NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES)

Amy McGovern, Ann Bostrom, Phillip Davis, Julie L. Demuth, Imme Ebert-Uphoff, Ruoying He, Jason Hickey, David John Gagne II, Nathan Snook, Jebb Q. Stewart, Christopher Thorncroft, Philippe Tissot, John K. Williams

a University of Oklahoma, Norman OK
b University of Washington, Seattle WA
c Del Mar College, Corpus Christi TX
d National Center for Atmospheric Research, Boulder CO
e Colorado State University, Fort Collins CO
f North Carolina State University, Raleigh NC
 g Google, Mountain View CA
h National Oceanic and Atmospheric Administration (NOAA), Boulder CO
i University at Albany, Albany NY
j Texas A&M Corpus Christi, Corpus Christi TX
k The Weather Company, an IBM Business, Andover MA

Corresponding author: Amy McGovern, amcgovern@ou.edu

Early Online Release: This preliminary version has been accepted for publication in Bulletin of the American Meteorological Society, may be fully cited, and has been assigned DOI 10.1175/BAMS-D-21-0020.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2022 American Meteorological Society
ABSTRACT: We introduce the National Science Foundation (NSF) AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES). This AI institute was funded in 2020 as part of a new initiative from the NSF to advance foundational AI research across a wide variety of domains. To date AI2ES is the only NSF AI institute focusing on environmental science applications. Our institute focuses on developing trustworthy AI methods for weather, climate, and coastal hazards. The AI methods will revolutionize our understanding and prediction of high-impact atmospheric and ocean science phenomena and will be utilized by diverse, professional user groups to reduce risks to society. In addition, we are creating novel educational paths, including a new degree program at a community college serving underrepresented minorities, to improve workforce diversity for both AI and environmental science.
1. AI Needs for Environmental Science

While Artificial Intelligence (AI) has demonstrably improved prediction and understanding of many environmental science phenomena (e.g., Ahijevych et al. 2016; Williams et al. 2016; McGovern et al. 2017; Gagne et al. 2017; Gagne et al. 2019; Lagerquist et al. 2019a; Barnes et al. 2019; Reichstein et al. 2019; Boukabara et al. 2021), there is often a lack of trust by environmental science decision-makers when it comes to relying on “black box” algorithms, especially in life-or-death situations (Karstens et al. 2018; Demuth et al. 2020). Developing AI that is trustworthy and useful for environmental risk management requires fundamental natural, mathematical, and social sciences research on the AI needs and perceptions of key users. These users’ judgments and decisions may depend on their expertise and context (Larkin et al. 1980; Chi et al. 1981; Payne et al. 1992). Such research should include a users’ understanding and perceptions of the AI method, its performance, and other factors emerging in empirical and theoretical research on AI (Mueller et al. 2019; Wang et al. 2019; Glikson and Woolley 2020).

We introduce the NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES), a national AI institute that conducts convergent research focused on creating trustworthy AI for the weather, climate, and ocean communities. AI2ES seeks to uniquely benefit humanity by developing novel physically-based AI techniques that are demonstrated to be trustworthy, and to directly improve prediction, understanding, and communication of high-impact environmental hazards.

Developing trustworthy AI, particularly for the weather, water, and climate communities, is an urgent and timely priority (Masson-Delmotte et al. 2018; Reidmiller et al. 2018; ERISS Corporation and The Maritime Alliance 2019) at the highest levels of government and industry. NOAA has identified AI as a high priority in their strategic AI plan (NOAA 2019). Similarly, the White House continues to prioritize the development of innovative AI (National Science and Technology Council 2019b; Office of Science and Technology Policy 2019; National Science and Technology Council 2019a); the National Academies of Science, Engineering and Medicine cites improved forecasting of extreme events as a critical task (National Academies of Sciences, Engineering, and Medicines 2016); and NOAA and the National Weather Service (NWS) have increasingly focused on providing impact-based decision support services (IDSS) to reduce weather risks (Uccellini and Hoeve 2019), which requires developing new and improved forecast information to meet IDSS...
needs (Demuth et al. 2020). The European Centre for Medium-Range Weather Forecasts places high importance on improving predictions and understanding (European Center for Medium-range Weather Forecasts 2016). The recently released ECMWF 10 year plan for AI states that “We anticipate that it will be increasingly difficult to distinguish between scientists working on machine learning and domain scientists in the future” (Düben et al. 2021).

