SciML - Trustworthiness, Ethics and Bias

Mark Asch - IMU/VLP/CSU

2023

Program

1. Ethics and responsible use of scientific machine learning:
   (a) Inference Cycle
   (b) SSL and LLMs
   (c) Bias in machine learning models.
   (d) Fairness in machine learning.
   (e) Transparency in machine learning.
   (f) Trustworthiness, explainability.
   (g) The ethical challenges of using scientific machine learning.
   (h) The responsible use of scientific machine learning.

Some sayings...

ML sucks. LLMs suck. [LeCun 2023]

Weapons of Math Destruction

Algorithms are not rascist.
Your skin is just too dark.

THE INFERENCE CYCLE

SciML and Digital Twins

- We begin with a quote [K. Willcox, 2019]

  Learning from data through the lens of models is a way to bring structure to an otherwise intractable problem: it is a way to respect physical constraints, to embed domain knowledge, to bring interpretability to results, and to endow the resulting predictions with quantified uncertainties.

- One possible definition of a Digital Twin is the following [AIAA, 2020]

Definition 1 (Digital Twin)**.** A Digital Twin is a set of virtual information constructs that mimics the structure, context and behavior of an individual/unique physical asset, or a group of physical assets, is dynamically updated with data from its physical twin throughout its life cycle and informs decisions that realize value., if

SciML and Digital Twins - pragmatic

- In a more prgamatic way, If we take into account:
  - ⇒ The availability of (large) volumes of (often real-time) data.
  - ⇒ The accessibility to this data.
  - ⇒ The tools and implementations of ML/AI-based algorithms.
  - ⇒ The body of knowledge of mathematical models.
  - ⇒ The readiness and low cost of computational devices.

- Then we have all the necessary ingredients for creating a virtual counterpart of a physical object or system.

- This **"digital twin"** can then accompany its physical counterpart, possibly throughout the object's life cycle.

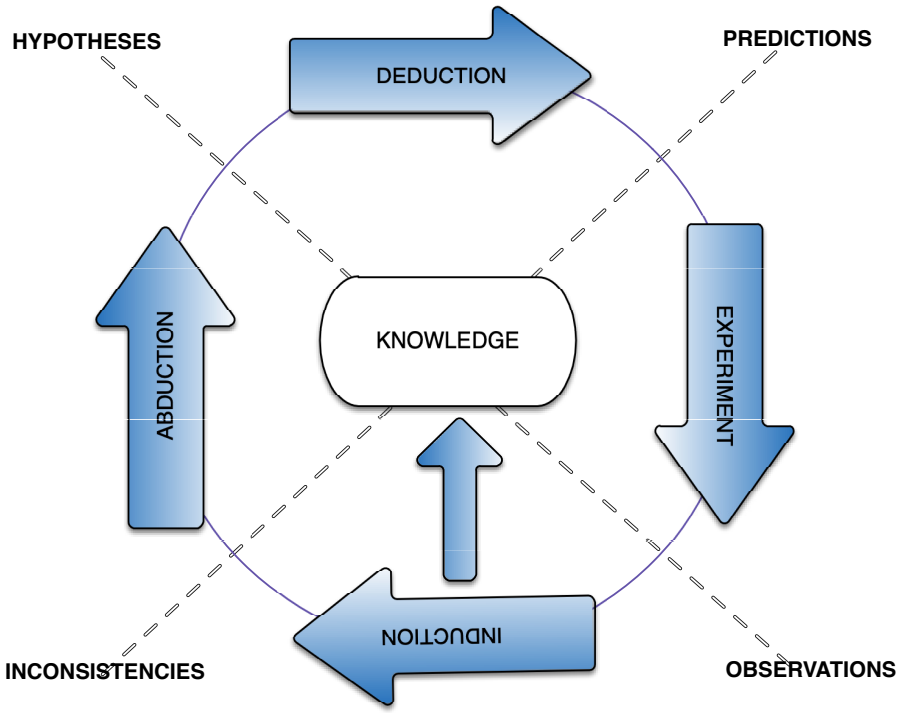
- In fact, recent developments in machine learning, could enable auto-repair and reprogramming of the twin as it ingests more and more data, and learns more and more about the physical plant.

- The twin then includes both static and dynamic parts.

  $\Rightarrow$ The static part consists of the initial model, process and design specifications.

  $\Rightarrow$ The dynamic part includes the simulation process, coupled with data acquisition, and finally auto-updating.

The Inference Cycle



- How can we conceptualize the approach to digital twins and SciML?

  $\Rightarrow$ A broader view will be indispensable because, as will be illustrated in the numerous examples, a digital twin based on SciML is a complex object that combines

  $\rightarrow$ modeling,

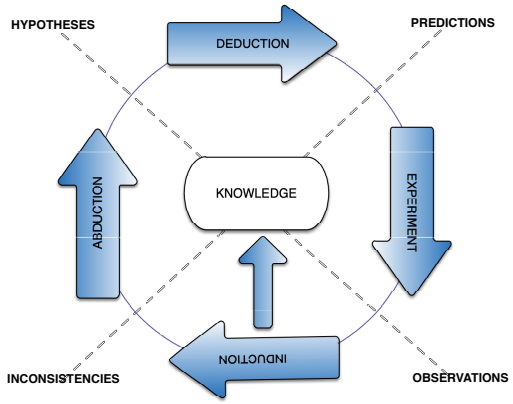
→ simulation,
→ optimization,
→ machine learning,
→ uncertainty quantification, etc.

⇒ So, it is important to see where we are, and where we are going.

- The inference cycle provides this global vision. The above Figure presents the key logical elements of the scientific process.

  ⇒ It expresses the classic view that the scientific method is a complex inferential process that seeks to improve our predictive understanding of nature by building on a foundation of thorough and carefully controlled observation.

The Inference Cycle - 3 forms of Inference



1. **Abduction**—going from (unexplained) effect to (possible) cause, i.e., guessing at an explanation.

2. **Deduction**—going from cause to effect, i.e., drawing out the necessary consequences of a set of propositions.

3. **Induction**—going from specific to general, i.e., making a sampling-based generalization. This is the process of evaluation, where we either accept the model and increase our state of belief, or we identify a mismatch, or inconsistency and we need to find a new hypothesis (by abduction) to test in the next deductive-inductive phase.

