Dear Ms. Tabassi,

UC Berkeley's Center for Human Compatible AI has prepared the following response to NIST's request for information (RFI) on the important topic of AI standards, as announced here: https://www.govinfo.gov/content/pkg/FR-2019-05-01/pdf/2019-08818.pdf

Our response is organized under the same headers as those used in the RFI itself, which are colored in mauve.

AI Technical Standards and Related Tools Development: Status and Plans

1. AI technical standards and tools that have been developed, and the developing organization, including the aspects of AI these standards and tools address, and whether they address sector-specific needs or are cross-sector in nature;

The cross-sector "standard model" for AI systems is based on the idea of optimal achievement of an explicit objective. This includes systems that find least cost solutions that achieve stated goals, game-playing systems that maximize win probability or expected score, decision-theoretic systems that maximize utility, dynamic programming and reinforcement learning algorithms that maximize expected rewards, and supervised learning algorithms that maximize predictive accuracy or minimize expected loss. This standard model is implicit in essentially all of the competitions and benchmarks in all fields of AI. The same basic template is followed in control theory (cost minimization), statistics (loss minimization), game theory (payoff maximization) and operations research (reward maximization). In all cases, the objective is assumed to be exogenously specified.

This standard model results from a mistake in framing and leads to systems that expose society to significant risks. The model is fundamentally inadequate for real-world AI systems, for the simple reason that humans are unable to specify objectives fully and correctly in real-world situations; indeed, humans may not know their own preferences regarding certain aspects of the world until they have acquired sufficient experience with those aspects. Inevitably, there is uncertainty as to the true objectives that the machine should optimize. Optimizing an incorrect objective can have arbitrarily negative consequences for humans; for example, social media content-selection algorithms that learn to maximize click-through have caused very serious harms to society. Thus, there is a disconnect between current AI methods and any notion that application of these methods will result in benefits to humans. In the future, as machines become more capable, the negative consequences of this disconnect could be much greater.

The Center for Human-Compatible AI (CHAI) is concerned with re-orienting the technical foundations of AI research in order to remove this disconnect: to ensure that, by design, AI systems are beneficial.

To eliminate the incorrect assumption of a known objective, CHAI has formulated the notion of an ***assistance game***, a formal game-theoretic model wherein one agent (typically, the machine) is tasked with assisting another (typically, the human). The machine's payoff is defined to be the payoff of the human, but the machine is initially uncertain as to what the human's payoff function is. In the assistance game framework, the machine defers to the human in a mathematically provable sense. For example, it will allow itself to be shut off when the human desires. The "standard model" is simply an extreme special case where the machine has perfect initial knowledge of the human payoff function.

Our current research is aimed at extending the assistance game paradigm -- for example, by allowing multiple humans and multiple machines and by accommodating human imperfection in a more general way -- so that it can become the dominant paradigm in AI system design. This would eliminate one major source of societal risk arising from the historical “objective maximization” framing.

The idea of systems that are beneficial by design also suggests a different notion of third-party audit for systems. Rather than testing only for bugs or security weaknesses, audits should (1) identify the scope of action of the system; (2) define notions of benefit and harm for any changes to world state (including human mental state) within the system’s scope of action; and (3) perform predictive and empirical analysis of whether the system is indeed beneficial within its scope of action. For example, does a video game induce addictive behavior or have a negative effect on user attention span? Does a content selection algorithm manipulate user opinions and attitudes? These kinds of questions are at the core of whether AI systems are truly assisting human beings and human society, and should be built into our technical definitions of success in AI system design.

2. Reliable sources of information about the availability and use of AI technical standards and tools;

The University of Oxford’s Center for the Governance of AI (https://www.fhi.ox.ac.uk/GovAI/) has written extensively on the potential impact of AI on society and how it might be governed, which is closely related to what standards are needed for AI systems to yield a positive benefit to society. Of particular note are:

- “AI Governance: A Research Agenda”: https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf
- “Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development”: https://www.fhi.ox.ac.uk/standards-technical-report/
- “Artificial Intelligence: American Attitudes and Trends”: https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/us_public_opinion_report_jan_2019.pdf
- “Syllabus: Artificial Intelligence and International Security”: https://www.fhi.ox.ac.uk/wp-content/uploads/Artificial-Intelligence-and-International-Security-Syllabus.pdf
- “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation”: https://maliciousaireport.com/
- “Deciphering China’s AI Dream”: https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf

