accenture

Dr. Walter Copan
Director and Undersecretary of Commerce for Standards and Technology
National Institute of Standards and Technology
100 Bureau Drive, Stop 200
Gaithersburg, MD 20899

RE: RFI: Developing Four Principles of Explainable Artificial Intelligence

As a leading global professional services company, Accenture provides a broad range of services and solutions in strategy, consulting, digital, technology, and operations that span multiple industries. We combine artificial intelligence (AI) with deep industry and analytics expertise to help our clients embrace these emerging, intelligent technologies confidently and responsibly.

Accenture is grateful for the opportunity to provide input on the National Institute of Standards and Technology's (NIST) Four Principles of Explainable Artificial Intelligence (XAI). The draft document constitutes a thoughtful and accurate survey of the current XAI landscape. The draft, and NIST's broader program to develop a much-needed comprehensive AI framework, should significantly contribute to the private and public sector's understanding of the many considerations necessary to implement AI, while also ultimately enabling broader, faster, and more responsible use of AI.

Accenture believes that, to be most effective, humans and machines should collaborate, combining their respective strengths to provide sustainable value for consumers, businesses, governments, and society. The attached memorandum provides additional comments to bolster the NIST's consideration of this process moving ahead.

Thank you for the opportunity to provide input. We stand ready to assist NIST and the Administration in ongoing efforts to ensure U.S. leadership in AI, which includes advancing trust and explainability so the American people may fully realize the positive benefits of these technologies.

Sincerely,

Jinsook Han
Chief Strategy Officer, Accenture

Fernando Lucini
Chief Data Scientist, Accenture

Memorandum Regarding NIST's Draft Principles of Explainable Artificial Intelligence

Comments

1. Introduction

NIST's introduction places the onus to achieve societal acceptance and trust on the developers of AI systems. Developers are not the only – or even primary – individuals that bear that responsibility. Accenture believes the introduction should be a call to action for business and government leaders to recognize the importance of gaining societal acceptance and trust, and for product managers, salespeople, and executives, in addition to developers, work individually and collectively on achieving these goals. In essence, societal acceptance and trust in AI systems should be part of strategic governance and strategy discussions.

Accenture recommends that executives evaluate the scope of AI decision-making within their organizations and prioritize post-implementation explainability solutions, such as counterfactual explanations, for those systems. In future AI projects, we recommend that enterprises plan for explainability by design and develop policies and principles that will guide the development, building, and implementation of new AI systems. Currently, only 23 percent of organizations report that they are preparing their workforce for collaborative, interactive, and explainable AI-based systems.[1]

On line 120, the phrase "high-stakes" needs to be defined. Generally, an AI decision is "high-stakes" when (1) we cannot entirely test for safety, (2) the normative concepts of justice that an organization's definition of fairness is trying to achieve are not disclosed or the definition of fairness is too abstract to be encoded in an explainable wrapper, and (3) an automated decision could affect a person's life, health, wellbeing, or on the trajectory of a person's life.

2. Four Principles of Explainable AI

Accenture notes that NIST's XAI principles are consistent with six measures we proposed in 2018 that can be applied to assess the value and effectiveness of XAI[2]. Those measures are:

1. **Comprehensibility:** How much effort is needed for a human to interpret it?
2. **Succinctness:** How concise is it?
3. **Actionability:** How actionable is the explanation? What can we do with it?
4. **Reusability:** Could it be interpreted/reused by another AI system?
5. **Accuracy:** How accurate is the explanation?
6. **Completeness:** Does the "explanation" explain the decision completely, or only partially?

While XAI will use techniques that address these questions, humans should still expect a trade-off between the various principles. It may not be possible in certain cases to obtain explanations

[1] Accenture Labs, "Understanding Machines: Explainable AI". 2018.
[2] Ibid.

that fulfill all of the criteria listed above. Instead, explanations may take a hybrid form, combining some of the measures listed above but not all. For example, an explanation might be complete, actionable and reusable but not succinct. It depends on the availability of data and context, together with the algorithms that are employed and modified. We encourage NIST to include all of the above techniques along with an explanation of their limitations and highlight the potential for trade-offs between them.

