wiley

NEWSLETTER

Can AI Explain Itself? NIST Outlines Potential Approach to AI Explainability Standards

September 2020

Privacy in Focus®

Last month, the National Institute of Standards and Technology (NIST) released a draft paper on "explainability" in Artificial Intelligence (AI) decisionmaking, seeking public comment and helping to kick off what is likely to be an extended and extensive collaboration on AI standards. NIST has focused on developing standards and tools for "trustworthy" AI, pursuant to a high-profile Executive Order from last year. As part of that process, NIST is likely to focus on a range of issues that included explainability, bias, data, privacy, security, accuracy, and reliability when it comes to AI. Like its cybersecurity and privacy work, which has been enormously influential, NIST's current focus on AI promises to be impactful as AI becomes more widely adopted throughout the industry and more closely examined by policymakers and regulators.

Below, we summarize NIST's most recent work on explainability, highlighting opportunities for stakeholders to engage with NIST. While this will certainly not be the last opportunity for stakeholders to weigh in on NIST's AI work, it is one of the first opportunities to provide input on a critical standard and provide direct input on NIST's approach.

What Is Explainability?

Explainability is the concept that AI algorithms should produce explanations for their outcomes or conclusions, at least under some circumstances. Explainability is a core component of the U.S.-backed OECD AI Principles, for example, and already mandated by certain sector-specific laws in the U.S. that would apply to AI decisionmaking (such as in credit decisions). Explainability will be a key focus for

Authors

Duane C. Pozza
Partner
202.719.4533
dpozza@wiley.law

Kathleen E. Scott
Partner
202.719.7577
kscott@wiley.law

Practice Areas

Privacy, Cyber & Data Governance

stakeholders, particularly as they seek ways to approach issues like bias or protecting the security of AI systems that could be aided by AI-generated explanations.

What Is NIST's Approach to Explainability?

Based on discussions to date, it appears that NIST intends to develop a risk-based framework—like it has done with the Cybersecurity Framework and the Privacy Framework—for AI standards. That risk-based framework would incorporate standards that would include explainability.

The current draft paper—NISTIR 8312—sets forth four fundamental principles for explainable AI systems and five different kinds of explanations. It provides a potential framework for how explainability standards might be organized.

The four suggested principles are:

- Explanation. The paper suggests that AI systems should be capable of delivering accompanying evidence or reasons for all their outputs.
- Meaningfulness. Systems should provide explanations that are meaningful or understandable to individual users. An explanation need not be one-size-fits-all, and indeed groups of users may require different explanations, and the definition of a meaningful explanation may change over time.
- Accuracy. The explanation correctly reflects the system's process for generating the output. There can be different accuracy metrics for different groups – some audiences will require simple explanations that focus on the critical points but lack nuances, while others need detailed explanations to be fully accurate.
- Knowledge limits. The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output. Therefore, if a system has insufficient confidence in its decision, it should not supply a decision to the user. This can happen when an algorithm is not designed to answer a specific question, or when it is not sufficiently certain of its conclusion.

The five types of explanations proposed in the paper are:

- User benefit: Informing a user about a specific output.
- Societal acceptance: An explanation that is "designed to generate trust and acceptance by society" – particularly in the case that the AI application generates an unwanted outcome.
- Regulatory and compliance: Explanations to assist with audits for compliance in specific sectors – for example, dealing with regulation of self-driving cars.
- System development: An explanation to assist debugging, improvements, and maintenance.
- Owner benefit: An explanation to benefit the operator of a system.

There is certainly room to debate how this framework would be applied in practice. To take a few examples:

- The framework appears to suggest that an explainability requirement might apply for all AI decisions, even if there is a low level of risk in the outcome – which may not be feasible or optimal for many AI systems.
- The framework might be read as requiring a wide range of different *kinds* of explanations in a way that would be overly burdensome to development of the technology.
- Portions of the framework venture into how humans understand AI explanations, which falls outside typical technical standard-setting, and which raises an additional set of issues about the best ways to communicate information to humans.

In general, NIST has succeeded with its Cybersecurity and Privacy Frameworks when avoiding making value-based, policy judgments, and it will need to navigate that concern here, even as it seeks to help establish easily understood, implementable, and interoperable standards.

The deadline for comment on NIST's draft explainability paper is October 15, and NIST is seeking broad input from stakeholders. We also expect additional workshops in the coming months – following two August workshops, including one on bias – which will be open for public input.

Wiley's Artificial Intelligence practice counsels clients on AI compliance, risk management, and regulatory and policy approaches. Please reach out to the authors for further information.