The Inference Cycle - Lessons

- We must not confuse abduction with deduction.
  - ⇒ That is, if your (deductive stage) model produces a result, you cannot assume that this is the truth.
  - ⇒ The truth has to be tested, then reprocessed through the induction stage, before updating your belief—nothing more.
- Knowledge is only a state of belief—what we *believe* to be right, to the best of our actual knowledge.
  - ⇒ That is why the cycle does not really end, we keep on going round and round, improving our state of belief at each cycle (see Figure).
- The inference cycle acts as a compass, not telling us how to get there, but rather where we are, where to go and what the next step should be.
- The cycle can be enriched with all the possible methods enabling the passage from phase to phase.

⇒ At the highest level, these are computational modeling and data analysis, or machine learning.

⇒ At the next level, we have to dig into the toolbox and extract the right methods from within these two.

⇒ Overall, we will probably/ideally need to mix and match, employing tools from both.

The Inference Cycle - Quotes

- From [Flach, Kakas, Ray 2006]

  Modeling a scientific domain is a continuous process of observing and understanding phenomena according to some currently available model, and using this understanding to improve the original domain model. In this process one starts with a relatively simple model which gets further improved and expanded as the process is iterated. At any given stage of its development, the current model is very likely to be incomplete. The task then is to use the information given to us by experimental observations to improve and possibly complete this description. The development of our theories is driven by the observations and the need for these theories to conform to the observations. This point of view forms the basis of many formal theories of scientific discovery (see Popper, Kuhn) in the sense that the development of a scientific theory is considered to be an incremental process of refinement strongly guided by the empirical observations.

- From [Peirce, Collected Papers, 1978]

> 141. All positive reasoning is of the nature of judging the proportion of something in a whole collection by the proportion found in a sample. Accordingly, there are three things to which we can never hope to attain by reasoning, namely, absolute certainty, absolute exactitude, absolute universality. We cannot be absolutely certain that our conclusions are even approximately true; for the sample may be utterly unlike the unsampled part of the collection. We cannot pretend to be even probably exact; because the sample consists of but a finite number of instances and only admits special values of the proportion sought. Finally, even if we could ascertain with absolute certainty and exactness that the ratio of sinful men to all men was as 1 to 1; still among the infinite generations of men there would be room for any finite number of sinless men without violating the proportion. The case is the same with a seven legged calf.

142. Now if exactitude, certitude, and universality are not to be attained by reasoning, there is certainly no other means by which they can be reached.

The Robot Scientist

- More formally,
  - ⇒ given a theory T describing our current (incomplete) model of the scientific domain under investigation, and
  - ⇒ a set of observations O,
  - ⇒ *abduction* and *induction* are employed in the process of incorporating the new information contained in the observations O, into the current theory T.
  - ⇒ They both synthesize new knowledge, H, that extends the current model to $T \cup H$ such that the union is consistent and obtained by using the *deductive* process.

- The concept of a "Robot Scientist" [King 2009] was used to test the automation of this abductive reasoning, where a hypothesis H is generated to explain a goal G with respect to a theory T. This is actually a form of Machine Learning, in its purest sense.

- The robot proceeds by:

  $\Rightarrow$ hypothesizing to explain observations,
  $\Rightarrow$ devising experiments to test these hypotheses,
  $\Rightarrow$ physically running the experiments using laboratory robotics,
  $\Rightarrow$ interpreting the results from the experiments,
  $\Rightarrow$ repeating the cycle as required.

- This sounds familiar—just replace the robot by a human research engineer trying to build a DT using SciML, and we are in the inference cycle of scientific research.

- Bayesian Optimization can be used to help in the choice of optimal parmeters for the nect cycle.

SSL and LLMs - the future?

Self-Supervised Learning (SSL) and Large Language Models (LLMs)

- [LeCun, 2023]

  ⇒ "ML sucks" (compared to humans and animals)
  - → a 17 year-old can learn to drive a car in 20h of practice, but we still don't have Level-5 autonomous driving

  ⇒ Self-Supervised Learning (SSL) (= filling in the blanks) has taken over the world
  - → performance is amazing ("we are fooled by their fluency"), but they (LLMs) make stupid mistakes
  - → they have no common sense and cannot plan their responses
  - → in general, they're very bad at math—but plugins, such as Mathematica, are changing this

  ⇒ "Large Language Models (LLMs) suck"
  - → good for: writing assistance (first draft, style polishing), code writing assistance
  - → bad for: producing factual and consistent answers

(hallucinations!), reasoning, planning, math, using tools such as search engines, calculators, databases

⇒ The world is only partially predictable, so how to deal with this?

⇒ Solutions:

→ JEPA (joint embedding architectures),

→ energy-based models (replace probabilistic models that are intractable)

→ spiked neural networks (closer to brain functions), etc.

→ ETHICS and BIAS analysis

LLM Impacts

- Significant impacts already in
  - ⇒ software productivity
  - ⇒ basic science and drug design: Alpha-Fold helping to design COVID vaccines
  - ⇒ transportation safety
  - ⇒ healthcare, radiology
  - ⇒ weather and climate predictions

LLMs in the (near) Future

- Small number of highly intelligent multimodal LLMs, 10x size of today's
- Open source, smaller LLMs that are downloadable and proliferate worldwide
- Many specialized LLMs for healthcare, legal work, finance, . . .
- LLMs using diverse plugins (calculators, route planners, databases, LLMs,. . .)
- Multi-modal models increasingly able to act in the physical world
- Personal LLMs for everybody, knowledgeable about each individual user.
- Wider use in education: how we educate and what we teach.

Conclusion

LLMs will require extensive norms and controls, in paricular with respect to ethics, privacy and bias.

TRUST AND TRUSTWORTHINESS

Definitions

These are current working definitions of key terms, from AI2ES[1] who are actively working to develop clear, shared definitions of these terms [5].

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.

[1] `https://www.ai2es.org/`

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) trustor'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., High-Level Expert Group on AI 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).

AI-Bias— AI bias, is a phenomenon that occurs when an algorithm produces results that are systemically prejudiced due to erroneous assumptions in the machine learning (ML) process.

Definitions of Trust

1. Trust is the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party. (e.g., Mayer et al 1995)

2. Trust: In the presence of uncertainty, the degree to which someone does or does not rely on, or put faith in, someone or something (Wirz et al.)

   (a) Definition is purposefully broad, so as to capture the many different definitions and related dimensions of trust.
   (b) Our definition of trust is designed to capture trust in all forms.

3. Trust is the relationship between a trustor and a trustee: the trustor trusts the trustee. Trust is dynamic, evolves with interactions, and is easier to lose than gain.

4. **AI2ES Definition**: Trust is the willingness to assume risk by relying on or believing in the actions of another party.