AI algorithms for the weather, water, and climate communities must not only be skillful, but also trustworthy. The European Commission’s report on Ethics Guidelines for Trustworthy AI notes that “AI systems need to be human-centric, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom” (European Commission 2019). AI2ES directly incorporates risk communication to connect the development of trustworthy AI to the human decision makers AI2ES aims to serve.

Before we proceed, a comment is in order regarding the terms Machine Learning (ML) versus Artificial Intelligence (AI). Machine learning is focused on the development of algorithms that allow computers to learn from data how to perform certain tasks without explicit programming. In contrast, artificial Intelligence is a much broader concept that seeks to create algorithms providing human-like reasoning abilities. For the applications considered here we are only concerned with the machine learning type of algorithms. From here on we use both terms, ML and AI, interchangeably, always with the understanding that AI only refers to the machine learning style of AI.

2. Building Trustworthy AI - Key components of AI2ES

Figure 1 provides an overview of the key components of AI2ES. Our research cycle integrates development of foundational new AI/ML methods, working with atmospheric scientists and risk communication researchers in a virtuous cycle where each thread informs the others. Specifically we seek to develop trustworthy AI approaches for environmental science by:

1. Integrating risk communication research to determine which AI and eXplainable (XAI) features promote trustworthiness and use by different user groups, such as forecasters, for managing risk;

2. Leveraging and expanding approaches from physics-based AI and XAI;

3. Raising awareness about negative side effects that AI has caused in other domains and helping the community to proactively avoid those;
The European Centre for Medium-Range Weather Forecasts places high importance on improving predictions and understanding (European Center for Medium-range Weather Forecasts 2016). The recently released ECMWF 10 year plan for AI states that “We anticipate that it will be increasingly difficult to distinguish between scientists working on machine learning and domain scientists in the future” (Düben et al. 2021).

AI algorithms for the weather, water, and climate communities must not only be skillful, but also trustworthy. The European Commission’s report on Ethics Guidelines for Trustworthy AI notes that “AI systems need to be human-centric, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom” (European Commission 2019). AI2ES directly incorporates risk communication to connect the development of trustworthy AI to the human decision makers AI2ES aims to serve.

Before we proceed, a comment is in order regarding the terms Machine Learning (ML) versus Artificial Intelligence (AI). Machine learning is focused on the development of algorithms that allow computers to learn from data how to perform certain tasks without explicit programming. In contrast, artificial Intelligence is a much broader concept that seeks to create algorithms providing human-like reasoning abilities. For the applications considered here we are only concerned with the machine learning type of algorithms. From here on we use both terms, ML and AI, interchangeably, always with the understanding that AI only refers to the machine learning style of AI.

2. Building Trustworthy AI - Key components of AI2ES

Figure 1 provides an overview of the key components of AI2ES. Our research cycle integrates development of foundational new AI/ML methods, working with atmospheric scientists and risk communication researchers in a virtuous cycle where each thread informs the others. Specifically we seek to develop trustworthy AI approaches for environmental science by:

1. Integrating risk communication research to determine which AI and XAI features promote trustworthiness and use by different user groups, such as forecasters, for managing risk;
2. Leveraging and expanding approaches from physics-based AI and XAI;
3. Raising awareness about negative side effects that AI has caused in other domains and helping the community to proactively avoid those;
4. Using a variety of specific use cases (applications) to guide, test and disseminate the AI approaches;
5. Taking a multi-sector approach, involving academia, government agencies, and the private sector from the very beginning to facilitate maximal transition of the developed approaches from research to operations.
6. Creating new education pathways for AI, environmental science, and risk communication at all levels to improve workforce development and participation.

In the remainder of this article we give a quick overview of each of these components.

3. Convergent, virtuous cycle involving risk communication research with users

AI2ES is focused on developing trustworthy AI for professional users, such as weather forecasters, emergency managers, transportation officials, ecological and water resource managers. These
professionals are close, direct users of AI forecast information, and their job responsibilities involve assessing risk and making decisions that have critical consequences for people’s well-being. It is therefore important to study their AI interpretations, perceptions, uses, and needs, as situated in their varied decision-making contexts, in order to guide development and refinement of AI forecast information that is trustworthy.

