Request for Information (RFI):
Developing a Federal AI Standards Engagement Plan

84 Fed. Reg 18490 (May 1, 2019)
84 Fed. Reg 25756 (Jun. 4, 2019)
Docket # 190312229-9229-01

June 10, 2019

Google welcomes the opportunity to provide comments in response to the National Institute of Standards and Technology's (NIST) Request for Information (RFI) on Artificial Intelligence Standards.

Google's mission is to make the world's information universally accessible and useful. In pursuing this mission, Google has continually invested in advanced technologies like AI, which have become integral to Google's core services, such as Search, Gmail, Translate, and Photos. Google has also invested in platforms and tools designed to enhance innovation in the broader AI ecosystem and help individuals and organizations develop and deploy AI in their respective domains. In 2015, Google open-sourced its own machine learning software library TensorFlow, which has since been used to advance AI in areas as diverse as protecting wildlife and detecting disease.[1] Additionally, Google Cloud offers a range of AI-powered products and services for the enterprise world, from hardware accelerators to train and run AI models to pre-built APIs for computer vision, natural language processing, and more.[2]

Driving all of these efforts is our vision to democratize access to AI and help realize the positive benefits of this technology for individuals and society. Just as AI is creating new opportunities, however, it is also raising questions about its responsible development and use, particularly as it pertains to areas such as explainability, unfair bias, privacy, and safety. It is essential for stakeholders across society to engage in an ongoing dialog on these questions and consider the full range of tools available to promote responsible AI.

Standards can play a constructive role in encouraging the adoption of technical best practices across the development and use of AI. This work is not starting from scratch. Industry and academia have already invested significantly in the development of technical best practices, and international standards bodies, from the Institute of Electrical and Electronics Engineers

[1] https://www.blog.google/technology/ai/tensorflow-smarter-machine-learning-for/
[2] https://cloud.google.com/products/ai/

(IEEE) to the International Organization for Standardization (ISO), have established workstreams focused on AI standards. NIST and the US Government should play a leading role in these fora to promote consistency among global standardization efforts.

It is also important to distinguish areas where standards for AI generally may improve the development and use of AI across a broad range of application areas, versus where it may be more appropriate to take a domain- or sector-specific approach to standards development. For example, NIST can play a constructive role in defining what the different forms and levels of explainability are for AI systems generally, but other agencies or entities may be better positioned to speak to which particular form or level of explainability is most appropriate for a specific application.

The first section of our comment focuses on Google's views on the main areas for developing responsible practices and standards for AI.[3] The second section focuses on Google's recommendations on how the US Government can promote healthy standards development for AI both domestically and internationally.

I. Google's recommended practices for Responsible AI

As a discipline of software development, AI development should follow general best practices for software systems,[4] together with practices that address considerations unique to machine learning. At Google, we have additionally identified four broad areas where AI systems require particular thought: fairness, explainability, privacy, and security. Though not all of these practices may be appropriate to be codified into standards, we share them here to provide a sense of where the current state of research is and inform potential standards as the field matures. This should not be viewed as a static list; rather, Google continues to invest in iterating on these practices and developing new ones, as the field of AI generally continues to grow and evolve.

Fairness

In addition to consumer applications, AI systems can be used for critical tasks, such as predicting the presence and severity of a medical condition, matching people to jobs and partners, or identifying if a person is crossing the street. Such computerized assistive or decision-making systems have the potential to be fairer and more inclusive at a broader scale than decision-making processes based on ad hoc rules or human judgments. The risk is that any unfairness in such systems can also have a wide-scale impact. Thus, as the impact of AI increases across sectors and societies, it is essential to work towards systems that are fair and inclusive for all.

[3] More examples of research, tools, and training materials we have developed to promote responsible AI practices can be found here: https://ai.google/responsibilities/responsible-ai-practices/
[4] https://techdevguide.withgoogle.com/

This is a challenging task. First, ML models learn from existing data collected from the real world, so a model may learn or even amplify problematic pre-existing biases in the data based on race, gender, religion or other characteristics. For example, a job-matching system might learn to favor male candidates for CEO interviews, or assume female pronouns when translating words like “nurse” or “babysitter” into Spanish, because that matches historical data.

Second, even with the most rigorous and cross-functional training and testing, it is a challenge to ensure that a system will be fair across all situations. For example, a speech recognition system that was trained on US adults may be considered fair and inclusive in that context. When used by teenagers, however, the system may fail to recognize evolving slang words or phrases. We might also discover unexpected segments of the population whose speech it handles poorly, for example people speaking with a stutter or uncommon regional dialect. Use of the system after launch can reveal unintentional, unfair blind spots that are difficult to predict.

Third, there is no standard definition of fairness, whether decisions are made by humans or machines. Identifying appropriate fairness criteria for a system requires accounting for user experience, cultural, social, historical, political, legal, and ethical considerations, several of which may have tradeoffs. Is it fairer to give loans at the same rate to two different groups, even if they have different rates of payback, or is it fairer to give loans proportional to each group’s payback rates? Are either of these the most fair approach? At what level of granularity should groups be defined, and how should the boundaries between groups be decided? When is it fair to define a group versus better factoring on individual differences? Even for situations that seem simple, people may disagree about what is fair, and it may be unclear what point of view should dictate policy, especially in a global setting.