Context and Trust

Trust is invaribaly context-dependent

Actors: Who is being expected to trust?

Targets: What are they being expected to trust?

Purpose: What should they trust something/someone for?

Reason: Why should they trust someone/something?

Setting: In what place or role are they being asked to trust?

Trustworthiness

- **Aim**: how to develop trustworthy AI/SciML for environmental sciences?

  $\Rightarrow$ the foundations of trustworthiness for AI
  $\Rightarrow$ explanatory AI (XAI): how explanations, physics, and robustness can help build trust in AI
  $\Rightarrow$ the relationship between ethics and trustworthiness
  $\Rightarrow$ how machine-learning systems have been developed for a range of environmental science applications

- XAI offers the opportunity for scientists to

  $\Rightarrow$ gain insights about the decision strategy of NNs,
  $\Rightarrow$ help fine tune and optimize models,
  $\Rightarrow$ gauge trust,
  $\Rightarrow$ investigate new physical insights to establish new science.

- Trustworthy Artificial Intelligence for Environmental Science (TAI4ES) Summer School

Definitions of Trustworthiness

- Trust is relational - there is an actor (trustor) and target (trustee)

  $\Rightarrow$ Who or what am I trusting, and what am I trusting it for?

- Trustworthiness is evaluative - why should I trust you?

- **AI2ES Definition**: Trustworthiness is a trustor's evaluation, or perception, of whether, when, why, or to what degree someone or something should or should not be trusted.

ETHICS AND BIAS

Definitions and Context

- "one reason to desire trust is an 'almost necessary' condition on ethical action: that the user has a reasonable belief that the system (whether human or machine) will behave approximately as intended." (Danks, AIES'19)

- Both bias and uncertainty (including error, or noise) can cause a system to behave in unintended ways.

- More broadly, whether an action is ethical may depend on either the process or the outcomes of the action:

  ⇒ utility/benefits (consequentialism),
  ⇒ whether it is virtuous/the right thing to do (virtue ethics), or
  ⇒ whether it is required by moral principles or duties (deontological ethics)

- Honesty is a deontological imperative, to respect others' rights and dignity, and the autonomy of their will. "Be honest" is also a virtue rule.

How can AI go wrong in Environmental Sciences?

Issues related to training data:

1. Non-representative training data, including lack of geo-diversity.
2. Training labels are biased or faulty.
3. Data is affected by adversaries.

Issues related to AI models:

1. Model training choices.
2. Algorithm learns faulty strategies.
3. AI learns to fake something plausible.
4. AI model used in inappropriate situations.

5. Non-trustworthy AI model deployed.

6. Lack of robustness in the AI model.

Other issues related to workforce and society:

1. Globally applicable AI approaches may stymie burgeoning efforts in developing countries.

2. Lack of input or consent on data collection and model training.

3. Scientists might feel disenfranchised.

4. Increase of CO2 emissions due to computing.

Bias

- Different types of bias:
  - ⇒ Computational/Model Bias
  - ⇒ Data Bias
  - ⇒ Decision-Making Bias
  - ⇒ bias-variance tradeoff (recall ML lectures)
- Sources of bias
  - ⇒ from training data
  - ⇒ from flawed data sampling
- One of the most complex steps is also the most obvious—understanding and measuring "fairness."
  - ⇒ causality and counterfactual fairness
- fairness vs. bias
- mitigation strategies and methods

Data Bias

- The data itself can contain biases, which affect the AI/ML model

  ⇒ Biases could be caused by underlying human biases (e.g. unintentional or intentional).

  ⇒ Biases can be caused by sampling and selection of data.

- Potential definition:

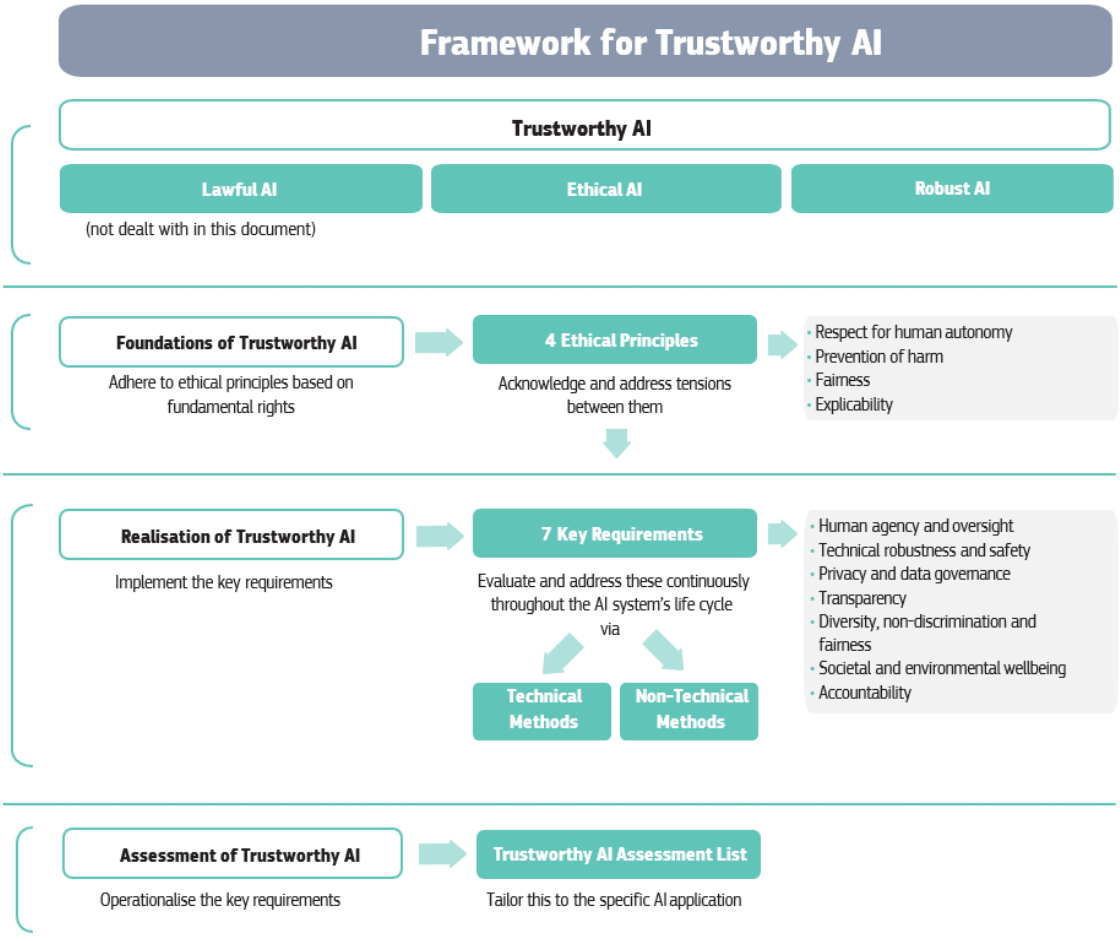
  ⇒ A class imbalance or distortion in the data from what we know is true, based on environmental and other knowledge about parameters of interest

