OpenAI

June 10, 2019

Elham Tabassi,
National Institute of
Standards and Technology
100 Bureau Drive, Stop 200
Gaithersburg, MD 20899

Dear Ms. Tabassi,

On behalf of OpenAI, we are submitting comments in response to the request for information regarding 'Artificial Intelligence Standards' (Docket No. 190312229-9229-01. Document citation: 84 FR 25756.)

This response will outline how NIST can help to create a constructive environment for the development of AI standards within America and internationally. This response will focus specifically on standards for "reliable, robust, and trustworthy" systems. NIST's support of these standards and associated measurement and benchmarking initiatives can contribute to a flourishing, safe global market for AI services and products, and can enhance the US government's capacity to effectively oversee increasingly rapid AI technology development.

For this response, we'll specifically focus on the relationship between standards, safe and robust AI technology, and the US's ability to be globally competitive at AI technology development.

About OpenAI

OpenAI is an artificial intelligence research company based in San Francisco whose mission is to ensure that artificial general intelligence benefits all of humanity, and is attempting to build safe and beneficial AGI. OpenAI's work is primarily built around three areas: technical capabilities research and development; AI safety research and development; and policy work, which is a mixture of advocacy for specific positions relating to building an informed, responsive international government-driven AI policymaking environment. For this response, we draw on our experience in developing cutting-edge technical AI systems, and developing approaches for creating increasingly autonomous systems that satisfy constraints relating to safety, namely predictability, robustness, and interpretability, among others.

OpenAI

What do we mean by "standards"?

For the purpose of this response, we define standards as follows: broadly agreed upon techniques for assessing the capabilities of a given artificial intelligence technique within a specific application context; and reference data and data-generating systems (eg: software simulators).

Standards for technical forecasting for safer AI development and deployment.

Question(s) this responds to:
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;

Judgment based forecasts and predictions of future capabilities, economic results, and geo-political outcomes form a critical input into AI policy, military, and industry decisions. However, many individuals and organizations make such predictions and it's difficult to know what weight to give their predictions when they disagree substantially. Accountability for such predictions is so bad that many pundits were shown to have no more foresight than "dart-throwing monkeys" by Phillip Tetlock in his book Expert Political Judgment in 2005.

A standard for judgment-based forecasts and a track record of foresight from NIST would help create the top-level demand for well-defined, accurate, decision-relevant forecasts from all parties looking to inform federal agencies. It would promote the measurement of and incentivization of good judgment from analysts, think tanks, and subject matter experts.

An IARPA competition[1] that ran from 2011, set the standards for a) a well-formed prediction and b) a track record of foresight. An example of a well-formed prediction follows: Good judgment open currently estimates a 5%[2] chance Bashar

[1] https://www.iarpa.gov/index.php/newsroom/press-releases-and-statements/970-iarpa-announces-publication-of-data-from-the-good-judgment-project?highlight=WyJnb29kIiwianVkZ21lbnQiLCJqdWRnbWVudCdzIiwiZ29vZCBqdWRnbWVudCJd

[2] https://www.gjopen.com/questions?filter=featured

al-Assad cease to be president of Syria before 1 January 2020? It's well formed, because (i) this future observation being predicted is unambiguous (ii) the prediction has probability rather than imprecision natural language used to convey uncertainty (iii) the prediction is time bound, so we can be sure that we'll be able to resolve it at some point. A relevant track record of foresight consists of good performance on past well-formed predictions using a proper scoring rule, like the Brier score, in the same domain.

Phillip Tetlock, working under an IARPA grant, led a team that demonstrated an impressive amount of skill in the task of making judgment based predictions[3] in winning that IARPA competition. The organization he leads, Good Judgment Inc[4], has an impressive track record in policy relevant judgment based forecasts. Standards around technical forecasting from NIST would incentivize more individuals and organizations to produce high quality forecasts, which would be a useful input into decisions around AI, but also the progress of technology generally and geopolitical outcomes.

It would be helpful for NIST to create standards around a) well formed predictions b) a track record of foresight. Such a standard would help ensure best practices are used when organizations forecast technical AI progress and AI policy outcomes.

Standards for trustworthiness and "AI safety"

Question(s) this responds to:

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.

By trustworthiness and AI safety, we mean the set of techniques that can give us sufficient confidence in an autonomous system that we can deploy it into the world. AI safety encompasses a broad set of situations, including things like: guaranteeing a system won't cause physical harms to humans when they interact with it; ensuring that systems will check-in with humans about critical decisions

[3] https://hbr.org/2016/05/superforecasting-how-to-upgrade-your-companys-judgment
[4] https://goodjudgment.com/

relating to actions they're about to take; and ensuring that the recommendations made by AI systems are respectful of human constraints and the larger contexts in which the systems operate. Further examples of what we mean by trustworthy and safe AI - and what happens when these features are not present - are available in *Concrete Problems in AI Safety* (Amodei et al, 2016) and *Specificiation gaming examples in AI (Krakovna, 2018)*[5].

In practice, this means techniques that can let us predict and to a lesser extent direct the ways it will solve tasks and the sorts of things it will and won't do while solving these tasks, as well as technical metrics to help us measure its performance attributes. We also mean safety techniques that are applied to machine learning-based systems, and primarily ones which rely on deep learning - stacked layers of neural networks.

At OpenAI, AI safety encompasses techniques that can increase the predictability of a given system, provide assurances that increasingly powerful systems will operate according to the (human) values imparted to them by their developers or operators, and evaluate increasingly capable systems.

