July 19, 2019

Comment Template for Draft Plan for Federal Engagement in Developing Technical Standards and Related Tools for AI Technologies

COMMENT #	NAME OF COMMENTER	TYPE i.e., Editorial Minor Major	LINE # PAGE etc.	RATIONALE for CHANGE	PROPOSED CHANGE (specific replacement text, figure, etc. is required)
1	Courtney Bowman & Anirvan Mukherjee , Palantir Technologies Inc.	Major	P. 5, Line 78	"Interoperability" is a concept that requires a clearer articulation in the context of AI technologies.	We recommend succinctly defining the word "interoperability" in the text and including it in a glossary.
2	Courtney Bowman & Anirvan Mukherjee , Palantir Technologies Inc.	Major	p. 6, lines 115 - 121	The nature of AI applications is such that useful standards may only have applicability within specific domains. By offering principles rather than standards, NIST can ensure that the principles will be flexible to accommodate a wide range of AI applications across industries, both now and in the future.	We recommend re-writing this section with the framing of "Horizontally Applicable Principles" rather than "Horizontally Applicable Standards" for AI. A good model for writing effective principles are the Fair Information Practice Principles, which have broad applicability because they can be appropriately transposed to any given domain without attempting to be overly prescriptive.
3	Courtney Bowman & Anirvan Mukherjee , Palantir Technologies Inc.	Major	P. 7, lines 164 - 166	Comment: AI safety is a good example of why standards are better applied vertically rather than horizontally. Safety/risk evaluations will necessarily be very domain/application specific.	[comment only; no change suggested]
4	Courtney Bowman & Anirvan Mukherjee , Palantir Technologies Inc.	Major	P. 9, lines 184 - 188	Standard data sets have a positive utility, but also carry downside risks, because they present opportunities for gaming evaluation systems and frameworks. There are two aspects of this: (1) Academic researchers are often incentivized to try to establish improvement over the literature in an "apples-to-apples" way, which has led to a focus on using the same datasets over and over. Updating benchmark datasets or adding new ones will help diversify research. (2) Currently, many researchers have to use the same data as previous researchers to build on their work because their code is not published. Encouraging publication of the code behind academic results to reproduce findings in papers will make it easier for researchers to compare their approaches to other approaches by running the code from other approaches on new data, rather than having to use the same data as whoever published results using other methods. We recommend including language about diversifying benchmark datasets and also recommend encouraging the publication of code to mitigate the risk of gaming data sets used for training and testing of AI systems.	We recommend including language about (1) preventing the over-reliance of AI algorithms on static benchmark datasets; and (2) encouraging publication of code to reduce this reliance.

July 19, 2019

Comment Template for Draft Plan for Federal Engagement in Developing Technical Standards and Related Tools for AI Technologies

COMMENT #	NAME OF COMMENTER	TYPE i.e., Editorial Minor Major	LINE # PAGE etc.	RATIONALE for CHANGE	PROPOSED CHANGE (specific replacement text, figure, etc. is required)
5	Courtney Bowman & Anirvan Mukherjee , Palantir Technologies Inc.	Major	P. 9, lines 189 - 191	This entire bullet point is unclear. In order to confidently compound upon AI outputs towards reasoning, it's critical to adopt technology for understanding the consequences of AI on downstream consumers.	We recommend offering a clarification of this bullet point. For example, it might be appropriate to briefly mention model branching and/or provenance here and encourage the use of technology that allows simulation of outputs to better understand potential consequences.
6	Courtney Bowman & Anirvan Mukherjee , Palantir Technologies Inc.	Major	P. 9, lines 196 - 200	Testing is not an isolated, one-time requirement for functional AI. It would be helpful for NIST to articulate a more robust framework for testing.	We suggest adding langauge to articulate the need to apply testing in at least four stages of AI deployment: 1) training, 2) production roll-out, 3) periodic, on-going assessments, and 4) whenever existing AI outputs are repurposed for a new task that is outside of their original intent (i.e., with a different risk-reward surface area).
7	Courtney Bowman & Anirvan Mukherjee , Palantir Technologies Inc.	Major	P. 9, lines 201-204	Metrics are most meaningful, best applied, and most applicable in the context of specific AI domains.	We suggest adding this language to refine the focus of metrics: "Metrics related to specific applications of AI technology (as opposed to hardware performance) should reflect domain-specific and other context-dependent normative considerations."
8	Courtney Bowman & Anirvan Mukherjee , Palantir Technologies Inc.	Major	P. 10-11, lines 248 - 251	"Domain-specific" AI evaluations should be treated as a feature, not a bug. AI applications, like the human systems they are intended to synthesize, augment, or even replace must be regarded as fundamentally situated and contextualized in domain-specific expertise and environments.	We suggest reframing these lines to focus on broad-ranging AI **principles** that cut across AI applications, while focusing **standards** on specific application verticals.