National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

October 15, 2020

BSA | The Software Alliance (BSA) appreciates the opportunity to provide comments to the National Institute of Standards and Technology (NIST) on the draft Internal Report "Four Principles of Explainable Artificial Intelligence" (Draft Report).[1] BSA is an association of the world's leading enterprise software companies that provide businesses in every sector of the economy with tools to operate more competitively and innovate more responsibly.[2] As companies at the forefront of AI innovation, BSA members recognize that trust is essential to the public's willingness to embrace these new technologies. Accordingly, BSA engages with governments around the world on the development of policies that both promote innovation and enhance confidence in the technologies that are driving economic growth.

The paper's release coincides with a time when public interest in AI systems is at an all-time high, making this discussion (among others around AI and ethics) particularly timely. The growing ubiquity of AI has energized conversations about how to instill trust in such systems. As AI is integrated into high-stakes decision-making processes that can have consequential impacts on the public, "explainability" has emerged as a foundational element for promoting trust. As leaders in the space, BSA members are pursuing a range of efforts to enable explainability, both by integrating it into the systems they develop and by creating tools and frameworks to help other organizations enhance the explainability of their own systems. For example, IBM Research created an open source toolkit with algorithms and metrics that can help developers address the unique needs of stakeholders that may be impacted by an AI system, including in in high-stakes areas like healthcare, human resources, or loan applications.[3] Microsoft's InterpretML toolkit helps developers

---

[1] https://www.nist.gov/system/files/documents/2020/08/17/NIST%20Explainable%20AI%20Draft%20NISTIR8312%20%281%29.pdf

[2] BSA's members include: Adobe, Atlassian, Autodesk, Bentley Systems, Box, Cadence, CNC/Mastercam, IBM, Informatica, Intel, Microsoft, Okta, Oracle, PTC, salesforce.com, ServiceNow, Siemens PLM Software, Sitecore, Slack, Splunk, Trimble Solutions Corporation, The MathWorks, Trend Micro, Twilio, and Workday.

[3] https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/

20 F Street, NW, Suite 800
Washington, DC 20001
P 202-872-5500
W bsa.org

better understand how an AI model functions at a "global" level (i.e., how the overall system functions) and a "local" level (i.e., how a system arrived at an individual decision).[4]

Although there is a consensus that "explainability" can play an important role in promoting trust in AI, there is as-yet no universal understanding of what it means for a system to be explainable. NIST's effort to survey the existing literature to identify the four principles that encompass the "core concepts of explainable AI" will help stakeholders coalesce around a shared foundational understanding. There is much to commend in the Draft Report. Rather than seeking to develop a static definition, the Draft Report rightfully characterizes "explainability" as a dynamic concept whose precise meaning will vary depending on context. Accordingly, rather than identifying specific metrics for assessing whether a system is "explainable," the Draft Report seeks to identify the common properties (i.e., principles) of explainability. Although these principles – Explanation, Meaningful, Explanation Accuracy, and Knowledge Limits – are characterized as "fundamental elements" of explainability, the Draft Report cautions against a one-size-all approach for evaluating whether any particular system is "explainable" because the "context of the application, community and user requirements, and the specific task will drive the importance of each principle."

In furtherance of the objectives of the Draft Report, we offer several recommendations for your consideration.

"Explanation" Principle Should Embrace Potential for "Global Explanations." In describing Principle 1, the Draft Report indicates that "the *Explanation* principle obligates AI systems to supply evidence, support, or reasoning for each output." By focusing only on the possibility of individual output-level explanations, the description seems to suggest that AI systems that are capable of "global" explanations (i.e, explanations that offer insight into the overall working of a system) may not fall within the spectrum of Principle 1. Limiting the scope of Principle 1 in this manner seems inconsistent with the overall approach of the Draft Paper and the acknowledgement in later sections that global explanations can themselves "be used to generate per-decision explanations." Accordingly, we encourage NIST to remove the "for each output" requirement from Line 174.

Expand Discussion of Trade-Offs and Value of Explanations. We concur strongly that the precise implementation of explainability must by guided by context and that the "practical needs of the system will influence how these principles are addressed (or dismissed)." We also agree that the four principles may, at times, be in tension with one another and that system designers may have to navigate trade-offs when external constraints require them to prioritize one principle over another: "For example, emergency weather alerts need to be meaningful to the public but can lack an accurate explanation of how the system arrived at its conclusions." In further developing that Draft Report, we encourage NIST to elaborate on

[4] https://www.microsoft.com/en-us/research/uploads/prod/2020/05/InterpretML-Whitepaper.pdf

the circumstances in which such trade-offs may arise and potentially explore the considerations that may warrant prioritization of specific principles and/or design choices.

As part of such a discussion, NIST should also explore the threshold question of the circumstances in which "explainability" itself is an important element of trust. NIST could, for instance, develop a framework or rubric to help developers think through how to determine whether the value of "explainability" outweighs the costs that might arise from adding in such a capability to a system. For AI systems that have no material impact on users – e.g., a system that recommends an emoji for use in a messaging app – the meager value of an explanation may be outweighed by the engineering costs it implicates.

Acknowledge International Developments. The Draft Report would benefit from a greater acknowledgement of international discussions on issues related to the principles of explainability. Specifically, we encourage NIST to add a brief analysis of the approach to explainability that the European High-Level Expert Group on AI (HLEG) has integrated into the Ethics Guidelines for Trustworthy AI[5] and the more recent Assessment List for Trustworthy AI (ALTAI).[6] These documents frame "explainability" as one key element – along with "traceability" and "open communication" – in promoting the broader principle of "transparency" in AI. Importantly, the Assessment List for Trustworthy AI acknowledges that output-level explanations may not always be possible as a technical matter, and explains that "in those circumstances, other explainability measures (e.g. traceability, auditability and transparent communication on the AI system's capabilities)" can help to achieve the goal of promoting trust.[7] Moreover, consistent with the discussion above about trade-offs and assessing the value of explainability, the ALTAI acknowledges that "the degree to which explainability is needed depends on the context and the severity of the consequences of erroneous or otherwise inaccurate output to human life."[8] We encourage NIST to include these unique contributions of the HLEG's discussion of "explainability" into Section 5 ("Overview of principles in the literature.") Similar to the HLEG's work in this area, the Artificial Intelligence Ethics Framework for the U.S. Intelligence Community also situates explainability as a key enabler of the broader principle of transparency.[9]

Greater Exploration of Nexus Between Explainability and Human-Computer Interaction. The Draft Report includes a valuable comparison of the explainability principles to the human decision-making process. The acknowledgement in Section 6 that "human-produced

---

[5] https://ec.europa.eu/futurium/en/ai-alliance-consultation

[6] https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[7] Assessment List for Trustworthy AI at pages 14-15.

[8] Id. at 15.

[9] https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community#Transparency

explanations for their own judgments, decisions, and conclusions are largely unreliable" provides an important frame of reference for setting expectations when it comes to AI system explanations. Ensuring that policies are informed by a realistic understanding and apples-to-apples comparison of computer- vs. human-administered decisions is critical. Establishing requirements for AI systems that could not be met by human decision-makers will do little to advance the public policy objective of promoting trust. We encourage NIST to expound on this discussion and explore in greater length the impacts that human-computer interaction can have on explainability.