Decision-Making Bias

- Heuristics in human decision making
  - ⇒ Perceptual biases - motion, color, orientation (Wolfe, Psych Bull & Rev 28[4], 2021) orientation
  - ⇒ Memory biases
    - → working memory (Miller's "magical number seven plus or minus two")
    - → categorization biases, determined in part by expertise
  - ⇒ Attribute substitution (Kahneman & Frederick, 2002), such as
    - → Representativeness heuristic
    - → Affect heuristic
  - ⇒ Anchoring and adjustment, for example -
    - → familiarity, salience
    - → Example: preference to use models and tools that are familiar to you
  - ⇒ Systemic biases stemming from social norms and institutions also affect decision making.

Ethics and Bias

- Ethical Principles can provide a foundation for trustworthy AI, as illustrated by the European Commission HLEG on AI 2019[2] Ethics Guidelines for Trustworthy AI



[2] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

- 4 Principles

1. Respect for human autonomy
2. Prevention of harm
3. Fairness
4. Explicability

- 7 Key Requirements

⇒ **Human agency and oversight** Including fundamental rights, human agency and human oversight
⇒ **Technical robustness and safety** Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
⇒ **Privacy and data governance** Including respect for privacy, quality and integrity of data, and access to data
⇒ **Transparency** Including traceability, explainability and communication
⇒ **Diversity, non-discrimination and fairness** Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
⇒ **Societal and environmental wellbeing** Including sustainability and environmental friendliness,

social impact, society and democrac

⇒ **Accountability** Including auditability, minimisation and reporting of negative impact, trade-offs and redress.

- Key Guidance for Realization of Trustworthy AI:
  - ⇒ Ensure that the AI system's entire life cycle meets the seven key requirements for Trustworthy AI.
  - ⇒ Consider technical and non-technical methods to ensure the implementation of those requirements.
  - ⇒ Foster research and innovation to help assessing AI systems and to further the achievement of the requirements; disseminateresults and open questions to the wider public, and systematically train a new generation of experts in AI ethics.
  - ⇒ Communicate, in a clear and proactive manner, information to stakeholders about the AI system's capabilities and limitations, enabling realistic expectation setting, and about the manner in which the requirements are implemented. Be transparent about the fact that they are dealing with an AI system.
  - ⇒ Facilitate the traceability and auditability of AI systems, particularly in critical contexts and situations.

⇒ Involve stakeholders throughout the AI system's life cycle. Foster training and education so that all stakeholders are aware of and trained in Trustworthy AI.

⇒ Be mindful that there might be fundamental tensions between different principles and requirements. Continuously identify, evaluate, document and communicate these trade-offs and their solutions.

- Key Guidance for Assessment of Trustworthy AI:

⇒ Adopt a Trustworthy AI assessment list when developing, deploying or using AI systems, and adapt it to the specific use case in which the system is being applied.

⇒ Keep in mind that such assessment list will never be exhaustive. Ensuring Trustworthy AI is not about ticking boxes, but about continuously identifying requirements, evaluating solutions and ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders therein.

Conclusion

Unethical and biased models should not be trusted!

EXPLAINABLE AND INTERPRETABLE

Explainable AI - Methods

- Theoretical topics:
  - ⇒ the difference between interpretable and explainable AI,
  - ⇒ trust vs. trustworthiness, and
  - ⇒ model-dependent vs. model-agnostic explanation.
- Practical explanation methods:
  - ⇒ the J-measure,
  - ⇒ permutation- and impurity-based variable importance,
  - ⇒ sequential forward and backward variable selection,
  - ⇒ partial-dependence plots,
  - ⇒ saliency maps,
  - ⇒ class-activation maps,
  - ⇒ integrated gradients,
  - ⇒ feature visualization,
  - ⇒ backward optimization, and
  - ⇒ novelty detection.
- Finally, augmented explanation methods, which include

$\Rightarrow$ physical constraints and
$\Rightarrow$ checks for statistical significance.

Conclusion

These require a separate course on Advanced Statistical Methods for model evaluation...

Global vs. Local Explainability

- **Global explanations** attempt to describe the model as a whole

  ⇒ What are the important features?
  ⇒ What relationship has been learned for this feature?

- **Local explanations** attempt to describe individual predictions

  ⇒ Which feature is making the biggest impact on the prediction for this example?
  ⇒ If this feature value was slightly different, how would it change the prediction?

Global Explainability

Global explainability methods can be divided into 3 categories:

1. Feature Importance/Relevance
   - Importance: How does this feature contribute to the model's performance?
   - Relevance: How does this feature contribute to the model's prediction?
2. Feature Effects
   - What is the relationship between this feature's values (or these set of features) and the model's prediction?
3. Feature Interactions
   - How is a feature's effect impacted by the effects of other features?

Local Explainability

The most common local explanation methods are known as feature attribution methods where we assume that a model's prediction$P$ can be interpreted as a linear combination of contributions from each feature

$$P = \phi_0 + \sum_{j=1}^{p} \phi_j,$$

where

- ϕ_0 is the average prediction
- the second term is the sum of the contributions from each feature