Additionally, this section should include the question of whether an AI system even needs to exist. Some questions that should be asked: is an AI system even needed? Who or what does the AI benefit? What are the potential benefits from it?

2.1 Explanation

On line 173, the word "obligates" may be too strong. AI systems cannot always deliver explanations, and some explanations are useful only under certain circumstances. In relatively trivial AI decisions, such as a song recommendation on a streaming app, an explanation may be unnecessary. In this example, a general explanation for how the algorithm works could be beneficial for the broader public but explanations about specific song recommendations to an individual should not be obligated because the recommendation has an insignificant impact on individuals and does not affect them financially, politically, or in terms of their personal well-being. In some cases, business leaders will have some incentive to use stronger explanations as a differentiator against their competitors. Conversely, in areas with significant impacts on human life, like medical diagnoses or mortgage applications, it becomes increasingly vital that entities can explain why AI systems reached a specific decision. NIST should add more context around when AI systems may or may not need to be explainable, and when explanations are critical.

On lines 179 – 180, Explanation Accuracy does not impose any metric of quality. This conflicts with lines 216 – 217 of Section 2.3 Explanation Accuracy.

NIST should also include data explainability in this section. The data used to train AI systems and the variation of the labels between annotators is rarely measured. Given that this is a source of bias in most systems, measurement is critical. [3,4]

2.2 Meaningful

Meaningfulness, as defined in the memorandum, is difficult to measure. Accenture recommends including the discussion referenced above about comprehensibility and actionability, to help further enable entities to characterize and measure meaningfulness.

[3] Nassar, J., Pavon-Harr, V., Bosch, M., McCulloh, I. (2019). Assessing Data Quality of Annotations with Krippendorff's Alpha for Applications in Computer Vision. In *Proc. AAAI 2019 Fall Symposium.* Arlington, VA: AAAI

[4] McCulloh, I., Burck, J., Behling, J., Burks, M., Parker, J. (2018) Leadership of Data Annotation Teams. In *Proceedings Social Sens 2018*. Orlando, FL: IEEE.

2.3 Explanation Accuracy

Accenture believes the “Explanation Accuracy” principle should actually be called “Explanation Quality.” Most current XAI research is focused on the quality of an explanation, not accuracy.

We also recommend that NIST include a mention of performance metrics.

This section opens a broader discussion on benchmarks for XAI pipelines, evaluation protocols, and metrics. Some protocols are purely synthetic and do not involve humans, while others involve expert or non-expert human annotators.

2.4 Knowledge Limits

In practice, “knowledge limits” is known in machine learning literature as “quantifying epistemic uncertainty.”

It appears that the Knowledge Limits section is saying that a system must understand if it is being weaponized. It is impractical to harden systems in this way. Safeguarding against weaponization is the job of responsible innovation or responsible product development,a human-centric process

On line 235, there is a third reason for low confidence related to the quality and size of the trained model, when it is not “big enough” to learn from the training set (i.e. the model is under parametrized or would benefit from better hyper parameter tuning).

3. Types of Explanations

This section should include a reference to counterfactual explanations, which can often be more actionable and beneficial than other explanations. Counterfactual explanations make human-machine collaboration possible even if the AI wasn’t designed to explain its decision making process. For example, a counterfactual explanation could tell a rejected loan applicant which inputs (income, assets, etc.) would have needed to change for the application to have been approved.[5] This provides the loan applicant with concrete actions they can take to alter the decision made by the AI system.

Some XAI systems aim to generate explanations that allow users to contest the decision (i.e. to have a legal recourse). In this case, the “user benefit” is the ability to use the explanation for recourse.