Addressing fairness and inclusion in AI is an active area of research, from fostering an inclusive workforce that embodies critical and diverse knowledge, to assessing training datasets for potential sources of bias, to training models to remove or correct problematic biases, to evaluating machine learning models for disparities in performance, to continued testing of final systems for unfair outcomes. In fact, ML models can even be used to identify some of the conscious and unconscious human biases and barriers to inclusion that have developed and perpetuated throughout history, bringing about positive change. Far from a solved problem, fairness in AI presents both an opportunity and a challenge. Google is committed to making progress in all of these areas, and to creating tools, datasets, and other resources for the larger community.

Recommended Practices for Fairness

1. Design models using concrete goals for fairness and inclusion:
    - Consider how the technology and its development over time will impact different use cases: Whose views are represented? What types of data are represented? What’s

being left out? What outcomes does this technology enable and how do these compare for different users and communities? What unfair biases, negative experiences, or discriminatory outcomes might occur?
- Set goals for systems to work fairly across anticipated use cases: for example, in X different languages, or to Y different age groups. Monitor these goals over time and expand as appropriate.
- Design algorithms and the objective function to reflect fairness goals.
- Update training and testing data frequently based on the diversity of people using the technology.

2. Use representative datasets to train and test models:
- Assess fairness in datasets, which includes identifying representation and corresponding limitations, as well as identifying prejudicial or discriminatory correlations between features, labels, and groups. Visualization, clustering, and data annotations can help with this assessment.
- Public training datasets will often need to be augmented to better reflect real-world frequencies of people, events, and attributes that your system will be making predictions about.
- Understand the various perspectives, experiences, and goals of the people annotating the data. What does success look like for different workers, and what are the trade-offs between time spent on task and enjoyment of the task?
- When working with annotation teams, partner closely with them to design clear tasks, incentives, and feedback mechanisms that ensure sustainable, diverse, and accurate annotations. Account for human variability, including accessibility, muscle memory, and biases in annotation, e.g., by using a standard set of questions with known answers.

3. Check the system for unfair biases.
- For example, organize a pool of trusted, diverse testers who can adversarially test the system, and incorporate a variety of adversarial inputs into unit tests. This can help to identify who may experience unexpected adverse impacts. Even a low error rate can allow for an occasional problematic result. Targeted adversarial testing can help find problems that are masked by aggregate metrics.
- While designing metrics to train and evaluate a system, also include metrics to examine performance across different subgroups. For example, false positive rate and false negative rate per subgroup can help to understand which groups experience disproportionately worse or better performance.
- In addition to sliced statistical metrics, create a test set that stress-tests the system on difficult cases. This enables quick evaluation of how well a system is doing on examples that can be particularly undesirable or problematic each time it is updated. Test sets should be updated as the system evolves, with added or removed features and user feedback.
- Consider the effects of biases created by decisions made by the system previously, and the feedback loops this may create.

4. Analyze performance.
   - Take the different metrics defined into account. For example, a system's false positive rate may vary across different subgroups in your data, and improvements in one metric may adversely affect another.
   - Evaluate user experience in real-world scenarios across a broad spectrum of users, use cases, and contexts of use (e.g., TensorFlow Model Analysis[5]). Test and iterate internally first, followed by continued testing after launch.
   - Even if everything in the overall system design is carefully crafted to address fairness issues, ML-based models rarely operate with 100% perfection when applied to real, live data. When an issue occurs in a live product, consider whether it aligns with any existing societal disadvantages, and how it will be impacted by both short- and long-term solutions.

Explainability

Explainability is essential to being able to question, understand, and trust AI systems. These issues apply to humans as well as AI systems—after all, it's not always easy for a person to provide a satisfactory explanation of their own decisions. For example, it can be difficult for an oncologist to quantify all the reasons why they think a patient's cancer may have recurred—they may just say they have an intuition, leading them to order follow-up tests for more definitive results. In contrast, an AI system can list a variety of information that went into its prediction: biomarker levels and corresponding scans from 100 different patients over the past 10 years, but have a hard time communicating how it combined all that data to estimate an 80% chance of cancer and recommendation to get a PET scan. Understanding complex AI models, such as deep neural networks, can be challenging even for machine learning experts.

Understanding and testing AI systems also offers new challenges compared to traditional software. Traditional software is essentially a series of if-then rules, and interpreting and debugging performance largely consists of chasing a problem down a garden of forking paths. While that can be difficult, a human can generally track the path taken through the code, and understand a given result.

With AI systems, the "code path" may include millions of parameters and mathematical operations, and it is much harder to pinpoint one specific bug that leads to a faulty decision. While this poses new challenges, the collective effort of the tech community to formulate guidelines, best practices, and tools is steadily improving our ability to understand, control, and debug AI systems.

