We thank the National Institute of Standards and Technology (NIST) for providing an opportunity to review and comment on the draft "Proposal for Identifying and Managing Bias in Artificial Intelligence" ("the proposal").

To provide some context for our perspective, Palantir Technologies, Inc. ("Palantir Technologies," "we") is a software company that provides data integration, analysis, and decision making platforms. Our platforms are often used as a data foundation, development, and deployment infrastructure for AI, giving us insight into the challenges and successes of AI programs. In that capacity — and as a company experienced in working with the government, commercial, and non-profit sectors on data challenges — we have broad insights into the process of developing, training, and operationalizing AI that we believe may be useful in refining the proposal.

As a general matter, we agree that standards for identifying and managing bias are crucial for deploying responsible AI systems. However, we respectfully posit that this proposal and its underlying intent can have a greater impact and be implemented more effectively by deliberately narrowing the scope to a specific class of biases, namely forms of statistical bias and coupling that with concrete domain driven metrics, techniques and procedures for examining bias as opposed to the much more generic approach taken in the paper. We think this falls more squarely within the historical mandate of NIST as a standards promulgating institution.

In particular, we recommend the following:

1. **Recognizing the forms of bias relating to the application of AI technologies are manifold (as helpfully enumerated in the draft's annex), nonetheless, at the outset, NIST should endeavor to limit its guidance to the subset of bias types that relate directly to statistical decision making.** The proposal in its present form appears in places to take an expansive, all-inclusive approach to addressing questions of bias as they relate to AI systems. At the same time, the closest it comes to actually defining what "bias" means as a term draws upon an ISO, statistics oriented definition. The annex further lays out a seemingly exhaustive set of types of bias spanning domains of statistics, aesthetics, cognitive sciences, behavioral sciences, etc. This lack of clarity creates considerable confusion for readers who are left with an uncertainty about what to actually do to address identified forms of bias they may encounter. We suggest that this guidance may be most useful and actionable if the focus is narrowed to address the set of biases that converge under the heading of statistical bias (e.g., amplification, detection, evaluation, exclusion, measurement, population, sampling, selection, temporal, etc.). By limiting to statistical bias, NIST is operating more squarely within its mandate as a measurement and standards setting body, as statistical bias is by definition what is measurable in data and through outcomes.

   To be clear, in proposing to narrow the scope of NIST's bias guidance, we're not advising that other, potentially non-statistical, downstream externalities and biases be ignored or forgotten, but rather that they be passed on to be addressed by other qualified experts. For instance, a federal mortgage loan program may induce downstream geographic distribution changes for individuals *not* part of the program so understanding of societal impact requires data and research exogenous to the AI system. We encourage agencies to fund research to better understand specific, context-dependent classes of societal impact.

2. **Working from a more clearly defined and more focused scope of bias, the proposal could then convey an important point of threshold guidance for AI systems developers: namely determining whether AI is the right tool for the job in the first place.** The majority of AI-driven decision systems make statistical predictions either directly via probabilities or indirectly via some version of a "score" or other quantitative metric. Statistical computation systems in general should be caveated with an acknowledged categorical limitation, i.e., that not all decisions are statistical in nature, and that statistical decision systems are not appropriate in all situations. For instance, legal determinations of guilt or innocence using juridical reasoning are not usually statistical in nature. In other words, sometimes the best approach to de-biasing AI is identifying when certain problems are simply not fit for AI solutions.

   Even when a problem passes a categorical threshold that doesn't mean that AI should be deployed and some of the worst impacts of AI (including issues that cascade from AI ingrained bias) can be avoided by considering a threshold question before a project is even started: Is AI appropriate for this decision given the tradeoffs and risks inherent with AI, including issues related to perpetuating bias? The same rubric should be applied to any decision making process:

is this methodology, given its limitations appropriate to use in this context? Eliminating a methodology does not necessarily mean that the problem is left unsolved. For example, we often find that decisions that are high-risk often do not warrant AI approaches, and can be more effectively addressed through other, lower-risk approaches including UX/UI improvements; more complete, timely, and accurate data; and post-hoc analytics.

Statistical tools, however, can play a valuable role in certain classes of decision making; and where statistical methodologies apply, there are a number of broadly understood and researched forms of bias that we advise comprise the core focus of this document, as noted in (1) above.