Feature Importance

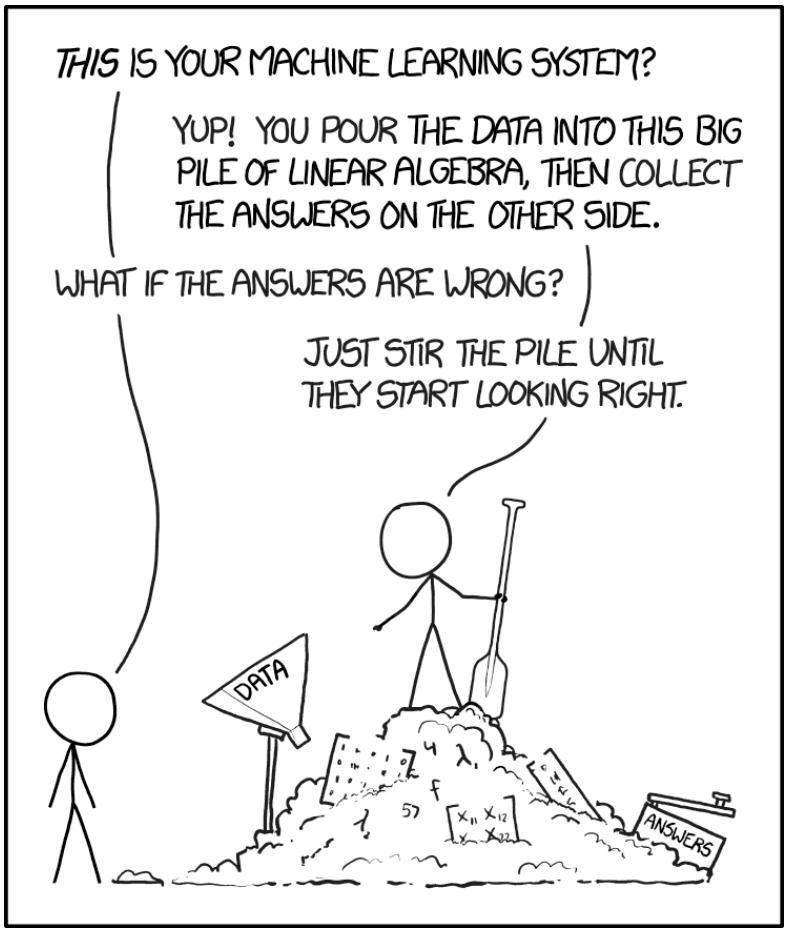
- Establishing the important features helps inform the explainability downstream.
- Given their greedy nature, ML models will tend to favor only a subset of the total features they are trained on.
- Thus, explaining an ML model largely comes down to explaining the top features.
- By knowing the top features we can ask the following questions:
  - $\Rightarrow$ How much more important are they than the less important features?
  - $\Rightarrow$ What are the learned relationships for these top features?
  - $\Rightarrow$ What features are interacting with them?

Other methods

- Permutation importance: determine the importance of a feature by removing it from the model and evaluating how the model performance suffers.
- Partial dependence: evaluate the sensitivity of a model's prediction to changes in the value for a particular feature.
- Shapley values: ϕ_i for feature x_i is the weighted average difference in model prediction when it is included and not included in some subset of features for all possible features subsets.

INTERPRETABILITY

Interpretability



This is definitely not the way to go!!!

- Meaning of interpretabilty

  $\Rightarrow$ dictionary: to explain or to present in understanble terms

$\Rightarrow$ ML: degree to which a human can understand the cause of a decision

Interpretability—importance

- With widespread use of machine learning (ML), the importance of interpretability has become clear in avoiding catastrophic consequences.

- Black box predictive models, which by definition are inscrutable, have led to serious societal problems that deeply affect health, freedom, racial bias, and safety.

- Interpretable predictive models, which are constrained so that their reasoning processes are more understandable to humans, are much easier to troubleshoot and to use in practice.

- It is universally agreed that interpretability is a key element of trust for AI models [Rudin 2022]

Interpretability—black boxes

Definition 2 (Black Box ML Model)**.** A black box machine learning model is a formula that is either too complicated for any human to understand, or proprietary, so that one cannot understand its inner workings.

- Black box models are difficult to troubleshoot, which is particularly problematic for medical or health data.
- Black box models often predict the right answer for the wrong reason (the "Clever Hans" phenomenon), leading to excellent performance in training but poor performance in practice—this is the brittleness phenomenon related to the bias-variance trade-off.
- There are numerous other issues with black box models.
  - ⇒ In criminal justice, individuals may have been subjected to years of extra prison time due to typographical errors in black box model inputs.
  - ⇒ Poorly-designed proprietary models for air quality have had serious consequences for public safety during wildfires.

⇒ Both of these situations may have been easy to avoid with interpretable models—see below for further explanations.

- In cases where the underlying distribution of data changes (called domain shift, which occurs often in practice), problems arise if users cannot troubleshoot the model in real-time, which is much harder with black box models than interpretable models.

- Standard performance metrics (such as area under the ROC curve – AUC) can be misconstrued as representing the value of a model in practice, potentially leading to overconfidence in the performance of a black box model. Specifically, a reported AUC can easily be inflated by including many "obvious" cases in the sample over which it is computed.

- Determining whether a black box model is fair with respect to gender, poverty or racial groups is much more difficult than determining whether an interpretable model has such a bias.

- In medicine, black box models turn computer-aided decisions into automated decisions, precisely because physicians cannot understand the reasoning processes of black box models.

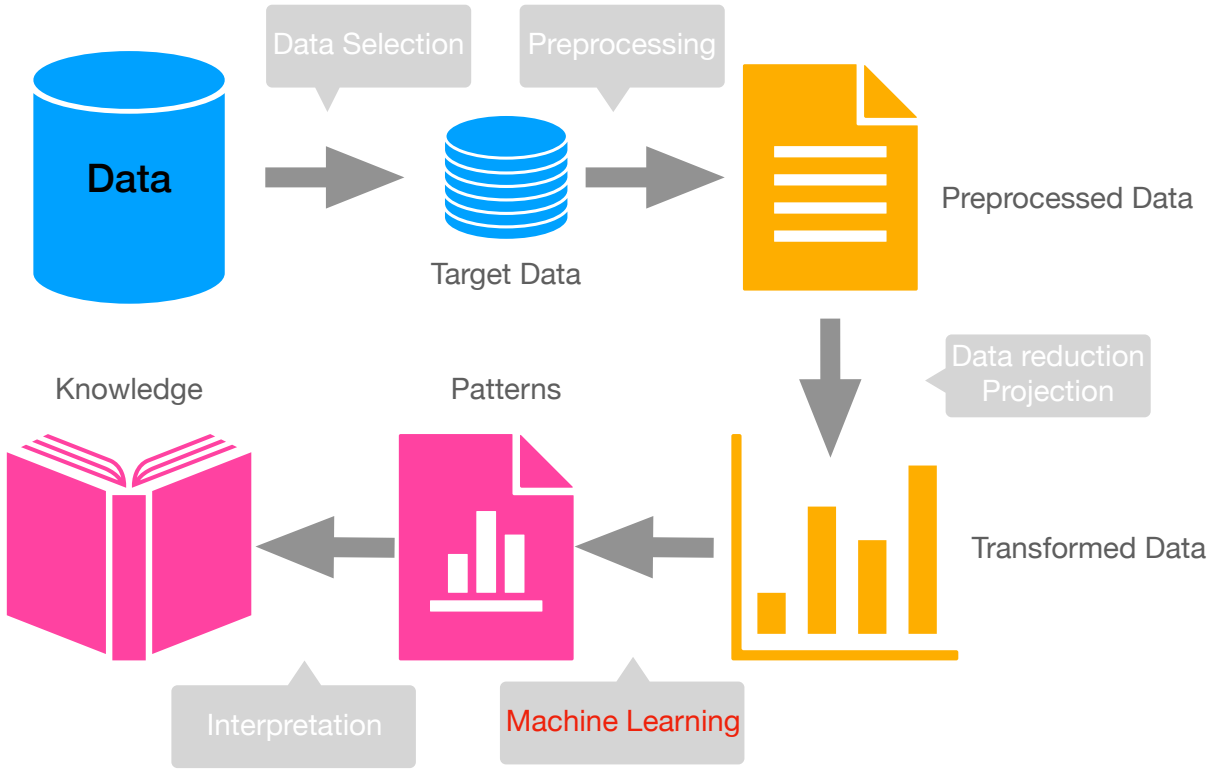
- While interpretable AI is an enhancement of human decision making, black box AI is a replacement of it.

