. *International Journal of Man-Machine Studies*, *27*(5-6), 479–525.

Roth, E. M., DePass, B., Harter, J., Scott, R., & Wampler, J. (2018). Beyond levels of automation: Developing more detailed guidance for

human automation interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 150–154.
Roth, E., Klein, D., Sushereba, C., Ernst, K., & Militello, L. (2022). Methods and Measures to Evaluate Technologies that Influence Aviator Decision Making and Situation Awareness. Roth Cognitive Engineering, Marimo Consulting, and Applied Decision Science, for the United States Army Aeromedical Research Laboratory, Fort Rucker, AL, USA. USAARL-TECH-CR-2022–22.
Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Introduction to special topic forum: Not so different after all: A cross-discipline view of trust. *Academy of Management Review. Academy of Management*, *23*(3), 393–404. http://www.jstor.org/stable/259285
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x
Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, *16*, 1–85.
Saßmannshausen, T., Burggräf, P., Hassenzahl, M., & Wagner, J. (2023). Human trust in otherware—A systematic literature review bringing all antecedents together. *Ergonomics*, *66*(7), 976–998. https://doi.org/10.1080/00140139.2022.2120634
Satellite Liaisons. (2013). *About the blog*. https://satelliteliaisonblog.com/about/
Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, *58*(3), 377–400.
Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (NIST) Special Publication (1270). National Institute of Standards and Technology, U.S. Department of Commerce. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf
Schwarz, N., & Clore, G. L. (2003). Mood as Information: 20 Years Later. *Psychological Inquiry*, *14*(3–4), 296–303. https://doi.org/10.1080/1047840x.2003.9682896
Seong, Y., & Bisantz, A. M. (2002, September). Judgment and trust in conjunction with automated decision aids: A theoretical model and empirical investigation. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. *46*, No. 3, pp. 423–427). Sage CA: Los Angeles, CA: SAGE Publications.
Siegrist, M. (2010). Trust and Confidence: The Difficulties in Distinguishing the Two Concepts in Research. *Risk Analysis*, *30*(7), 1022–1024. https://doi.org/10.1111/j.1539–6924.2010.01454.x
Siegrist, M. (2021). Trust and risk perception: A critical review of the literature. *Risk Analysis*, *41*(3), 480–490.
Singer, S. J., Kellogg, K. C., Galper, A. B., & Viola, D. (2022). Enhancing the value to users of machine learning-based clinical decision support tools: A framework for iterative, collaborative development and implementation. *Health Care Management Review*, *47*(2), E21–E31. https://doi.org/10.1097/HMR.0000000000000324
Singhee, A., Li, Z., Koc, A., Wang, H., Cipriani, J. P., Kim, Y., Kumar, A. P., Treinish, L. A., Mueller, R., Labut, G., Foltman, R. A., & Gauthier, G. M. (2016). OPRO: Precise emergency preparedness for electric utilities. *IBM Journal of Research and Development*, *60*(1), 6:1–6:15. 10.1147/JRD.2015.2494999
Slovic, P. (1993). Perceived risk, trust, and democracy. *Risk Analysis*, *13*(6), 675–682.
Slovic, P. E. (2000). *The perception of risk*. Earthscan Publications.
Sobash, R. A., Schwartz, C. S., Romine, G. S., Fossell, K. R., & Weisman, M. L. (2016). Severe weather prediction using storm surrogates from an ensemble forecasting system. *Weather and Forecasting*, *31*(1), 255–271.
Spiegelhalter, D. (2017). Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, *4*(1), 31–60.
Stanton, B., & Jensen, T. (2021). *Trust and artificial intelligence* (NIST Interagency/Internal Report (NISTIR) 8332). National Institute of Standards and Technology. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087
Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST Trustworthy and Responsible AI, National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
Tullio, J., Dey, A. K., Chalecki, J., & Fogarty, J. (2007, April). How it works: a field study of non-technical users interacting with an intelligent system. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 31–40).
van der Bles, A. M., van der Linden, S., Freeman, A. L. J., Mitchell, J., Galvao, A. B., Zaval, L., & Spiegelhalter, D. J. (2019). Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, *6*, 181870. https://doi.org/10.1098/rsos.181870
van der Bles, A. M., van der Linden, S., Freeman, A. L., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences*, *117*(14), 7672–7683.
Varshney, K. R. (2022). *Trustworthy machine learning*. Independently Published. http://www.trustworthymachinelearning.com/trustworthymachinelearning.pdf
Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–39.
White House Office of Science and Technology Policy. (2022, October). Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. The White House. https://www.whitehouse.gov/ostp/ai-bill-of-rights/
White, M. P., & Eiser, J. R. (2006). Marginal trust in risk managers: Building and losing trust following decisions under uncertainty. *Risk analysis*, *26*(5), 1187–1203.
White, M. P., & Johnson, B. B. (2010). The intuitive detection theorist (IDT) model of trust in hazard managers. *Risk Analysis: An International Journal*, *30*(8), 1196–1209.
Wirz, C. D., Demuth, J. L., Bostrom, A., Cains, M. G., Ebert-Uphoff, I., Gagne, D. J. II, Schumacher, A., McGovern, A., & Madlambayan, D. (2023). *(Re)Conceptualizing trustworthy AI as perceptual and context-dependent* (Working paper).
Williams, J. K. (2014). Using random forests to diagnose aviation turbulence. *Mach Learn*, *95*, 51–70.
Wormwood, J. B., Siegel, E. H., Kopec, J., Quigley, K. S., & Barrett, L. F. (2019). You are what I feel: A test of the affective realism hypothesis. *Emotion*, *19*(5), 788–798. https://doi.org/10.1037/emo0000484
Yousefzadeh, R., & Cao, X. (2022). To *what extent should we trust AI models when they extrapolate?* https://doi.org/10.48550/arXiv.2201.11260

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bostrom, A., Demuth, J. L., Wirz, C. D., Cains, M. G., Schumacher, A., Madlambayan, D., Bansal, A. S., Bearth, A., Chase, R., Crosman, K. M., Ebert-Uphoff, I., Gagne, D. J., Guikema, S., Hoffman, R., Johnson, B. B., Kumler-Bonfanti, C., Lee, J. D., Lowe, A., McGovern, A., … Williams, J. K. (2024). Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences. *Risk Analysis*, *44,* 1498–1513. https://doi.org/10.1111/risa.14245