MITRE

MITRE Response to NIST Call for Comments: Draft NISTIR 8312

Author:
Michael Hadjimichael

October 2020

Document No: MP200929
McLean, VA

Table of Contents

Introduction

The MITRE Corporation appreciates the opportunity to comment on the draft National Institute of Standards and Technology Interagency or Internal Report (NISTIR) 8312, Four Principles of Explainable AI. This is a timely discussion in a rapidly developing area of artificial intelligence (AI) research and development. We provide comments and feedback in this document and invite the NISTIR authors to contact us for clarifications or further discussion.

The NISTIR 8312 draft introduces four principles for explainable AI (XAI), based on concepts from computer science, engineering, and psychology. It expands on their types and usage and compares the ideas with analogous concepts in human explanations.

This response document includes some high-level commentary on the concepts of XAI as provided by the NISTIR and discussion of the individual principles and other sections of the draft, and it notes other organizations working in the field that may be of interest or useful to the authors. Overall, the four principles proposed provide a solid and interesting foundation for discussion of the necessity, requirements, and benefits of XAI. Our comments call for more detailed definitions and an enhanced integration of contextual factors in the specification, consideration, and interpretation of principles of explainability.

Explainable AI

Principles in Context

As AI permeates most aspects of society and people become more aware of the impacts of AI on their lives, society has rightfully become more skeptical of the answers that AI provides. To develop general principles of XAI that apply across all sectors of society is a worthy goal, but the principles become useful only when they recognize and are adapted to the various motivations and requirements of each sector. User communities and sectors, such as healthcare, military, transportation, and government, have different needs and uses for XAI, and these needs and uses will drive different interpretations of the principles of XAI.[1]

Consider, for example, the explanatory needs in a military combat environment. A commander receives an AI-generated mission plan, along with an explanation justifying it. There is time between plan generation and execution during which the explanation can be fully understood, and the result evaluated. In contrast, an AI system that is controlling a drone fighter jet in combat may have no person to whom to provide an explanation, and a person monitoring the system's behavior from afar would likely not have adequate time for reaction in any case.

This NISTIR will be more useful to the various user communities if, along with general principles, it acknowledges the context in which the principles will be applied and provides guidelines for their interpretation and implementation according to the sector in which they are applied.

[1] A. H. Michel. 2020. "The Black Box, Unlocked: Predictability and Understandability in Military AI." Geneva, Switzerland: United Nations Institute for Disarmament Research. DOI: 10.37559/SecTec/20/AI1.

Need for Explainability

Beginning at the foundation, consider the need for or potential benefits of XAI. This NISTIR hypothesizes that trust is one benefit of XAI, as a reaction to issues of fairness and bias observed in many AI uses.

It is assumed that XAI is necessary for trust, but the reality may be more complex, as introduced above. The NISTIR (line 125) assumes that failure to articulate the rationale for an answer can affect the level of trust granted to a system, and XAI is one of several properties that characterize trust in AI systems (line 133). There is additional subtlety to be explored here to determine if XAI is necessary or sufficient for a trustworthy AI system, and under what conditions or user communities. The answers will come from the field of in human-machine teaming–the interaction of user and use (the context). When the need for XAI is justified under certain conditions, then that need, and those conditions will also dictate the form and properties of explanation required to make it effective for its purpose. Alternatively, it might be that principles can be identified that are context free.

There is room for discussion on the need for explainability, founded on the questions of how systems and humans interact.

1. What empirical evidence shows that explainability leads to trust? Many AI-based systems (e.g., map route planners) provide no explanation but are trusted by millions of users daily. How does criticality of the use case (context) modify the degree of explainability required?
2. Explainable AI is valued to the extent that it contributes to trust or to sensible decisions about the trustworthiness of a system. Trustworthiness and explainability balance against value of decision being made (risk versus benefit, tolerance). An interesting discussion motivating XAI is in Adadi et al. (2018).[2]
3. Some systems may not require explainability because the cost of error is low, and the model is validated empirically.
4. Auditability might be a more appropriate property so that an outside subject matter expert can evaluate the system without going through the code.
5. Which comes first: trust or explainability? Where does predictability fit into the picture?
6. What is special about these principles with respect to AI?
7. What guarantees, if any, are principles intended to provide?

These questions and likely others are important to understand the utility of AI explainability. Gaining this understanding will require a multifaceted group of experts consisting possibly of AI and information technology experts, ethicists, decision-makers, and those schooled in the disciplines of epistemology and semantics. Such an understanding might be gained for a subsequent iteration of the NISTIR through a consensus study conducted for NIST by the National Academies of Sciences, Engineering, and Medicine or other analogous organizations.