- Explaining black boxes, rather than replacing them with interpretable models, can make the problem worse by providing misleading or false characterizations.

Conclusion

There is a clear need for innovative machine learning models that are inherently interpretable.

Knowledge Discovery Process



In a full data science process, interpretability plays a key role in determining how to update the other steps of the process for the next iteration. One interprets the results and tunes the processing of the data, the loss function, the evaluation metric, or anything else that is relevant, as shown in the diagram. How can one do this without understanding how the model works?

Interpretability—5 principles

Principle-1 An interpretable machine learning model obeys a domain-specific set of constraints to allow it (or its predictions, or the data) to be more easily understood by humans. These constraints can differ dramatically depending on the domain.

Principle-2 Despite common rhetoric, interpretable models do not necessarily create or enable trust – they could also enable distrust. They simply allow users to decide whether to trust them. In other words, they permit a decision of trust, rather than trust itself.

Principle-3 It is important not to assume that one needs to make a sacrifice in accuracy in order to gain interpretability. In fact, interpretability often begets accuracy, and not the reverse. Interpretability versus accuracy is, in general, a false dichotomy in machine learning.

Principle-4 As part of the full data science process,

one should expect both the performance metric and interpretability metric to be iteratively refined.

Principle-5 For high stakes decisions, interpretable models should be used if possible, rather than "explained" black box models.

Interpretability in the Knowledge Discovery Process

- The knowledge discovery process in Figure 1 explicitly shows important feedback loops.

- In practice, iti useful to create many interpretable models (satisfying the known constraints) and have domain experts choose between them.

  $\Rightarrow$ Their rationale for choosing one model over another helps to refine the definition of interpretability.

  $\Rightarrow$ Each problem can thus have its own unique interpretability metrics (or set of metrics).

Interpretability Setup for Supervised Learning

- Suppose we have

  $\Rightarrow$ data $\{y_i\}_i$
  $\Rightarrow$ models from a function class $\mathcal{F}$
  $\Rightarrow$ a loss function $\mathcal{L}$

- Then the interpretable, supervised learning process can be defined as

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_i \mathcal{L}(f, y_i) + \alpha P(f),$$

  where P is an interpretability penalty, subject to the interpretability constraint,

$$C(f)$$

- The loss function, as well as soft and hard interpretability constraints, are chosen to match the context.

- The goal of the constraints is to make the resulting model f or its predictions more interpretable:

  $\Rightarrow$ The constraints would generally help us find models that would be interpretable (if we design them well), and

  $\Rightarrow$ We might also be willing to consider slightly suboptimal solutions to find a more useful model.

  $\Rightarrow$ The constant α trades off between accuracy and the interpretability penalty, and can be tuned, either by cross-validation or by taking into account the user's desired tradeoff between the two terms.

- Creating interpretable models can sometimes be much more difficult than creating black box models for many different reasons including:

  $\Rightarrow$ (i) Solving the optimization problem may be computationally hard, depending on the choice of constraints and the model class $\mathcal{F}$.

  $\Rightarrow$ (ii) When one does create an interpretable model, one invariably realizes that the data are problematic and require troubleshooting, which slows down deployment (but leads to a better model).

⇒ (iii) It might not be initially clear which definition of interpretability to use. This definition might require refinement, sometimes over multiple iterations with domain experts.

Operational Interpretability

- Define your need:

  ⇒ what your definition is and what you are optimizing
  ⇒ evaluate with your end-task in mind

- The user/client does not need to undersatnd every single thing about the model—all that is needed is knowing enough for one's objectives.

  ⇒ "Enough" is for what we are trying to do... and no more.

Interpretability is NOT

- about making ALL models interpretable
- about understading EVERY SINGLE line of a model
- against developing complex models
- only about gaining user trust and fairness...

Need for Interpretability

We may not really need/want interpretabilty

- no significant consequences, or low stakes—eg. advertising
- prediction is what we care about
- abundance of empirical evidence (logical/expected reponse)
- model is 100% reliable
- prevent users from manipulating/gaming the system

Conclusion

We do not always need interpretability.

Interpretability - Performance Trade-off

- "It is a myth that there is necessarily a trade-off between accuracy and interpretability." [Rudin 2019] There is no scientific evidence for this.

- Carefully adding structure to the model—architecture, prior, loss function—has long been done to improve performance without interpretability in mind.

Conclusion

An interpretability and performance trade-off often does not exist. One should not equate interpretability with model sparsity.