The following report from the Center for New American Security makes the important argument that international standards for the safe and appropriate use of emerging technologies in general are needed to protect American interests:
https://s3.amazonaws.com/files.cnas.org/documents/CNASReport-Technology-Roulette-DoSproof2v2.pdf

Georgetown University’s recently formed Center for Security and Emerging Technology also promises to be a leading source of information on policy relating to emerging technologies, including AI:
https://cset.georgetown.edu/research/

3. The needs for AI technical standards and related tools. How those needs should be determined, and challenges in identifying and developing those standards and tools;

At least three new types of standards seem particularly important to be established:

3a) Published deliberation of side-effects.

i) **In industry:** When an AI system is deployed in the United States—by either a domestic or foreign institution—its creators should be required to write about and publish a deliberation of its potential negative or otherwise significant side effects on individuals and American society. Currently, companies are expected only to advertise the benefits of their technology. But, as designers of the technology being deployed, tech companies will always hold an earlier and greater understanding of the details of what they are building than outsiders will possess. As such, companies are in a privileged position to notice and warn society about the potential for negative side effects of their AI systems upon the United States, and the world more broadly. As such, NIST should establish an expectation that technology companies will use the privileged position to warn society about the potential negative side effects of their technology. Similar requirements already exist for pharmaceutical companies to establish safety as well as efficacy, and for construction projects to file environmental impact reports analyzing possible negative consequences.

Once the impacts are better understood, perhaps companies could be required to warn their users about the impacts. For instance, if social media technologies are found in general to be addictive for some users, social media companies could be required to issue a warning to their users, such as:

"Social media apps may be addictive for some people. Users addicted to a social media app might continue to use the app even if the way they use the app makes them unhappy or unable to do their jobs. They might also use the app in ways that make their friends, family, or coworkers upset with them, but be unable to bring themselves to stop using it. Please pay attention to whether you are addicted to this app, and seek the help of friends and/or professionals to stop using it if you find you might be addicted, or call 1-800-APP-STOP."

(This example is provided for the sake of illustrative concreteness, rather than as a specific recommendation.)

ii) **In academia:** Academics are currently also expected to advertise only the positive applications of their research. In fact, there is a social expectation that discussion of negative side effects from the misapplication of research outputs could impair funding opportunities relative to other projects that do not acknowledge any downside-risk. This social standard needs to be changed. For example, currently, every proposal to the National Science Foundation must include a "broader impact" statement briefly summarized as follows:[1]

[1] "Perspectives on broader impacts," National Science Foundation, 2015.

"Broader impacts—the potential to benefit society and contribute to the achievement of specific, desired societal outcomes."

Notice that this requirement does not include any obligation to discuss potential negative consequences and approaches to mitigate those consequences. While some might argue that all knowledge is good, there is no doubt that research on, say, gain-of-function modifications to virulent disease organisms or low-cost, low-tech methods of uranium enrichment should clearly state and analyze the potential negative consequences for humanity. NIST should anticipate a time (if it is not already here) when AI systems present comparable risks to human well-being.

We recommend, therefore, that standards for proposal preparation and evaluation should be modified to reward frank discussion of potential risks and constructive approaches to risk mitigation and to punish concealment of risks or failure to address mitigation. To draw out discussion of risks, funding agencies could then ask concrete questions to stimulate discussion. Establishing this new professional standard will help to create an earlier, more nuanced awareness of how research should be applied judiciously to benefit individuals and society. It would also trigger earlier research innovations for addressing and/or mitigating negative side effects of technologies: the sooner it becomes common knowledge among a field of researchers (such as AI researchers) that a certain negative side effect of their work is looming, the sooner the researchers can support and collaborate with each other on designing solutions to mitigate that effect.

3b) User cohort audits in industry. Any AI-based software system used in the United States to interface with a sufficiently large number of American users should be required to commission a third party auditor to conduct a controlled user-cohort study assessing the impacts and correlates of the software's usage.

