600 14th St. NW, Suite 300
Washington, D.C. 20005

October 13, 2020

U.S. Department of Commerce
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

Subject: **Draft NISTIR 8312, "Four Principles of Explainable Artificial Intelligence"**

Dear Director Copan:

On behalf of International Business Machines Corporation (IBM), we welcome the opportunity to respond to the National Institute of Standards and Technology's (NIST) draft report, "Four Principles of Explainable Artificial Intelligence" (hereafter, "draft report").

In general, we believe the four principles outlined in this draft report – explanation, meaningfulness, explanation accuracy, and knowledge limits – can help promote meaningful explainability in AI systems. We also agree on the importance of using humans as a baseline comparison for informing the development of the characteristics that define explainable AI systems.

Of course, as the draft report notes, there is still a great deal of work to be done in this field, and we look forward to opportunities to continue engaging with NIST as it works through these important questions.

To that end, we believe the following suggestions and recommendations can significantly strengthen the draft report:

- Incorporate a stronger focus on explainability in decision-making by human-AI collaborations;
- Clarify the definitions of different user groups by simplifying the audience taxonomy to "providers," "owners," and "users"; and
- Discuss how different audiences and types of explanations would link to the nature of an AI application's risk profile.

IBM commends NIST for its work on this timely and important topic. We thank you in advance for considering these comments and welcome the opportunity to engage with the agency as it moves forward in this process.

Respectfully,

Christina Montgomery
Vice President and Chief Privacy Officer
IBM Corporation

Francesca Rossi
IBM Fellow and AI Ethics Global Leader
IBM Research

IBM Response to Draft NISTIR 8312, "Four Principles of Explainable Artificial Intelligence"

IBM applauds NIST's work in developing this draft report. We are broadly supportive of the four principles – explanation, meaningfulness, explanation accuracy, and knowledge limits – NIST identifies as key to promoting meaningful explainability in AI systems. In particular, we agree that "humans as a comparison group for explainable AI can inform the development of benchmark metrics for explainable AI systems; and lead to a better understanding of the dynamics of human-machine collaboration."[1] This is an appropriate framing, as much of the value to be unlocked from AI applications will come from their use in augmenting, not replacing, human judgement and decision-making.

As the draft report notes, there is still a great deal of work to be done in "understanding general principles that drive human reasoning and decision making" in order to help better inform further study of explainable AI.[2] And because "humans demonstrate only limited ability to meet the principles," it is important for NIST to similarly recognize that the ways in which AI is used by humans should not necessarily be held to a higher standard than humans working alone, when assessing the performance of a similar task. As such, it may be worth focusing on those situations in which both the use of an AI system and a human acting without AI assistance operate at parity across the four principle domains.

We also agree with the draft report's recognition of the need for context in assessing these comparisons, especially as it pertains to "provid[ing] explanations that are intelligible and understandable" in meeting the "meaningful" principle.[3] In maximizing the potential for an AI system (or a human-led activity using AI in a defined role) to provide comprehensible explanations that meet all four principles, it is imperative to appropriately assess both the context of a given decision and its intended recipient.

To help clarify how developers, providers, and owners of AI systems can better assess the trade-offs associated with these complex considerations, we suggest NIST point to particular tools that can assist individuals at various stages of AI

[1] P. Jonathon Phillips, et al., "Four Principles of Explainable Artificial Intelligence," Draft NISTIR 8312, U.S. Department of Commerce, National Institute of Standards and Technology, p. 13, available at https://doi.org/10.6028/NIST.IR.8312-draft. (hereafter "Draft NISTIR 8312")
[2] *Id.* at 16.
[3] *Id.* at 13.

lifecycle development. For example, the draft report's footnotes point to IBM Research's own "Trusting AI" resource page, which includes resources on toolkits and methods that can aid developers and researchers engaged in model creation and deployment.

One such resource in that clearinghouse is AI Explainability 360 – an open source toolkit originally developed by IBM researchers that provides developers with a Python package suite of algorithms that can aid them in making these decisions.[4] We also recently debuted a separate website that specifically guides individuals to resources for addressing issues related to AI explainability, including a decision tree for guidance on how developers can consider which algorithms may be most appropriately suited for a given application's explainability needs.[5]

Recommendations

This draft report serves as an important development on the path towards creating a framework for trustworthy AI. In order to help advance those efforts, we offer a number of recommendations for improving on its strong foundation.