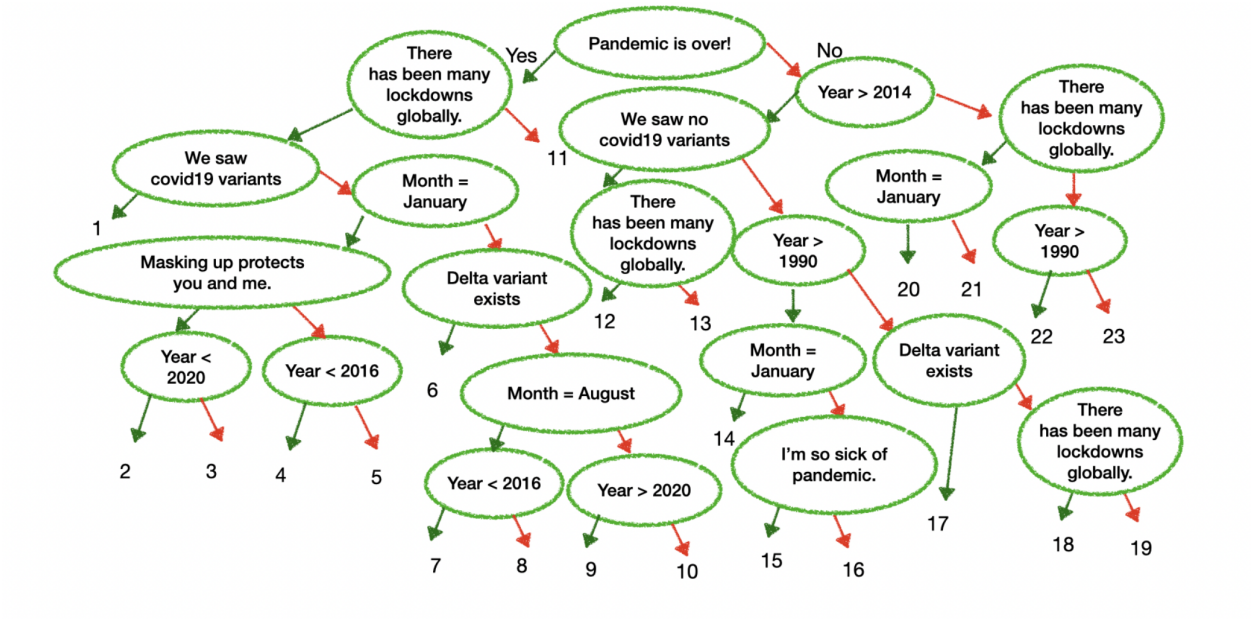
Interpretability and Trust?

- Trust, fairness and interpretability are NOT the same thing...

  $\Rightarrow$ interpretability may improve them, but difficult to formalize

  $\Rightarrow$ once formalized, you may not need interpretability...

Interpretability Methods



- Can you not just use decision trees?

  $\Rightarrow$ they can become very intricate and difficult to follow

  $\Rightarrow$ they are not robust unless we use aggregation, but then we loose the basic "interpretabilty"

Conclusion

Decision trees are not always interpretable—depends on your goal.

- Other methods:

  $\Rightarrow$ linear classifiers
  $\Rightarrow$ causality analysis

Types of Interpretability Methods

- Explaining the data—EDA.
- Building inherently interpretable models
- Post-training intepretability methods

See work of [Doshi-Velez, Been Kim]

Explainable vs Interpretable

- Explainable AI (XAI) attempts to explain a black box using an approximation model, derivatives, variable importance measures, or other statistics),
- Interpretable Machine Learning is creating a predictive model that is not a black box.
- Unfortunately, these topics are much too often lumped together within the misleading term "explainable artificial intelligence" or "XAI" despite a chasm separating these two concepts.
- Explainability and interpretability techniques are not alternative choices for many real problems; one of them (XAI) can be dangerous for high-stakes decisions to a degree that the other is not.
- History: Interpretable ML is not a subset of XAI. The term XAI dates from ∼2016, and grew out of work

on function approximation; i.e., explaining a black box model by approximating its predictions by a simpler model, or explaining a black box using local approximations. Interpretable ML also has a (separate) long and rich history, dating back to the days of expert systems in the 1950's, and the early days of decision trees. While these topics may sound similar they differ in ways that are important in practice.

Conclusions

When one explains black boxes, one expects to lose accuracy, whereas when one creates an inherently interpretable ML model, one does not. In fact, black boxes are generally unnecessary, given that their accuracy is generally not better than a well-designed interpretable model.

DATA AND WORKFLOWS

Trustworthy Data and Workflows

- **Recall**: the complete Knowledge Discovery Process can impact trustworthiness.

- Understand your data source

  $\Rightarrow$ observational
  $\Rightarrow$ simulated/synthetic
  $\Rightarrow$ crowd-sourced

- See above: how can AI go wrong?

Note

AI in environmental applications takes a diverse set of data in the development and evaluation process. Building trust should start from accounting for the quality and limitations of these data sources.

Observational Data

- Environmental AI applications often rely on in situ observation data to train or evaluate the model.

  1. How representative (spatially and/or temporally) are these data?
  2. Are there any systematic bias/error of these data?
  3. What is the uncertainty of these data?

- Satellite data are another set of observational data often used. You should consider:

  $\Rightarrow$ What is the quality of single-sensor data?
  $\Rightarrow$ How consistent are data from multiple sensors?
  $\Rightarrow$ Is the satellite observation the same as what you want?

Simulated Data

Why use simulated/synthetic data?

- Real data can be incomplete or inaccessible.
  - ⇒ Obtaining real data may be unethical—e cannot just infect people to test a new vaccination strategy...
- Real data cannot be directly used (due to restrictions such as privacy)
- Common examples:
  - ⇒ Radiative transfer model simulation to simulate satellite data for retrieval algorithm development
  - ⇒ Reanalysis data (e.g., ERA-5)
  - ⇒ Climate model simulations (e.g., CMIP6)
  - ⇒ Large eddy simulation (LES)
  - ⇒ Epidemiological projections (SIR)

Things to consider:

- There is an algorithm behind the simulated/synthetic data (including input and output).

- All of this must be trustwortty—see VV and UQ [Asch2022]

Crowd-Sourced Data

- Crowd-sourced data provides unique opportunity to fill the data gap in traditional data collection methods.
- It is very challenging to establish consistent data quality for crowd-sourced data...

Biases in Data

- See above: how can AI go wrong?
- Non-representative training data: rare events, non-uniform sensors, remote areas not covered, phenomena not well-represented at nigh, under-sampling
- Human-created biases:
  - ⇒ human labels can be wrong and the distribution discrete rather than continuous,
  - ⇒ hail size is continuous yet people cluster labels to common objects
- Temporal or seasonal biases
- Adversarial data:
  - ⇒ crowd-sourced data can be hacked
  - ⇒ false damage/storm/health reports
  - ⇒ insurance fraud

Data Quality

- Data quality factors to consider for YOUR use cases:
  - ⇒ Bias/accuracy
  - ⇒ Completeness/coverage
  - ⇒ Resolution/frequency
  - ⇒ Consistency
  - ⇒ Timeliness

- Datasheets or Data Management Plans (DMP)—"Documentation to facilitate communication between dataset creators and consumers."
  - ⇒ Motivation (why & who created it)
  - ⇒ Composition (what is in it)
  - ⇒ Collection process (how it is created)
  - ⇒ Preprocessing/cleaning/labeling (provenance)
  - ⇒ Uses (intended & not-suitable)
  - ⇒ Distribution (who can use it)
  - ⇒ Maintenance (dataset sustainability)

- Data Splitting is indispensable for model training, validation and testting

  ⇒ spatial considerations: First Law of Geography (Tobler, 1970): everything is related to everything else, but near things are more related than distant things.
  - → Random split often ignores this spatial autocorrelation aspect of the geospatial data and causes spillover effect. The result can lead to decreased model performance in unseen situations.
  - → Environmental data often have temporal autocorrelation – the data that from previous time periods are related to the current time – random splitting may not account for this information. Sometimes, a chronological splitting can be more appropriate.

