Comments of Microsoft Corporation on the National Institute of Standards and Technology (NIST) Report on *Four Principles of Explainable Artificial Intelligence* Draft NISTIR 8312

October 15, 2020

Microsoft appreciates the opportunity to comment on the National Institute of Standards and Technology Report (NIST) Report on *Four Principles of Explainable Artificial Intelligence*. We agree with NIST that the ability to explain AI can impact the trust users have in the AI system. Making AI systems understandable is fundamental to transparency, a foundational principle that is central to the responsible use of AI.

Given the importance of transparency, NIST's focus on this issue is welcome. When AI is used to help make decisions that impact people's lives, it is critically important that people understand, in a manner that takes into account those individuals, how the decisions are made. What is needed for individuals will likely differ from what researchers need, what would be useful for policymakers, or what is needed for a specific purpose. Further research on how to make the behavior of AI systems understandable - or "intelligible" - will contribute to a more informed approach to methods for enabling that intelligibility[1]. This suggests that reframing of the approach taken in the Report.

Achieving intelligibility can be complex. It is dependent on a host of variables and there will not be a "one-size-fits-all" approach. While the approach taken in the report is thoughtful, we offer the following recommendations:

1. **Expand the focus from "explainability" to "transparency."** Explainability is one part of transparency. Centering on the elements of transparency – explainability, traceability, and communication – provides a more robust framework to create human-understandable explanations of AI systems.

2. **Reframe the approach – looking to who you want to be intelligible for and for what purpose.** The approach to intelligibility may differ to meet specific stakeholder needs.

[1] The term "intelligibility," often used interchangeably with the terms "explainability" and "interpretability," refers to the concept of making the behavior of AI systems, or components of systems, understandable to humans.

3. **Facilitate further research** on transparency – explainability, traceability, and communication – that can build on current scholarship. Likewise, exploring the predictability of models or testing metrology related to the aspects of transparency should be considered.

4. **Foster additional dialogue** among stakeholders, including convenings , like conferences, workshops, and other meetings, on implementing transparency.

NIST can and should continue to contribute to and inform this important issue.

Explainablity is Part of Transparency.

The Report is currently framed "explainability." Explainability is a core component of transparency. While transparency provides visibility to the features, components and procedures of an AI system, explainability means that people should be able to understand, monitor, and respond to the technical behavior and decisions of AI systems.

Both transparency and explainability have technical and non-technical elements to them. Both need to be able to cater to stakeholders of different technical or legal levels of expertise. The difference between the two is in their purpose. Explainability denotes the ability to justify ("explain") an AI system output. Such ability is especially important for AI systems that use opaque machine learning models, such as neural networks.

There are other components of transparency besides explainability. First, transparency relies on a foundation of traceability, with teams clearly documenting their goals, definitions, and design choices, and any assumptions they have made. Second, transparency requires communication—those who build and use AI systems should be forthcoming about when, why, and how they choose to build and deploy them, as well as their systems' limitations.

Transparency became an umbrella term for providing any information about an AI system that would be helpful to or required by its stakeholders. As such, transparency might include the information about system's decomposition, ML model(s), training data, performance, e.g. accuracy benchmarks, applicability, and management practices of an organization responsible for the AI system. Such transparency components are essential to a broad spectrum of AI systems that use ML models of different level of "opaqueness".

The goal of making AI understandable to humans is fundamental to "transparency." Transparency represents not only the idea that people should be able to understand and monitor how AI systems behave, but also that those who use AI systems are deliberate and forthcoming about when, why, and how they choose to deploy them. All three components of transparency – explainability, transparency, and communication are fundamental to the trustworthiness of AI systems.

The Threshold Question: For whom should an AI system be intelligible and for what purpose?

The Report could be improved by reframing it in terms of stakeholders and their goals. The rationale for this approach is that transparency is driven by stakeholders and the stakeholder's goals. That framing enables exploring important aspects and trade-offs, for example the types of evidence provided, how accurately the simple explanation reflects the complex model, and the like – what is meaningful to the particular stakeholder and the explainability goals for that purpose.

At a high level one could break down "types of explanations" by common stakeholders and goals. Being more crisp about what is meant by a "user" is helpful. Is it the person using AI to make a determination or is it the person impacted by the system? Ultimately, the goal is an explanation type that covers decision makers using AI to help make their decisions.

The need for intelligibility can arise in an almost limitless range of scenarios involving any number of human actors. In most cases, in fact, intelligibility will often be a tool for achieving other human objectives, such as ensuring the fairness of decisions or the reliability and safety of AI systems operating in physical environments.

Selecting an intelligibility approach therefore starts with asking questions about the people at the center of an AI system's development and use: Who are they? What do they want to know? Why do they want to know it? Do they need to understand the overall behavior of an entire AI system, or do they require more specific understanding of a particular output or prediction? Do they need to know key characteristics of the data used to train a particular machine learning model? Once those aspects are considered then specific ways to communicate with the relevant stakeholder(s) can be determined. For example, one way to convey capabilities and limitations to users is through the use of Transparency Notes.