To stimulate action-oriented discussion, we describe here an example of what could be involved in such cohort studies; the example is meant to be illustrative, and is not intended as a specific recommendation as to what details such a study should or should not be required to include:

User cohorts could be studied over time spans of one month, one year, three years, and ten years. The studies could measure correlations with variables that the company cannot easily control (e.g., who chooses to use the company's software) as well as the causal impacts of variables that the company can randomize and control for (e.g., which individuals the company chooses to advertise to). The correlates and impacts measured could include variables relevant to the vigor and resilience of American individuals and institutions, such as:

i) ***Addiction.*** *Do American users regret using this software? Are they able to stop using the software in ways that they regret, or that upset other people? Do they wish other people would stop using the software?*

ii) ***Cognitive capabilities.*** *Are American users of this software system experiencing working memory impairments, or improvements? Long-term memory impairments, or improvements? Attentional deficit, or enhancement? Decline or advancement in literacy and/or numeracy?*

iii) ***Interpersonal capabilities.*** *Are American users of this system able to work well or poorly with others on collaborative tasks? Do they experience increased or decreased*

anxiety in commonplace social situations? Are they more or less able to understand and empathize with the beliefs and desires of their friends, family, and co-workers?

3c) Responsible publication standards. If a researcher or company develops an AI capability that could be easily and widely misused by enemies of the American public, what process can they use to decide whether that risk outweighs the benefit of publication? Currently there are almost no institutionalized procedures at universities or large tech companies for evaluating such questions, although one exception can be seen demonstrated by OpenAI (see #4 below). Just as universities have committees for assessing the appropriate use of human subjects in experiments, so, too, should companies and universities appoint committees or develop other organized processes for helping researchers to evaluate whether their work might pose a risk to the American or global public if released. OpenAI and the Partnership on AI have begun to experiment with such an organized process:
https://www.partnershiponai.org/when-is-it-appropriate-to-publish-high-stakes-ai-research/

Eventually, oversight of AI development could be evaluated by a body similar to the National Science Advisory Board for Biosecurity (NSABB), who issued the following recommendations in 2016 regarding Gain of Function Research in genetics:
https://osp.od.nih.gov/wp-content/uploads/2016/06/NSABB_Final_Report_Recommendations_Evaluation_Oversight_Proposed_Gain_of_Function_Research.pdf
In particular, NSABB has recommended that "The U.S. government should undertake broad efforts to strengthen laboratory biosafety and biosecurity and, as part of these efforts, seek to raise awareness about the specific issues associated with GOF research of concern."

4. AI technical standards and related tools that are being developed, and the developing organization, including the aspects of AI these standards and tools address, and whether they address sector-specific needs or are cross sector in nature;

Regarding responsible publication standards (#3c above), OpenAI has recently begun to experiment with possible responsible publication models. Specifically, OpenAI made the decision not to release the weights of a particular neural network called GPT-2, "Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale":

- https://openai.com/blog/better-language-models/

Whether or not OpenAI's technology was in fact dangerous for society in this case, we believe their decision not to publish sets a valuable and important social precedent that encourages researchers to think about the impact of their work before releasing it. The details of their decision have fallen under some criticism; e.g. if the neural weights were dangerous to release, perhaps it was unwise to release the algorithms as well. This controversy suggests a need to reach broad agreement on standards for risk evaluation and publication. To this end, the Partnership on AI met with OpenAI and simulated a hypothetical standardized review process for deciding when to publish highly impactful AI research findings:

- https://www.partnershiponai.org/when-is-it-appropriate-to-publish-high-stakes-ai-research/

5. Any supporting roadmaps or similar documents about plans for developing AI technical standards and tools;

CHAI personnel and partners have been involved in producing the following calls to action for research to develop standards for safer, more robust AI systems:

- An open letter on “Research Priorities for Robust and Beneficial Artificial Intelligence” https://futureoflife.org/ai-open-letter/

- Research Priorities for Robust and Beneficial Artificial Intelligence: https://futureoflife.org/data/documents/research_priorities.pdf

We also recommend the following research agenda released by researchers who are now mostly affiliated with OpenAI:

- Concrete AI Safety Problems: https://openai.com/blog/concrete-ai-safety-problems/

6. Whether the need for AI technical standards and related tools is being met in a timely way by organizations;

It does not seem to us that standards and related tools for addressing the concerns we have raised in this response are being developed in a timely manner.

