CENTER FOR DEMOCRACY & TECHNOLOGY

*October 15, 2020*

National Institute for Standards and Technology
100 Bureau Drive (Mail Stop 8940)
Gaithersburg, Maryland 20899-2000

**Re: Four Principles of Explainable Artificial Intelligence**

**To: P. Jonathon Phillips, et al., at the Information Technology Laboratory**

The Center for Democracy & Technology (CDT) thanks the National Institute for Standards and Technology (NIST) for establishing the groundwork for this important aspect of developing trust in automated systems, and appreciates the opportunity to provide comments on NIST's Four Principles of Explainable Artificial Intelligence.[1] We offer the following comments in support of NIST's principles, along with our hope that NIST will continue to lead this and other conversations to establish standards, best practices, and common understanding around the development, use, and assessment of automated systems. As always, CDT looks forward to future engagements with NIST and offers the expertise of its staff and that of the GRAIL Network wherever it can be helpful.[2]

CDT generally agrees with the four principles set forth by NIST, which establish that the fundamental purposes of explanations require that they provide accurate information about the reasoning leading to an output and that the information disclosed should give the audience the right level of knowledge and understanding for their circumstances, including understanding whether certain inputs or outputs fall within the system's range of competence. These principles should provide a solid foundation on which to build a larger, more detailed discussion of NIST's approach to building trust in automated systems. However, we suggest the following ideas to improve the utility of NIST's principles:

[1] NIST, Call for Comments on Four Principles of Explainable Artificial Intelligence, (August 17, 2020), https://www.nist.gov/topics/artificial-intelligence/ai-foundational-research-explainability.
[2] The Governance and Research in Artificial Intelligence Leadership Network is a project of CDT and the R Street Institute. GRAIL consists of leading experts from the academic and research communities studying AI, from the computer and data science foundations to the public policy implications, and is intended to facilitate more informed policy discussions by helping researchers engage with policy makers. Learn more at https://grailnetwork.org.

**NIST should be more explicit in articulating the idea that explanations about AI systems will necessarily address all four principles.** For example, an explanation that provides accurate, meaningful information to the relevant audience is incomplete unless it also indicates whether the particular inputs and outputs fall within the system's knowledge limits.

**We urge NIST to expand and refine the "knowledge limits" principle, and perhaps the "meaningful" principle, to include the idea that explanations should help people understand whether a particular decision tool is fit for the purpose for which they intend to use it.** We agree with NIST that systems should be able to describe whether a given input is appropriate, in terms of data type, subject, or format. However, a narrowly focused principle that addresses only whether a system is competent to process a particular input fails to consider whether the decision tool is appropriate for the purpose and context of its use. A complete explanation should include information to help people assess whether the factors that a decision tool weighs are actually relevant to the purpose the tool purports to achieve. For example, algorithmic systems used in hiring processes may make assessments based on factors correlated with traits identified in a training data set, yet employers and job seekers may lack sufficient information to judge whether the system's predictions actually assess a candidate's ability to perform the essential functions of the job— a legal requirement for employee selection procedures.[3] In such cases, a system may produce undesirable results, or be inappropriately used, even though the inputs were within the system's knowledge limits. Therefore, a complete explanation should include a consideration of a system's fitness for the given purpose.

**We urge NIST to review and acknowledge the emerging legal standards for due process in the context of the principles for AI explainability.** As NIST points out, some decisions and the automated

[3] Civil Rights Principles for Hiring Assessment Technologies- July 2020, The Leadership Conference Education Fund, http://civilrightsdocs.info/pdf/policy/letters/2020/Hiring_Principles_FINAL_7.29.20.pdf; U.S. Equal Employment Opportunity Commission, *Employment Tests and Selection Procedures*, https://www.eeoc.gov/laws/guidance/employment-tests-and-selection-procedures.

systems supporting them are already subject to legal obligations to provide some degree of explanation for each determination.[4] In CDT's view, automated decision systems used by the government should be held to the highest standards in terms of their explainability, regardless of legal obligations. More specifically, any government decision affecting people's benefits or other government-granted property interest is subject to the constitutional right to due process. This requires meaningful notice, explanation and the use of ascertainable standards. There is a growing body of case law about what that means, so it is critical that those involved in the technical community are aware of the developing legal standards for explainability.[5]

**NIST, in conjunction with agencies that use AI systems, should apply the NIST principles of explainability to decisions made by those systems.** This collaborative effort would help build trust in the government's automated decision systems and help NIST to further develop and refine these principles. In keeping with the "whole of government" approach adopted by the Executive Order on Maintaining American Leadership on Artificial Intelligence,[6] this exercise could also expose which agencies administer, oversee, or regulate the use of AI decision systems, enabling a coordinated effort to make those systems explainable according to the NIST principles. At a minimum, NIST should describe how it envisions the explainability principles being used and what steps NIST plans to take to develop, refine, and apply the principles.

**NIST should make clear how different methods of explainability will impact the value of the explanation, and incorporate guidance for how AI developers should express the limitations of any explanation they provide**. For example, using an algorithm to explain the functions of a neural network may allow for a clearer explanation than would otherwise be possible, since the internal processes of

[4] *Four Principles of Explainable Artificial Intelligence* at 1 and FN 1, citing FCRA and GDPR.
[5] CDT will discuss this developing body of law in an upcoming paper, *Challenging the Use of Algorithm-driven Decision-making in Benefits Determinations Affecting People with Disabilities,* available soon at https://cdt.org/insights/report-challenging-the-use-of-algorithm-driven-decision-making-in-benefits-determinations-affecting-people-with-disabilities/.
[6] Executive Order on Maintaining American Leadership on Artificial Intelligence, (Feb. 11, 2019), https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/.

neural networks can be opaque and non-linear, but such an explanation is also limited in its accuracy, because there are necessarily contexts where the explainer algorithm might perform differently than the neural net. An explicit discussion of these trade-offs, how to weigh them, and whether there are practical thresholds related to accuracy or interpretability would help guide practitioners toward more consistent and more useful explanations.

**We caution NIST against using humans' abilities to explain their own decisions as a baseline for AI explainability**. We also urge caution, generally, when drawing analogies between the properties of automated systems and humans. Human decision makers are often unable to fully account for their decisions, even to themselves, relying on factors such as "gut feelings" to explain how they arrived at a conclusion or choice. Additionally, humans are known to exhibit dangerous and harmful biases, whether or not they are aware of them. Indeed, we understand far less about the inner workings of human decision making than we do about most AI systems. In contrast, as designers and developers of AI systems, humans should be able to exert more control over their development and use to ensure that AI systems are more explainable than human beings. Therefore, we should hold AI systems to a significantly higher standard for explainability.

As humans, our ability to adjust and improve algorithms also allows us to adjust and improve their explainability—and there is a pressing need to do so. AI systems, and machine decision making more generally, often confers a veneer of objectivity that makes humans less critical of decisions produced by these systems than they would be of the same decision from a human. Sadly, AI systems are rarely as objective as they may appear, making it even more important to expose their reasoning processes.

We thank NIST for taking up the issue of AI explainability and look forward to additional opportunities to engage with NIST in the future.

Respectfully submitted,

Stan Adams

1401 K Street NW, Suite 200 Washington, DC 20005