The diversity of reasons underlying requests for intelligibility, as well as the breadth of potentially relevant information, are illustrated by a few examples:

- **Intelligibility can improve the robustness of an AI system by making it easier to identify and fix bugs.** For example, there could be a predictive feature believed important for a machine learning model, yet adding this feature does not improve performance. Understanding how the model is using this feature will help the engineer determine how to proceed.

- **Intelligibility can help users decide how much to trust an AI system**. Suppose an AI system makes a prediction. Providing an explanation for the prediction, including a description of the factors most influential to that prediction, allows for better understand why the prediction was made and even to assess how to use that information before making a decision.

- **Intelligibility can uncover potential sources of bias or unfairness**. For example, by examining the characteristics of the data used to train models, issues related to unfairness may be identified and corrective action taken.

- **Intelligibility can help demonstrate compliance with regulatory obligations**. For example, examining the training data and understanding how that data influences the system's recommendations might help detect when a system is unintentionally using certain features, such as zip code, as a proxy for race, which the law excludes from consideration in certain decision-making, like lending.

The Report notes the accuracy and the purpose of the explanation varies based on the given task of AI systems and models. While metrics can be useful as a supporting tool to identify the list of characteristics and/or elements of explanations to meet the objectives of the AI system provider and developer those metrics must be appropriately scoped. To use such a metrics for the purpose of evaluation of an explanation accuracy would be applicable only for a very narrow scoped context. Since the large diversity of the situations and the stakeholders, universal metrics for evaluation are impractical.

Taking into account stakeholders' goals as well as the context of a given system, should help enable the articulation of risk that appropriately apply to that given system. Thus, what is provided to make a system understandable can be more germane and effectiveness improved. We suggest the same conceptual support for intelligibility that it be risk-based (or analogous) in approach. This is similar to a risk-based thinking that informs NIST's approach to both cybersecurity and privacy. The same conceptual support for intelligibility may prove useful here.

Facilitate Additional Research on Intelligibility

More research is needed to understand which approaches do and do not help people achieve the end goals for which they need intelligibility.

As the Report notes, there is existing work to draw on. For its part, Microsoft researchers are focused on various aspects of intelligibility. Because intelligibility is a fundamentally human concept, it's crucial to take a human-centered approach to designing and evaluating methods for achieving intelligibility. Microsoft researchers are questioning common assumptions about what makes a model "interpretable," studying data scientists' understanding and use of existing intelligibility tools and how to make these tools more useable, and exploring the intelligibility of common metrics like accuracy.

Tools have also been created to provide ways to implement intelligibility. Microsoft researchers have released InterpretML, an open-source Python package that exposes common model intelligibility techniques to practitioners and researchers. InterpretML includes implementations of both "glassbox" models (like Explainable Boosting Machines, which

Generalized Additive Models) and techniques for generating explanations of blackbox models (like the popular LIME and SHAP, both developed by current Microsoft researchers).

Beyond model intelligibility, a thorough understanding of the characteristics and origins of the data used to train a machine learning model can be fundamental to building more responsible AI. The Datasheets for Datasets project proposes that every dataset be accompanied by a datasheet that documents relevant information about its creation, key characteristics, and limitations. Datasheets can help dataset creators uncover possible sources of bias in their data or unintentional assumptions they've made, help dataset consumers figure out whether a dataset is right for their needs, and help end users gain trust. In collaboration with the Partnership on AI, interested stakeholders are developing best practices for documenting all components of machine learning systems to build more responsible AI.

NIST should explore ways to facilitate or incentivize research on intelligibility to further hone the principles and inform efforts to put them into practice.

Furthering Public and Stakeholder Dialogue

Fostering a more informed public understanding of these issues can help policymakers and regulators develop effective policy frameworks that best address societal needs and promote innovation. Similar to past NIST convenings, future forums and facilitations, like workshops, with the stakeholder community on intelligibility will further refine and contribute to the development of a principled framework to address the transparency of an AI system.

Comment on Human Explainability

We appreciate NIST addressing whether humans can meet the same set of principles set forth for AI. Such benchmarks – comparing abilities – are informative. We agree that moving forward incorporating the strength of both the AI systems and humans may help improve explainability beyond the capability of either in isolation. More research in this area could be instructive.

Conclusion

Transparency is fundamental to the responsible use of AI and, therefore, it's trustworthiness. Through this Report NIST is contributing to the dialogue about how to put transparency into practice. With further refinement, additional research, and stakeholder deliberations, we anticipate more positive progress can be made to form guidance and achieve that goal. We look forward to working with NIST and the stakeholder community on this important topic.