The risk communication research in AI2ES brings to bear risk analysis, perception, and communication research theories and methods, with the goals of understanding how different features of AI and XAI influence trustworthiness of, trust in, and willingness to use AI guidance (e.g., Jacovi et al. 2021; Mueller et al. 2019; Glikson and Woolley 2020). This research is being conducted through structured interviews, experiments, and surveys in naturalistic settings for (a) different professional user groups, (b) across different weather (severe convection, winter weather, tropical cyclones), coastal, and ocean hazards, and (c) with different AI and XAI techniques. For example, in our initial, structured interviews, we are evaluating how weather forecasters’ perceived trustworthiness is influenced by the training approaches and datasets used for developing AI guidance, different bulk and case-study verification statistics, who developed the guidance, and interactivity with the output (including the ability to interrogate a source about how an answer was derived). Additional work will be conducted with XAI techniques to evaluate users’ perceptions of, for example, algorithmic transparency, variable importance, and visualization techniques.

Conducting such user-oriented research is essential for multiple reasons. It will guide development of AI features and provision of AI guidance that aligns with users’ key decision-making needs and contexts. Further, it will help improve users’ evaluation and use of AI models and output, which intrinsically have limitations as all models do, to increase trust and use when warranted.

4. Leveraging Physics-Based AI and Explainable AI

In addition to ensuring that the AI is skillful, AI2ES is developing new explainable and interpretable AI techniques (McGovern et al. 2019) focused on the needs of environmental science end-users, developing new approaches to integrating the laws of physics into AI methods, developing novel ways to quantify and communicate model uncertainty, and ensuring that the AI is robust to natural and adversarial variations of AI model inputs.
As part of improving trust, many end-users want to understand what the model has learned and if it is physically plausible (Jacovi et al. 2021). Explainable AI (XAI) provides an approach to looking inside the “black box” of ML models (Molnar 2018; McGovern et al. 2019), but many existing XAI approaches do not adequately address environmental science phenomena such as spatiotemporal events. As noted above, our risk communication research agenda includes exploring professional users’ perceptions and interpretations of XAI techniques applied to these and other similar contexts. Additional research on what constitutes good XAI from users’ and human-AI teaming perspectives is needed (for examples and discussion of this see e.g., Lu et al. 2020; Klein et al. 2021; Mueller et al. 2021; National Academies of Sciences, Engineering, and Medicine 2021; Schwalbe and Finzel 2021). We are developing new XAI approaches as well as investigating interpretable AI methods (Rudin 2019), which are designed from the start to be more human-understandable models. We are also ensuring that the AI and XAI approaches follow the laws of physics (e.g., Karniadakis et al. 2021; Jia et al. 2018, 2021), helping to further improve trust by ensuring the models cannot predict scenarios which are impossible and cannot learn non-physical relationships for prediction (e.g., Yuval and O’Gorman 2020; Gettelman et al. 2021; Beucler et al. 2021a; Yuval et al. 2021). Murphy (1993) identifies three components of a good forecast: quality, or how well the forecast corresponds with observations; consistency, or how well the forecast corresponds with the forecaster’s prior judgment; and value, or how much the forecast benefits users. Standard ML models focus on optimizing quality by fitting closely to observations often at the expense of consistency in space and time. Prior physical constraints on different parts of the ML pipeline have shown promise at improving consistency while maintaining or even further improving quality (e.g., Beucler et al. 2021b; Willard et al. 2020). While one could focus entirely on model skill, previous work has shown that forecasters and other scientific end-users prefer a model based on physics.

5. Ethical and Responsible AI and Trust

“If not us, then who? If not now, then when? If not here, then where?” These inspiring words by Yeb Saño were spoken at the 2012 UN climate summit regarding the need to come together and combat climate change. We embrace these words to ensure that AI will be used in a responsible way in the environmental sciences. Namely, as the NSF center for creating trustworthy AI for environmental science, we are taking a lead on creating awareness and guidelines for the ethical
and responsible use of AI for the weather, climate, and ocean (McGovern et al. 2021). Not only is it integral to the creation of trustworthy AI, but it is critical that we ensure that we are cognizant of and avoid unintentional negative consequences resulting from the introduction of AI. This in turn will avoid inadvertently creating - or increasing - environmental injustice through the use of AI.