  ⇒ Imbalanced samples need to be dealt with caution to avoid the artificial impact on the model performance.

  ⇒ Finally: AVOID OVERFITTING.

COMMUNICATION

UQ Methods

1. Quantile regression (also works for ML models other than NNs)
2. CRPS loss function
3. Parametric prediction
4. Deep ensembles
5. Monte Carlo dropout
6. Bayesian neural networks

- Please consult

  ⇒ [Asch2022]
  ⇒ blog:

https://www.inovex.de/de/blog/uncertainty-quantification-deep-l

Communicating Risk and Uncertainty

- AI2ES Definition: Trust is the willingness to assume risk by relying on or believing in the actions of another party.

- How big the uncertainties are, matters in decision making!

- Communicating numerical risks:

  ⇒ Use absolute risks (but also provide relative risks when dealing with potential catastrophic events).
  ⇒ For single unique events, use percent chance if possible, or if necessary, "1 in X."
  ⇒ When appropriate, express chance as a proportion, a frequency, or a percentage—it is crucial to be clear about the reference class.
  ⇒ Precision
    → highest is a full, explicit probability distribution
    → lowest is explicit denial that uncertainty exists
    → inbetween, we have: distribution summaries, range or order-of-magnitude assessment, qualifying ver-

bal statement, list of possible scenarios, informal mention of uncertainty

Recommendations

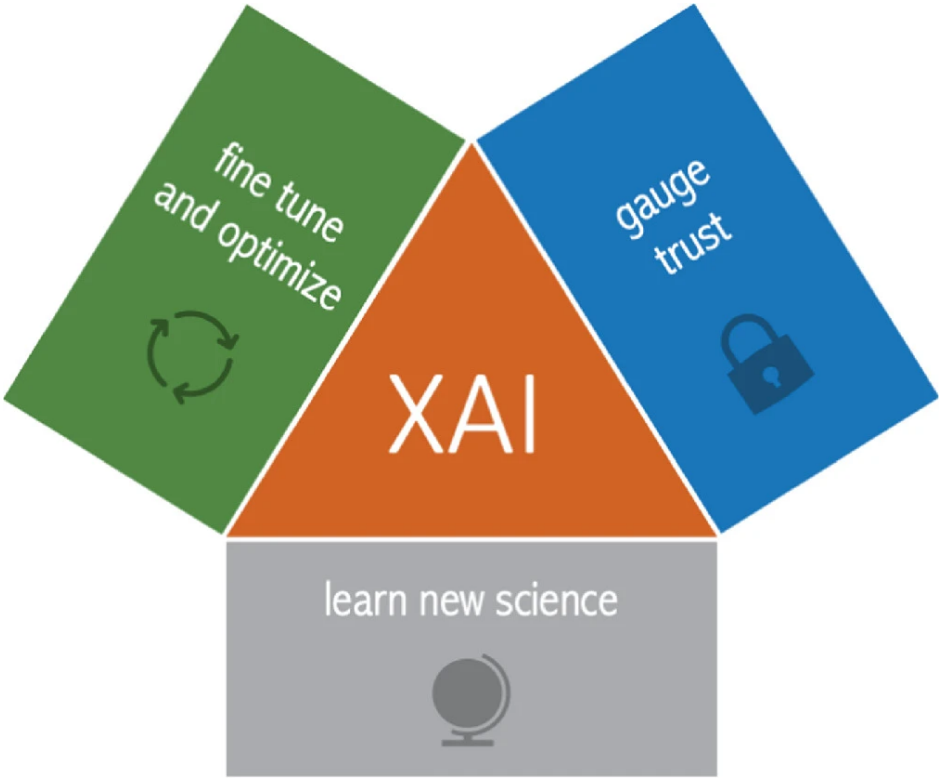
1. Communicating uncertainty can increase trust in the information, affect attitudes toward the messenger, and may sometimes delay decision making.

2. Align the format with the decision.

3. And test your communications!

Graphics


Trustworthy AI
Physics-based
Explainable
Robust
Uncertainty
Risk Communication
Transparency
Attitudes & Perceptions
Decision Making
Environmental Science
Convective Hazards
Winter Weather
Coastal Environment
Tropical cyclones
S2S
Broadening Participation
URM partnerships
Industry Partnerships
Mentoring
Workforce Development
Leadership training
Community college
Continuing education
ARTIFICIAL INTELLIGENCE
RISK COMMUNICATION
ENVIRONMENTAL SCIENCE
Ethical, Responsible, and Use-Inspired AI
ai2es.org


fine tune and optimize
gauge trust
XAI
learn new science

Bibliography

References

[1] M. Asch. *Digital Twins: from Model-Based to Data-Driven.* SIAM, 2022.

[2] A. Holzinger, et al. (editors). *xxAI - Beyond Explainable AI*. Springer. 2022

[3] McGovern, A., Gagne, D.J., Wirz, C.D., Ebert-Uphoff, I., Bostrom, A., Rao, Y., Schumacher, A., Flora, M., Chase, R., Mamalakis, A. and McGraw, M. (2023) *Trustworthy Artificial Intelligence for Environmental Sciences: An Innovative Approach for Summer School.* Bulletin of the American Meteorological Society, Apr 2023.

[4] McGovern, Amy, Imme Ebert-Uphoff, David John Gagne II and Ann Bostrom (2022) *Why we need to*

focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science, Environmental Data Science, 1, E6.

[5] McGovern, A., and Coauthors, 2022: *NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES)*. Bull. Amer. Meteor. Soc., 103, E1658–E1668

[6] C. Rudin. C. Chen. Z Chen. H Huang. L Semenova. C Zhong. "Interpretable machine learning: Fundamental principles and 10 grand challenges." Statistics Surveys, 16 1 - 85, 2022. https://doi.org/10.1214/21-SS133.