What standards have to do with AI safety:

Question(s) this responds to:

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

[5] Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, "Concrete problems in AI safety" (2016). https://arxiv.org/abs/1606.06565 . Krakovna, Victoria, "Specification gaming examples in AI" (2018). https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/

Standards are one of the mechanisms by which we can hope to increase the predictability, robustness, and reliability of an AI system. At OpenAI, we are conducting research into AI safety with the goal of developing a set of tools, techniques, and procedures for researching and developing 'safe' AI systems. Because AI research is a fast-moving area, there is today insufficient agreement between different technical actors about what constitutes a standard for assuring the safety of a given AI system.

To get a sense of how some existing research could eventually contribute to the formation of a standard, here is an outline of an existing (as-yet unpublished) OpenAI research project and its potential relationship to standards.

- Environments for safe exploration[6]:
  - Description: These are virtual worlds full of hazards for a simulated AI agent, where the goal is to learn to explore the world safely, and achieve a given task without generating negative externalities.
  - Motivation: Today, we want to develop more sophisticated algorithms capable of exploring an environment without generating unexpected problems. By designing a suite of environments to test against, we can develop algorithms and test their performance here, and can also learn how to build more sophisticated environments to test for more advanced capabilities.
  - Relationship to standards: Today, such environments are an area of active research, but it's likely that in the future the community could standardize on some set of environments. We may want to, for instance, benchmark agents to be deployed in a given context (eg, a

[6] For additional references on algorithms and environments that relate to safe exploration, please check out: 'Constrained Policy Optimization', published to the Berkeley AI Research blog in late 2017 https://bair.berkeley.edu/blog/2017/07/06/cpo/ ; 'AI safety gridworlds' from DeepMind https://deepmind.com/research/publications/ai-safety-gridworlds/; 'Safe Exploration in Continuous Action Spaces' from DeepMind https://deepmind.com/research/publications/safe-exploration-continuous-action-spaces/; and 'A Lyapunov-based approach for safe RL algorithms' from Facebook AI Research https://ai.facebook.com/blog/lyapunov-based-safe-reinforcement-learning/.

factory), against a standardized set of environments to test for the algorithm's ability to safely explore the space and achieve its goals.

Similarly to the above example, we can imagine wanting to have standard methods of evaluation to help us think about the advancement of other AI capabilities. For instance, we may ultimately want a standard way to measure the improvement in the outputs of generative models, for instance synthetic imagery generated by computers - to do this, we'll need standardized methods to use to measure the capabilities of AI systems in this regard. Today, that's a research problem, with multiple groups coming up with their own assessment criteria. In a few years, it's likely that a standardized assessment criteria will emerge here, and NIST may want to help support such a standard.

Specific things NIST can do to further the formation of standards for AI safety:

Question(s) this responds to:

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

NIST has a range of powerful levers to encourage the formation of standards for AI safety. One of the most powerful ones is its ability to convene people across industry, academia, and the government to work on this. It would be helpful for NIST to coordinate and host workshops for a mixture of industry, academia, and government participants to provide an overview of existing efforts towards work on AI safety, and potential routes to standardization.

The "AI Safety Testbed". In the same way that NIST today hosts the 'Manufacturing Robotics Testbed', we could imagine NIST also investing in a 'AI Safety Testbed', which would likely be a facility involving the following ingredients:

- A dedicated compute cluster for benchmarking and assessing systems.
- A small robotics facility (which could be within the Manufacturing Robotics Testbed) for evaluating certain AI safety techniques on real robots.

- A large, open-to-all (including industry) software-based 'AI safety testing suite', which could host a standardized set of benchmark environments to test safety properties within.

We imagine that the formation of an AI Safety Testbed could naturally complement NIST's coordination work in the area of AI safety, and would help industry and academia to align - via NIST - on common approaches to common problems. Such an initiative will likely require further federal funding on safety research in order to support a more timely delivery of standards.

How NIST can help further the American AI ecosystem through international engagement

Question(s) this responds to:

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

AI safety is, by nature, a borderless concept - AI systems today are deployed in a transnational manner, with organizations typically developing a system, then deploying it widely. Additionally, while some standards can be applied to AI products and services after they have been developed, our suspicion is that many of the techniques needed to guarantee the safety of a system will need to be applied during the development of the system itself - for instance, by exposing a certain system to a pre-agreed upon set of standardized environments during training (as discussed above), or potentially by training the system with additional objectives designed to encourage pre-agreed-upon safe behavior.

For that reason, it's critical that the US take a leadership position in the formation of international standards for AI technology. Our suspicion is that multiple actors will develop large-scale AI systems in the coming years and will eventually want to deploy such systems across the globe. To do this, they'll naturally seek to make it easier for regulatory infrastructures to accommodate such systems. We think that international standards are one bit of work that can be done today which can prepare regulatory infrastructures for large-scale, multi-purpose AI systems.

OpenAI

By contributing to the development of international standards for the safety of AI systems, NIST can achieve the following:

- Guaranteed American involvement in the formation of international standards, which will be crucial to the safe deployment and trade of AI systems.
- Create an international community of concern focused on the technical safety and assurance of AI systems; such a community will invariably be useful for further standards and assessment measure development in the future.
- Increase NIST's own capacity to effectively oversee the formation of standards in AI; by developing more expertise oriented around encouraging the creation of international standards for AI safety, NIST will naturally further develop its own internal talent with regard to AI measurement, assessment, and standard-setting.

3180 18th Street, Suite 100 San Francisco, California 94110