Although it might seem as if the environmental sciences, due to their use of scientific methods and observations, are immune to the danger of increasing bias through AI, that is not the case. For example, while it might seem that the weather affects everyone equally, the effects of extreme weather are felt disproportionately by vulnerable communities and individuals (Environmental Protection Agency 2021). Likewise, if a tool uses data that is not available in areas where the predictions would be most needed, additional care needs to be taken to develop approaches that would work for all areas. It is well known that AI algorithms tend to reinforce and solidify unintentional biases in data (O’Neil 2016; Benjamin 2019). Given that we know there are existing unintentional biases in weather data, such as the population biases shown in hail and tornado reports (Allen and Tippett 2015; Potvin et al. 2019), one of the goals of AI2ES is to ensure that AI developers for weather, climate, and ocean applications have the knowledge and tools to create AI that can counteract these effects, to make the AI both ethical and responsible and to minimize bias. For example, we aim to develop a tool that would identify potential biases in data automatically to facilitate the developer counteracting these biases when training the AI model. This tool could identify that a data set had a population bias or a bias towards specific sensors or specific times of data (all real examples of biases we have identified in weather and climate data) and encourage the developer to counteract these biases through over/under sampling and data augmentation. Tools and principles such as these will help to ensure our AI is more trustworthy and will also help to address environmental justice needs (McGovern et al. 2021).

It could also be argued that in the absence of empirical evidence of model interpretability by users, modelers have a duty to develop XAI that informs the user what the model is doing, sufficiently to respect their decision making autonomy, as suggested by deontological ethics. AI that convinces the user without actually informing their decision making, may blur persuasion and manipulation. This potential lack of honesty can create distrust (Lamb 2017). For example, Generative Adversarial Networks (GANs, Goodfellow et al. (2014)) and other generative AI models can create small-scale features in simulated imagery that look extremely realistic without necessarily
being accurate to the same degree. This could potentially mislead forecasters, at least with regard to the appropriate level of confidence in the output images, and thus also in their interpretation of the model outputs. In contrast, XAI that informs user decisions thereby respects decision making autonomy, contributes to ethical AI, and should contribute to trustworthy AI. Five principles for science communication emerge from deontological ethics (Keohane et al. 2014), of which honesty is primary and imperative, the others being: precision, audience relevance, process transparency and specification of uncertainty about conclusions. All of these are relevant for XAI seen as a form of science communication, and conform with findings to date regarding what might contribute to human trust in AI (e.g., Glikson and Woolley 2020).

6. Grounding AI Development through Use Cases

AI2ES is principally focused on 5 environmental science applications: convective weather, winter weather, subseasonal to seasonal prediction, tropical cyclones, and coastal oceanography. We briefly outline our work in each of these areas. Note that the applications described here are not meant to cover all important areas in weather, climate and coastal environments. Rather they should be seen as case studies that - while being important applications in their own right - serve the main purpose to ground the development of trustworthy AI methods in real-world environmental applications. Thus these topics were selected to cover many different types of problems (e.g., covering a large variety of meteorological phenomena, including a large range of temporal and spatial scales), requiring many different types of AI approaches (e.g., from generating simulated satellite imagery to using AI to learn new physics in S2S applications), and to be supported by the expertise of the founding members (PIs and co-PIs). The variety of case studies and approaches offers further opportunity to ground the research and development with the aforementioned different professional users, yielding fundamental and applied research that is actionable.

a. Convective Weather

In the area of convective weather, AI2ES is performing research on improving the skill and trustworthiness of AI predictions of weather hazards including tornadoes, hail, and severe wind. On the topic of tornado prediction (Lagerquist et al. 2019b), AI2ES is focusing on utilizing short-range (0-3 hour) NWP forecasts and dual-polarization radar observations to both produce skillful
tornado forecasts using deep learning and investigating how to best communicate these predictions to human forecasters. In the area of hail prediction (Gagne et al. 2017; Burke et al. 2020), AI2ES is investigating both short-range (0-6 hour) and multiple-day (24-48 hour) hail prediction. Hail prediction is also being used as a testbed for transition of AI2ES research results to industry applications.

The work in convective weather provides our initial test case for the synergistic research cycle involving AI, atmospheric scientists, and risk communication researchers. We are interviewing forecasters and emergency managers to understand their trust in several AI convective weather products and identifying how that trust varies as a function of XAI, visualization and interactivity with the model, and model performance (Cains et al. 2022). The results of this will inform our AI model development for all of AI2ES.