3. **Even with a limited focus on addressing forms of *statistical bias*, NIST should be careful in passing *a priori* judgement about the ethical ramifications and impact of presence of bias in the abstract.** Bias is not inherently problematic and must be evaluated in the larger context and use of an AI system. AI systems never manifest in the abstract and invariably apply as situationally, contextually dependent tools in the world. This implies that even generalized insights about statistical bias may apply universally. Rather, insights about statistical bias must always factor in the context of application to determine whether the relevant forms of statistical bias are *in situ* desirable or undesirable.
    a. For example, the BERT AI model could be applied to classroom transcripts to investigate if teachers talk to black and white students (about the same lesson) differently. In this case, BERT is being used to discover societal bias. However, BERT itself is well known to have issues with bias representation. Thus, using BERT on the same data as a basis for predicting whether students will need extra help is inappropriate.
4. **The extension of the points above implies a recognition that the work of exhaustively addressing bias considerations in AI is a broader, unfinished project that will require further effort by bodies other than NIST.** Put simply, addressing the potential societal impacts of bias in AI systems writ large is beyond the reasonable scope of NIST as a standards-promulgating body, or any single entity for that matter. Even more so, no single codifiable set of standards across all domains of AI application will make sense (without being reduced to the most diluted and inactionable form), because bias assessments tend to require deep domain specialization and input.
    a. NIST's work is thus most impactful when establishing better *domain-specific statistical bias* standards by looking in detail at particular use cases, data sets, and the outcomes they generate. Other regulatory and decision making bodies are better suited to then apply, extend, or adjust these standards, and to consider if adjustments should be made to existing law or regulations on domain-driven bases. To illustrate why domain-specific considerations of bias are important, consider "FinTech" loan platforms vs. diagnostic assistants for radiologists vs. facial recognition on public transit. Each distinct area of AI application needs to take into account different processes, interaction modes with the AI, and societal impact. The guidance should help not just AI engineers, but also organizations purchasing and using these AI solutions and regulators overseeing their use. NIST should prioritize high-impact, risky, or sensitive use cases for deeper investment in standards. In all cases, recommendations should be specific about (1) the kinds of biases to check for; (2) the benchmark metrics to measure; and (3) the associated societal risks *in the applicable context*.

5. **The report conflates bias associated with AI systems with that associated with automation or any decision process. In most cases, standards developed should apply regardless of the decision mechanism**. We think it's worthwhile to consider which bias issues are particular to AI systems versus those that are particular to automation or decision processes broadly. By framing (statistical) bias as primarily related to the use of AI, NIST perhaps misses the opportunity to land a broader point about where similar concepts should apply. In the ideal situation, decision making, by any mechanism, should record contextual information relevant to the decision, the decision, and the outcome. This allows for, at a minimum, a post-hoc understanding of (statistical) bias issues. While AI deserved particular scrutiny in this regard on account of its reliance on potentially problematic data sets, standards around (statistical) bias should not be restricted to AI systems. Limiting to AI unnecessarily hems in the scope of this work.

6. **The document as written focuses exclusively on standards for bias in AI development, without recognizing the need for continued standards upon AI deployment and refinement.** AI systems are never finished and, if anything, often require *even greater* scrutiny and monitoring once deployed. In fact, many issues in the underlying AI mechanics

may not be observable or detected until the capability is fully deployed in production settings, even given the most rigorous testing for accuracy and bias in the lab. Our experience is that the emphasis, and the crux of building a good AI system, is the monitoring and iteration on deployed models, not merely refining models in the lab. The same goes for addressing issues of potential statistical bias in AI systems — these assessments, refinements, and corrections are a continuous process that extend well beyond the development phase.

This structure is further complicated in that AI models have many stakeholders: data scientists who develop them, engineers who deploy them, domain experts who vet them, end users/operators who use them, and people in society who feel the effects of those decisions. Each of these groups has a different understanding of the impact of bias that needs to be accounted for. The full, cumulative weight of embedded (statistical) biases may not be wholly analyzable until the complete system — with all of its accreted atomic parts — is in active use in the field.

We wish to reiterate that we applaud NIST's efforts to provide guidelines for addressing an important set of challenges in an evolving area of technology development and we thank NIST once more for this opportunity to provide feedback on its draft proposal. Our remarks are intended in the spirit of honing NIST's guidance to make it as useful and meaningful as possible for researchers, practitioners, users, and society at large. We welcome further opportunities to contribute to this and related efforts.

Sincerely,

Anthony Bak, Head of AI Engineering
Courtney Bowman, Global Director of Privacy and Civil Liberties Engineering
Megha Arora, Privacy and Civil Liberties Engineering Lead