Symptom 1: No urban planning for AI. There is as yet no equivalent of “urban planning” in the deployment of AI systems. We would never allow a private company to build a new urban development affecting the lives and environments of thousands of people without a degree of accountability to local and state governments, e.g., in the form of zoning laws and environmental impact reports. By comparison, it seems foolish to allow technology companies to deploy AI-based applications with widespread impact upon (literally) billions of people with no process for auditing the decision processes behind those deployments. Yet, this is the status quo.

Symptom 2: No expectation to reflect on negative side effects. A related problem is the widespread professional expectation—essentially a ‘social standard’—that companies and researchers do not openly discuss the potential negative side effects of their work, for fear of losing funding or public favor relative to other researchers or companies. Specifically:

- researchers at present mostly only discuss the potential benefits of their own work, with warnings of potential downsides elided or fully excluded from their own publications; and
- companies at present mostly only discuss the positive applications of their software, with reasoning about misuse and/or side-effects kept internal to the company and usually completely unpublished unless public outcry demands accountability.

Researchers will from time to time criticise *other researchers*, or develop techniques to address perceived safety issues with the methods of *others*. But there is little expectation that a researcher should use their own privileged position of understanding their own work in order to warn others about their work’s potential negative side effects upon the American or global public. The same is true of technology companies: no one expects a tech company to warn the public about how their products could be used in ways that would harm either Americans specifically, or the world at large.

7. Whether sector-specific AI technical standards needs are being addressed by sector-specific organizations, or whether those who need AI standards will rely on cross-sector standards which are intended to be useful across multiple sectors.

The suggested standards described in #3 above should be sector-general, namely, Published Deliberation of Side Effects, User Cohort Studies in Industry, and Responsible Publication Standards.

8. Technical standards and guidance that are needed to establish and advance trustworthy aspects (e.g., accuracy, transparency, security, privacy, and robustness) of AI technologies.

"Red team, blue team" methods from cybersecurity should also be used as a widespread standard for assessing the "trustworthiness" of AI systems along dimensions of safety, transparency, security, privacy, and robustness. That is to say, for any important application, the researchers assessing the trustworthiness of an AI system should not be the same as the researchers who developed it.

9. The urgency of the U.S. need for AI technical standards and related tools, and what U.S. effectiveness and leadership in AI technical standards development should look like;

The United States should seek to institute *international* standards for AI development that will protect American interests. There are at least two channels of impact through which internationally adopted standards will affect the United States:

- *Impacts directly on Americans.* Standards for software used by American companies and individuals should apply irrespective of where the software was manufactured.
- *Impacts on US security.* For instance, international standards to enable tracking the manufacturing and distribution of AI-relevant hardware could make the use of AI technology in the global marketplace more traceable and accountable, thereby increasing the difficulty for malicious actors to assemble sufficient AI resources to threaten US security specifically.

As such, where possible, NIST and the Federal government should consider ways in which American standards could be adopted and/or enforced internationally.

10. Where the U.S. currently is effective and/or leads in AI technical standards development, and where it is lagging;

Lagging. The US seems to be lagging behind Europe in the establishment of transparency and accountability standards for technology companies, as evidenced, for instance, by the instatement of the General Data Protection Regulation in Europe:
https://eugdpr.org/

Leading. The US remains competitive in the establishment of standards for training and testing AI systems. For instance, the widely used MNIST training dataset was developed by researchers in New York in 1999, from images taken from a larger dataset provided by NIST:
https://en.wikipedia.org/wiki/MNIST_database
http://www.pymvpa.org/datadb/mnist.html

The ImageNet dataset has been developed primarily at Stanford, Princeton, CMU, Michigan, and UNC Chapel Hill, since 2006:
https://en.wikipedia.org/wiki/ImageNet
http://image-net.org/about-people

Canada, a close political and economic ally of the United States, was home to the development of the CIFAR-10 dataset, in 2009:
https://en.wikipedia.org/wiki/CIFAR-10
https://www.cs.toronto.edu/~kriz/cifar.html

More recently, in 2016, the San Francisco-based company OpenAI has developed standards for the testing and benchmarking of reinforcement learning systems, called OpenAI Gym, which is now widely used in academia and industry:
https://gym.openai.com/

11. Specific opportunities for, and challenges to, U.S. effectiveness and leadership in standardization related to AI technologies;

Challenges. As long as companies and researchers are not expected to reflect upon and disclose the potential negative impacts of their own products, discourse will always lag behind impacts, and standards will always be developed in a reactive rather than proactive mode, as is the case today.