Additionally, the NISTIR would benefit from a formal definition of explainable AI. Is it simply AI that satisfies the four principles? In line 133, XAI is described as a property, which we infer is required to satisfy the four principles. A useful definition of the interpretability (equated with

[2] A. Adadi, M. Berrad. "Peeking Inside the Black-Box: Survey on XAI." IEEE Access, v6, DOI: 10.1109/ACCESS.2018.2870052

explainability) of a model is provided in the NISTIR reference Miller 2019: "Interpretability is the degree to which a human can understand the cause of a decision."

We also recommend that the report draw any necessary distinctions between explainability and interpretability. NISTIR reference Rudin 2019 tries to discuss the distinction. Reference Michel 2020 draws an additional, useful distinction between explainability and understandability.[3]

Conclusion

In conclusion, the field of XAI is developing rapidly, and it is a good time to propose a set of guiding principles. The NISTIR authors have captured much of the recent work in the Reference section and have drawn on it to develop and present a set of reasonable principles to guide future research in the field. The discussion below will address those principles in more detail.

Finally, as noted above, the principles of explainability may look or function differently in various sectors of society. Many organizations are investigating explainability in the context of trustworthy, fair, and ethical AI. These include the Department of Defense's Joint AI Center, the Defense Advanced Research Projects Agency (DARPA),[4] the intelligence community, the Defense Innovation Board, the National Security Commission on AI, the Advanced Technology Academic Research Center, and of course academia and corporate interests. Others have moved explainability or transparency to the forefront of their AI concerns and include organizations such as the Food and Drug Administration,[5] U.S. Patent and Trademark Office, and the Department of Justice. MITRE is engaged with many of these institutions to varying degrees and is available to help develop the principles of XAI to the benefit of AI and to fulfill MITRE's mission to work in the public interest to solve problems for a safer world.

The remainder of this document reviews and comments on each section of the NISTIR.

Section 2

Section 2 discusses the four principles. There is discussion of influences on the principles and the definition of system output. It would be helpful to understand better the broad set of motivations, reasons, and perspectives. This would lead to a better evaluation on the correctness and completeness of the principles.

What are the connections among the type of task, output, and AI and the requirements of the principles? The output described in this section (lines 159–164) references classifiers and recommenders, but what about AI methods such as reinforcement learning and autonomous systems planning? Is their behavior so complex that no meaningful explanation exists?

Principle 1

This principle requires the AI system to provide evidence, support, or reasoning for each output (lines 173–174). Can "explanation" be more crisply defined? Are these terms synonymous, and how are they defined? We look for more information in Section 3, titled Types of Explanations,

[3] A. H. Michel. 2020. "The Black Box, Unlocked: Predictability and Understandability in Military AI." Geneva, Switzerland: United Nations Institute for Disarmament Research. DOI: 10.37559/SecTec/20/AI1

[4] D. Gunning, D. W. Aha. 2019. "DARPA's Explainable Artificial Intelligence Program." AI Magazine, v40 n2.

[5] U.S. Food and Drug Administration. 2019. "Clinical Decision Support Software, Draft Guidance for Industry and Food and Drug Administration Staff," Docket Number FDA-2017-D-6569.

but it refers primarily to the users and uses of explanations. There is room for a deeper discussion on the meaning of "explanation." What properties that describe an explanation, independent of meaningfulness and accuracy? Additional discussion describing what is meant by "explanation" would help set the stage for the material that follows.

Principle 2

This principle requires that an explanation be meaningful for individual users. It is not clearly defined. As the report notes, it is not necessary that one size fits all. Background knowledge, context, and user goals help define what is meaningful to the user. However, we would prefer to see a justification and clearer definition of "meaningful." A more operationally oriented definition might be more useful: *An explanation is meaningful if and only if, for the intended user with sufficient contextual background, it allows the user to decide if the AI system should be trusted.*

Additionally, while a meaningful explanation is intuitively important, the requirement that explanations are understandable to individual users is not well-defined; what is meant by "individual users"? Perhaps this would be clarified by specifying that it be understandable to **intended** users. Section 3 describes five categories of explanation, defined by use/user case. Is it necessary to be understandable to all individual users or perhaps to just the engineers whose job it is to debug the system? Is more information required than that which is necessary to determine the validity of the AI output?

Principle 3

This principle requires that an explanation be accurate–that is, correctly explain the system's process for generating its output. There is a trade-off between explanation accuracy and meaningfulness (simplicity). Increasing the accuracy and perhaps complexity of explanation may lead to an explanation that is beyond the comprehension of most users. Conversely, an explanation comprehensible to all may be of little benefit to anyone. The report would benefit from a discussion of this trade-off and its impacts on the principle of explanation accuracy.