4. Overview of principles in the literature

NIST has produced a strong overview of current XAI literature. However, Accenture recommends including work on generating surrogate models from black boxes, specifically a

[5] Costabello Luca, McGrath, Rory. “Interpreting AI ‘Black Boxes’ with counterfactual explanations.” 31 July 2019.

paper titled “Demystifying Black-box Models with Symbolic Metamodels” by Ahmed M. Alaa and Mihaela van der Schaar[6].

We are encouraged to see Doshi-Velez and Kim referenced on line 341; their paper is among the first to propose a taxonomy of explanation tasks (with or without humans, with or without experts, etc.).

5. Overview of Explainable AI Algorithms

There is no universally agreed-upon definition of “interpretability” or “explainability” within the XAI community. The explanation must be in service to those impact by the model. It is one thing to explain why the machine made the decision; it’s quite another say how it affects the future trajectory of someone’s life. Interpretability should be in service to individual agency and should not be a regulatory checkbox.

5.1 Self-Explainable Models

On lines 409 – 410, it is unclear what “accurate” means in this context. Instead of “accuracy,” we recommend using the term “fidelity.”

On lines 418 – 419, NIST should mention that the accuracy-interpretability tradeoff is still being heavily debated in the machine learning community.

We commend the inclusion of Generalized Additive Models (GAM) on lines 455 – 456, as GAMs are a promising direction in transparent design.

5.2 Global Explainable AI Algorithms

No comments

5.3 Per-Decision Explainable AI Algorithms

The draft states, “Good counterfactuals answer the question ’what is the minimum amount an input would need to change for the system to change its decision on that input?’” This is simply the definition of a counterfactual explanation. Researchers, including Accenture Labs, are currently experimenting with different ways of going beyond the notion of distance, for example by injecting human annotators’ preferences so we get a “good counterfactual” – a counterfactual explanation which is more actionable and relevant to affected parties.

5.4 Adversarial Attacks on Explainability

On lines 524 – 527, it is unclear what “100 percent” accurate means. Accuracy is still an open question/problem in the research community. There are currently no universally agreed-upon benchmarks.

[6] Alaa, Ahmed M. “Demystifying Black-box Models with Symbolic Metamodels”. https://papers.nips.cc/paper/9308-demystifying-black-box-models-with-symbolic-metamodels

6. Humans as a Comparison Group

Accenture commends NIST for using humans as the baseline for measuring AI systems, as AI systems are far too often measured against a baseline of perfection. Instead of asking whether an AI system is perfect, we should be asking whether the system is an improvement over the human-centric baseline.

6.1 Explanation

No comments

6.2 Meaningful

Identifying meaningfulness is one of the hardest tasks for XAI researchers.

Accenture recommends asking several questions when considering the meaningfulness of XAI: How do you quantify meaningfulness? Which protocols should be used? Which questions should we ask of human annotators? How do we sanity-check those questions? How do we sample the audience? How do you handle human disagreement?

6.3 Explanation Accuracy

As stated previously, Accenture believes this principle would be more useful and accurate if it was called "Explanation Quality."

6.4 Knowledge Limits

No comments

7. Discussion and Conclusions

In Four Principles of Explainable Artificial Intelligence, NIST states, "By understanding the explainability of both the AI system and the human in the human-machine collaboration, this opens the door to pursue implementations which incorporate the strengths of each, potentially improving explainability beyond the capability of either the human or AI system in isolation." Accenture strongly supports this statement and believes that collaboration between humans and machines is key to unlocking the full potential of AI. Indeed, in *Human+Machine* Accenture's Technology Group Chief Executive Paul Daugherty wrote, "Humans and machines aren't adversaries, fighting for each other's jobs. Instead, they are symbiotic partners, each pushing the other to higher levels of performance."[7] Accenture views XAI as the next stage of human augmentation by machines, when AI will provide humans with explanations and thus empower humans to take corrective actions. XAI and more responsible AI will be the backbone of the intelligent systems of the future that enable the intelligent enterprise.

[7] Daugherty, Paul & Wilson, H. James. "Human + Machine: Reimagining Work in the Age of AI" (Boston, MA, Harvard Business Review Press, 2018) p.8.