Opportunities. The above challenge could be addressed at least in academia through Federal funding initiatives which require individual researchers to consider and report on the potential negative side effects of each their own research outputs. These reports could trigger a wave of innovation to develop new and interesting techniques for addressing the reported side effects.

12. How the U.S. can achieve and maintain effectiveness and leadership in AI technical standards development.

See “Opportunities” under question 11.

13. The unique needs of the Federal government and individual agencies for AI technical standards and related tools, and whether they are important for broader portions of the U.S. economy and society, or strictly for Federal applications;

To the extent that America depends upon the preeminence of its technology sector as a source of economic of strategic advantage, America will need to deploy Federal funding to support AI researchers and companies in their efforts to protect American society specifically from negative side effects of their own AI research outputs.

Regarding academia, this will require significant federal funding increases in the areas of AI safety, AI transparency (especially mechanistic transparency, i.e., transparency that gives accurate insight into the internal processes of an AI system viewed as a machine), and human/AI interaction. It will also require changes to grant proposal expectations from institutions like the NSF and DARPA, to encourage

researchers to consider and discuss the potential negative side effects of their own work, as described in #3a above.

Regarding industry, this should involve Federal funding and support for user cohort studies as described in #3b above.

See also:
https://www.cnas.org/publications/reports/technology-roulette

14. The type and degree of Federal agencies' current and needed involvement in AI technical standards to address the needs of the Federal government;

The Federal government should seek to employ and retain more expert computer scientists in high ranking advisory roles to assist America in understanding its position in the technology sector, and the potential positive and negative impacts of the domestic and international technology sector upon the American population.

15. How the Federal government should prioritize its engagement in the development of AI technical standards and tools that have broad, cross-sectoral application versus sector- or application-specific standards and tools;

It seems to us that broad, cross-sectional promotion of principles from the Federal government would be most appropriate and effective, rather than industry-specific micro-management. This should involve the employment of computer science experts from a variety of industries, who are expected to report frankly and without bias to the Federal government on the potential positive and negative impacts of their industry or sector upon American individuals and society.

16. The adequacy of the Federal government's current approach for government engagement in standards development,[4] which emphasizes private sector leadership, and, more specifically, the appropriate role and activities for the Federal government to ensure the desired and timely development of AI standards for Federal and non-governmental uses;

Expertise from private sector leaders is needed to inform the Federal government's opinions on the impact of nascent AI technologies. However, incentives in the private sector are driven by profit rather than by side effects upon the welfare of individuals or the United States as a whole. If the technology sector were to collectively present a threat to American well-being or security, there may be little short-term incentive for multinational corporations to raise or validate this concern. As such, the Federal government should seek separately the expertise of computer scientists and AI researchers who are not employed in the private sector, such as those in academia. To reduce the incentives and social pressure for researchers in academia to give overly optimistic views of the impact of AI research and technology, changes to expectations in Federal funding proposals will be needed, as described in #3a above.

17. Examples of Federal involvement in the standards arena (e.g., via its role in communications, participation, and use) that could serve as models for the Plan, and why they are appropriate approaches; and

Involvement in communications. We suspect that broad Federal endorsement of principles such as 'AI safety', 'AI transparency', and 'AI accountability' could be highly effective and influential in both academia and industry. Researchers are more likely to work on topics where they expect to find collaboration and recognition. Since statements from the Federal government can rightly be expected to correlate with Federal funding priorities and hence collaboration opportunities for researchers, statements from the Federal government help to catalyze researchers to begin developing an interest in topics that might otherwise present uncertain career prospects. "Human/AI interaction", "AI safety", "AI transparency", "AI accountability", and are among topics that could benefit from such statements of support and funding.

Involvement as a buyer. The Federal government, including the DoD, could require certain standards of safety, accountability, and transparency in the technology it purchases from the private sector. For instance, if an AI system fails catastrophically in a combat scenario that harms American troops or America's relationship with another nation, there should be systems in place for transparency and accountability of the AI system, its components, and its manufacturers.

18. What actions, if any, the Federal government should take to help ensure that desired AI technical standards are useful and incorporated into practice.

These are covered in #17.