Principle 4

This principle requires that a system identify those cases where it was not designed or approved to operate. Unlike meaningfulness and accuracy, this is not truly a property of an explanation but rather a requirement that the system must govern its own usage.

This principle is problematic for two reasons: It is not clear what benefits this principle provides nor if the principle is feasible.

Users will certainly trust an AI system better if they know that they will be warned when it is being used in a way for which it was not designed. Model Cards are motivated in part by this desire: Tell how the model was trained and what its intended use is.

However, does this fourth principle add anything practical to the concept of explainability? Consider the environment in which the system is trained. In a closed world where a system is trained to identify different species of birds, the model has no concept of "not a bird"–it cannot help but find a bird. In an open-world example, a low probability score for every bird class may be interpreted as a situation where the classification answer will not be reliable.

In the first case, there is no concept of knowledge limits, rendering the principle impossible to implement. In the second case, the unreliability of the answer is communicated by the probability scores, and the principle has not added any benefit.

Perhaps one could revise the statement above to, “The system should be operated only under conditions for which it was designed.” This is a reasonable and universal principle and not specialized for XAI.

Additionally, it is arguable that many systems successfully operate in ways not intended by the designer. Humans relate to tools in complicated ways. Users want a tool to operate within its knowledge or capability limits, which may not correlate with what the system was designed to do.

Section 3

In this section, labeled Types of Explanations, explanations are categorized by the intended use or the consumer and beneficiary of the explanation. The section might be more accurately labeled as such.

User Benefit is about providing explanations on a case-by-case basis. Societal Benefit is about explanations in the aggregate. Regulatory and compliance and System development are about providing auditability to prove that a system is correct and to debug the system. Auditability is often closely associated with interpretability.

The distinction between User Benefit and Owner Benefit is unclear. Although they are different beneficiaries of an effective explanation, the output explanation is provided to and understood by the user of the system. The owner’s benefit is only in selling the service (in the movie example provided).

Section 4

This section provides an overview of some principles found in the literature.

Line 327 seems to be missing a word: Add “not” to “… argue that explanations do **not** need to meet …”

Section 5

This section provides an overview of categories of algorithms that output some form of explanation. It would be interesting to see if there is a useful mapping between XAI algorithm types and the types of explanations examined in Section 3.

Lines 393, 412: Given the variety of uses in the literature of the term “interpretable,” it may be beneficial to devote greater space to the definitions and distinctions between “explainable” and “interpretable.”

Line 538: “Fairwash” is a recently coined word and would benefit from a definition or reference.

Section 6

This section explores the explainability of human decisions. It is a long section which, although interesting, does not seem to contribute much to the discussion. It would be useful to see a greater motivation for covering this work, including:

- What parallels exist between human and machine decision-making, and between human and machine explanations?
- How useful are the parallels between them?
- Can any conclusions be drawn from this comparison, and has anything been learned that contributes to the principles for XAI?

Line 640: Add text: "This principle states that the system operates only under the conditions **for which** it was designed …"

Section 7

Line 680 seems to paraphrase principle 4 in a different way from originally defined. As expressed, principle 4 requires an XAI system to recognize and enforce its knowledge limits. This is a stronger requirement than that expressed in line 680, where it is required only to *express* its knowledge limits.

Summary of Specific Suggested Actions

Location	**Suggestion**
entire document	Discuss the context in which the principles will be applied and provide guidelines for their interpretation and implementation according to the sector in which they are applied.
Introduction	Is XAI necessary or sufficient for trustworthy AI?
Introduction/line 133	Add formal definition of XAI.
Introduction	Differentiate, explainability from understandability.
Section 2 (lines 159–164)	Add discussion of AI beyond classifiers and recommenders.
Section 2 (lines 173–174)	Define or differentiate evidence, support, and reasoning.
principle 2 (line 166)	Clarify individual intended users.
principle 3	Discuss trade-off between maximizing usefulness to the audience and being accurate.
principle 4	Consider if truly necessary.

line 271	Consider if owner benefit is a necessary distinction.
line 327	Add missing word “not.”
lines 393, 412	Compare and contrast interpretable and explainable.
line 538	Define “fairwash.”
Section 6	Eliminate to answer questions discussed earlier in this document to justify its presence.
line 640	Add words “for which.”
line 680	Reconcile paraphrased principle 4 with original definition.

MITRE Acknowledgments

The author, Dr. Mike Hadjimichael, wishes to acknowledge the assistance of Dr. Alexander Kott of the US Army Research Laboratory, and The MITRE Corporation staff: Lisa Bembenick, Dr. Eric Bloedorn, Dr. Michael Cohen, Dr. Eric Hughes, Dr. Peter Sharfman, Anne Townsend, and Samuel Visner. This NISTIR 8312 Response document benefited greatly from their insight.