Incorporate a stronger focus on explainability in decision-making by human-AI collaborations. Although the draft report discusses humans as a comparison group to AI systems at length in Section 6, there is little discussion of the specific value in human-AI collaboration in decision-making and its impact on the accuracy of explanations. In particular, Section 6.3 – "Explanation Accuracy" – discusses the pitfalls inherent to human decision-making, as contrasted to a discussion of an AI system's accuracy in Section 2.3. However, there is a considerable lack of analysis about the implications on accuracy when *both* AI and humans are involved in a decision. One example to which such an analysis could be applied is "intelligent automation," which refers to a confluence of capabilities, from robotics process automation to AI and predictive analytics, that can help optimize traditional work processes by offloading redundant tasks to technological decision-making and augmenting human capabilities.

IBM strongly believes that the purpose of AI is to assist human endeavor rather than replace it, so we recommend the draft report include a stronger focus on explainability in the context of human-AI collaborative decision-making, rather than treating humans and AI as separate categories. At a bare minimum, we would recommend the report consider adding an additional section (or sub-sections to

[4] See http://aix360.mybluemix.net/.
[5] See http://aix360.mybluemix.net/resources#guidance.

each of the headers in Section 6) that expands on how benchmarks for explainable AI systems may be impacted by human-AI collaboration in the decision-making process.[6]

Clarify the definitions of different user groups by simplifying the audience taxonomy to "providers," "owners," and "users." As the draft report focuses heavily on contextual AI decision-making for explainability purposes, it is imperative to create a clear and consistent taxonomy: Particularly regarding the use of the terms "consumers" and "users." Currently, "user" is defined in multiple ways – as a safety regulator, a developer, an auditor, and an "end-user" more broadly. We certainly agree, as the draft report notes, that explanations are not one size fits all, and that different users (and user groups) will require different explanations, based on a variety of factors. However, recognizing that "users" and "consumers" exist all along the various links of the supply chain, it is imperative to establish a clearer definition of what differentiates, e.g., an AI developer from an operator, and a regulator from an auditor.

Given this confusion, we recommend simplifying the taxonomy by reorganizing the various disparate types of individuals and entities into three categories: providers, owners, and users. In a February 2020 Policy Lab essay, IBM briefly described the contours of differentiating "providers" and "owners." Providers are those organizations that "contribute research, the creation of tooling, and APIs," while owners tend to be organizations that "train, manage, and control, operate, or own the AI models that are put to real-world commercial use."[7] Users can then be more accurately categorized as those at the tail end of the supply chain that make direct use of the AI systems. These categories will also help to clarify the roles and responsibilities of those organizations at different layers of the AI developmental lifecycle.

Discuss how different audiences and types of explanations would link to the nature of an AI application's risk profile. In general, we support the recognition that explanations will differ based on the intended audience. Developers further upstream on the AI supply chain will likely require more detailed explanations regarding a system's decision-making process, whereas end-users and consumers

---

[6] To offer a real-world example, consider the collaboration between doctors and AI diagnostic tools in metastatic breast cancer. The error rate for doctors acting alone is 3.5 percent; for AI diagnostics acting alone, the error rate is 7.5 percent. When acting in collaboration, however, the error rate plunges to 0.1 percent. See https://www.cnbc.com/2017/05/11/from-coding-to-cancer-how-ai-is-changing-medicine.html.

[7] Ryan Hagemann and Jean-Marc Leclerc, "Precision Regulation for Artificial Intelligence," IBM Policy Lab, 21 Jan. 2020, available at https://www.ibm.com/blogs/policy/ai-precision-regulation/.

will generally need (and prefer) simpler, more generalizable explanations for a decision. However, the draft report fails to make any mention of the risk profile of a given AI application and how that may impact the expectation of explainability for a given audience.

Low-risk uses of AI may require a simple explanation – or possibly none at all – detailing how the system arrived at a given decisions, whereas more impactful AI applications, such as those used in making loan determinations, are likely to require more detailed explanations. (As the IBM Policy Lab noted in the previously mentioned policy essay on AI regulation, "any AI system on the market that is making determinations or recommendations with potentially significant implications for individuals should be able to explain and contextualize how and why it arrived at a particular conclusion."[8])

To that end, we recommend that Section 3 – "Types of Explanations" – be expanded to include an assessment of how an AI application's risk profile, in addition to use-case and audience type, would potentially impact the need for, quality, and scope of explainability.

Conclusion

IBM commends NIST for its work on this timely and important topic. We thank you for considering these comments and welcome the opportunity to engage with the agency as it moves forward in this process.

[8] *Id.*