b. Winter Weather

Winter weather is a major hazard in the US. Heavy snowfall, freezing rain and extreme cold can all have severe impacts in many areas including travel (e.g. road, air, rail), utilities, commerce and public health. AI2ES will be using AI to generate solutions to improve response and resilience to winter weather, with an early emphasis on road weather and related decision-making for public safety (e.g. through prediction of precipitation type and snow amounts) and environmental conservation (e.g. through improved efficiency of salt usage). We will also be addressing the needs of the National Weather Service and potentially other winter weather sensitive sectors including energy (e.g. utilities). AI-empowered winter weather analyses and predictions will be developed to provide trustworthy, customized weather information to support decision-making ahead of, during and after storm (recovery). The work will develop trustworthy products that exploit New York State Mesonet (NYSM, Brotzge et al. 2020), and the Oklahoma State Mesonet (McPherson et al. 2007), together with outputs from traditional observations and numerical weather prediction models as well as non-traditional user-provided data sources (e.g. road temperatures, snowplow speeds, salt activation, car sensors such as windshield wipers, etc). AI is currently being used to extract weather information, such as visibility and precipitation, from the frequent camera images provided by the NYSM in a longer-term effort to extract such information from roadside cameras monitoring traffic and road conditions. The next phase of this work will exploit more data sources,
to include more emphasis on precipitation type as they affect road conditions and decision-making needs. Automakers and insurers additionally are interested in this work as it improves safety and automation.

c. Tropical Cyclones (TCs)

Proper representation of the convective structure of TCs is important for the analysis and prediction of TC intensity and TC intensity change. However, existing satellites cannot observe TC convective structure at high temporal resolution. Namely, infrared imagery from geostationary satellites provides high spatial and temporal resolution, but upper-level cirrus obscures the underlying convective structure. In contrast, microwave imagery obtained from polar orbiting satellites reveals the TC convective structure, but has very low temporal resolution. Our AI2ES team seeks to use AI algorithms to combine the best of both worlds by learning to generate simulated microwave imagery from the geostationary imagery, thus yielding imagery of TC convective structure at high temporal resolution. In the next phase of the project we will seek to study this imagery to develop a better scientific understanding of the evolution of TC structure and to develop better prediction tools for TC intensity and intensity change (Slocum and Knaff 2020; Haynes et al. 2021). We also revisit the task of predicting TC intensification directly (without microwave imagery) using physics-based AI to address the challenge that TCs are behaving differently from year to year due to rising ocean temperatures, as documented by Schaffer et al. (2020).

d. Subseasonal to Seasonal Prediction (S2S)

Making predictions in the range of two weeks to two months is known to be particularly challenging (National Academies of Sciences, Engineering, and Medicine 2016). Furthermore, it has become clear that sometimes it is possible to have good forecast skill, but not at other times (Albers and Newman 2019; Mayer and Barnes 2021). Times at which good skill is possible provide forecasts of opportunity (Mariotti et al. 2020) and one of the goals of the AI2ES team is to use machine learning to identify those conditions under which good forecast skill exists, and then use XAI techniques to understand the physical processes at play. Research by members of the AI2ES team is exploring the concept of abstention networks to identify such conditions where skillful forecasts may be possible (Barnes and Barnes 2021a,b). Abstention networks (Thulasidasan 2020)
are neural networks that are trained to make predictions, but that additionally have the option to abstain, i.e. say "I don't know" in cases where their confidence for a skilled prediction is low. In doing so, the network is able to learn the more predictable behaviour better than it would without abstention. Barnes and Barnes (2021a,b) explore how to apply abstention networks to earth science applications, including S2S.

e. Coastal Ocean Environment

The coastal environment intersects oceans, land, and atmosphere and is home to critical ecosystems, large industrial facilities, and ports. Environmental datasets coming from in situ observing networks and satellite remote sensing platforms, are still underutilized and represent a compelling opportunity to apply new AI methods to support better marine environment forecasting, science discovery, and stakeholder engagement. AI2ES is carrying out active research to combine physically based AI and machine learning with conventional numerical modeling to improve prediction skill and trustworthiness for a suite of coastal applications. These include problems ranging from marine ecology (e.g., predicting cold stuns to save sea turtles and fisheries), to marine transportation and offshore safety (e.g., marine fog forecast, ocean current and eddies prediction), to water quality (e.g., harmful algal blooms), to coastal hazards (e.g., compound flooding). And when coastal AI predictions significantly outperform existing models, such as for the prediction of coastal fog Kamangir et al. (2021), XAI has the potential to bring new insights to the dynamics of the processes including air-sea interactions.

7. Taking a Multi-sector approach

AI2ES takes a multi-sector approach (McGovern et al. 2020), one where researchers from academia, private industry, federally funded research centers, and government all work together to solve challenging problems. This approach represents the future for large-scale research initiatives as it brings together practitioners across the spectrum, from basic research all the way to operations. By working together, we can inspire new foundational research and transition research all the way to operations (R2O) as well as to other end users (R2X). For example, Google’s flood forecasting work involves the cooperation of the Indian government as well as private industry and researchers
(Matias 2021) and Schumacher et al. (2021) has demonstrated the critical need to work with directly with the targeted forecasters to develop an operational product.

As our climate changes, there are a number of wicked problems\(^1\) that must be addressed using a convergent multi-sector paradigm (Bendito and Barrios 2016). For example, the changes to high-impact weather including severe storms, heat waves, drought, and torrential rain, all require a collaborative approach to identifying the best long-term solutions that will facilitate climate resiliency and provide environmental justice. If only one agency or one sector studies the problems, they will not develop general solutions, and they may miss the inspiration of specific use-cases to drive foundational research as well as opportunities to bring the research to end-users through operational use.

8. Workforce Development and Broadening Participation

Broadening participation for both the AI and ES workforces is a major goal of AI2ES. We ensured that this goal was shared by all of our initial team members, including academic partners as well as private industry. We welcome additional partners and have already grown tremendously since our creation, including starting partnerships with NOAA Cooperative Science Centers and AMS’s outreach and education programs.

AI2ES is developing and pilot testing an occupational skills award (OSA) for community colleges, a set of five classes open to a broad range of students, including non traditional ones, taking the students from basics to implementing AI projects within a geographic information system software. This OSA award is being developed by Del Mar College with collaboration from Texas A&M University-Corpus Christi. Both are Hispanic serving and minority serving institutions and the goal of the OSA is to specifically create a pipeline of underrepresented minority students trained in AI and environmental science. The OSA debuts in Fall 2021. Once the curriculum is tested, AI2ES will share it nationwide with other community colleges.

AI2ES is also developing multiple online workforce development modules. These include virtual summer schools, full university-level classes, and short courses. Our 2021 summer school focused on trustworthy AI for the environmental sciences. Our short courses are focused tutorial sessions, facilitating a deep dive into a specific topic. Each short course also includes example Jupyter notebooks so that participants can try out the ideas on their own environmental science

\(^1\)https://en.wikipedia.org/wiki/Wicked_problem
phenomena. We recently completed a short course on XAI and have additional topics planned. Finally, university level courses are being developed and shared online so that anyone around the world can use the material for learning and retraining. All of our material is available publicly on our website, ai2es.org.

9. Sidebar

Below are our current working definitions of key terms. Note that as definitions in the literature vary and we gain additional experience, AI2ES is actively working to develop clear, shared definitions of these terms for use among our community.

- **Explainable AI** - an explainable AI method is one that can be explained post hoc, after training, in a way that makes it understandable (Schwalbe and Finzel 2021; Mueller et al. 2021). This includes methods to promote transparency into the black boxes, such as the ability to measure the importance of a variable or to see the effect of the values of that variable on the model as well as methods that allow a user to visualize patterns of activation in neural networks.

- **Interpretable AI** - an interpretable AI method is a model that is designed to be understood by humans without additional explanation. This does not include methods with large numbers of hyperparameters such as neural networks.

- **Interactivity** - the more interactive a method is, the more an end-user can change parameters, select features, change weights on data points or parameters, visualize and select a specific model or ensemble of models, and change how they view the explanation and AI output (Rudin et al. 2022).

- **Trustworthiness** - Trustworthiness and trust are related, yet distinct, concepts. Trust is *relational*, in that it is ‘given to’ or ‘placed in’ someone or something, and trustworthiness is *evaluative*, in that it is a perceived characteristic of someone or something. With this in mind, trustworthiness is a (potential) trustee’s evaluation, or perception, of whether, when, why, or to what degree someone or something should or should not be trusted. Current efforts to develop standards for trustworthiness (e.g., Commission 2019) may lead some to confuse the broader concept of perceived trustworthiness with assessment of compliance with formal standards or policies for trustworthiness. A key distinction is that trustworthiness is a subjective evaluation.
that is largely dependent on the perceptions, values, experiences, and context of the assessor, which may or may not be influenced by standards or policies for trustworthiness.

- Deontological - derived from the Greek word for duty (deon). Deontological ethics are rule-based ethics, or moral duties, such as the moral duty to be honest (Alexander and Moore 2021).

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. ICER-2019758

References