Recommended Practices for Explainability

---

[5] https://medium.com/tensorflow/introducing-tensorflow-model-analysis-scaleable-sliced-and-full-pass-metrics-5cde7baf0b7b

1. Plan out options to pursue explainability, including before, during, and after the design and training of a model:

- What degree of explainability does a system need? This will vary across applications and domains (e.g., medical devices in healthcare, shopper recommendations in retail).
- Is it possible to analyze the training/testing data? Anomalous behavior can often be explained by quality issues or gaps in the data. However, when working with private or sensitive data, it may not be possible to fully investigate the input data.
- Is it possible to change the training/testing data, to gather more training data for certain subsets or test data for categories of interest?
- Is it possible to design a new model or is the effort constrained to an already-trained model?
- Are there ways in which the information made available for explainability will open up vectors for abuse?

2. Treat explainability as a core part of the user experience:

- Iterate with users in the development cycle to test and refine assumptions about user needs and goals.
- Design the user experience so that users build useful mental models of the AI system. If not given clear and compelling information, users may make up their own theories about how an AI system works, which can negatively affect how they try to use the system.
- Where possible, make it easy for users to do their own sensitivity analysis: empower them to test how different inputs affect the model output.
- Draw from relevant UX resources, including designing for human needs,[6] user control,[7] teaching an AI,[8] habituation,[9] fairness,[10] representation.[11]

3. Design the model to be explainable:

- Use the smallest set of inputs necessary for your performance goals to make it clearer what factors are affecting the model.
- Use the simplest model that meets your performance goals.
- Learn causal relationships instead of correlations when possible (e.g., using height instead of age to predict if it is safe for a child to ride a roller coaster).
- Craft the training objective to match your true goal (e.g., train for the acceptable probability of false alarms, not accuracy).

---

[6] https://design.google/library/intro-to-hcml/
[7] https://design.google/library/ux-ai/
[8] https://design.google/library/designing-and-learning-teachable-machine/
[9] https://design.google/library/predictably-smart/
[10] https://design.google/library/fair-not-default/
[11] https://ai.googleblog.com/2017/05/neural-network-generated-illustrations.html?m=1

- Constrain your model to produce input-output relationships that reflect domain expert knowledge (e.g., a coffee shop should be more likely to be recommended if it's closer to the user, if everything else about it is the same).
- Analyze the model's sensitivity to different inputs, for different subsets of examples.

4. Choose metrics to reflect the end-goal and the end-task:
    - Metrics should address the particular benefits and risks of the application in question. For example, a fire alarm system would need to have high recall, even if that means the occasional false alarm.

5. Communicate explanations to model users:
    - Provide explanations that are understandable and appropriate for the user (e.g., technical details may be appropriate for industry practitioners and academia, while general users may find UI prompts, user-friendly summary descriptions or visualizations more useful).
    - Identify if and where explanations may not be appropriate (e.g., where explanations could result in more confusion for general users, nefarious actors could take advantage of the explanation for system or user abuse, or explanations may reveal sensitive or proprietary information).
    - Consider alternatives if explanations are requested by a certain user base but cannot or should not be provided, or if it is not possible to provide a clear, sound explanation. In such cases, accountability may be achievable through other mechanisms, such as auditing or allow users to contest decisions or to provide feedback to influence future decisions or experiences.
    - Prioritize explanations that suggest clear actions that can be taken to correct inaccurate predictions going forward.
    - Ensure explanations do not conflate causation and correlation.
    - Recognize human psychology and limitations (e.g., confirmation bias, cognitive fatigue) when crafting explanations.
    - When using visualization to provide explanations, use best practices from HCI.
    - Understand that any aggregated summary may lose information and hide details (e.g., partial dependency plots).
    - Recognize that the ability to understand the parts of the ML system (especially inputs) and how all the parts work together ("completeness") helps users to build clearer mental models of the system. These mental models match actual system performance more closely, providing for a more trustworthy experience and more accurate expectations for future learning.
    - Be mindful of the limitations of explanations (e.g., local explanations may not generalize broadly, and may provide conflicting explanations of two visually-similar examples).

6. Test repeatedly and follow software engineering best test practices:
    - Conduct rigorous unit tests to test each component of the system in isolation.

- Proactively detect input drift by testing the statistics of the inputs to the AI system to make sure they are not changing in unexpected ways.
- Use a gold standard dataset to test the system and ensure that it continues to behave as expected. Update this test set regularly in line with changing users and use cases, and to reduce the likelihood of training on the test set.
- Conduct iterative user testing to incorporate a diverse set of users' needs in the development cycles.
- Apply the quality engineering principle of poka-yoke: build quality checks into a system so that unintended failures either cannot happen or trigger an immediate response (e.g., if an important feature is unexpectedly missing, the AI system will not output a prediction).
- Conduct integration tests: understand how the AI system interacts with other systems and what, if any, feedback loops are created (e.g., recommending a news story because it's popular can make that news story more popular, causing it to be recommended more).

Privacy

AI models learn from training data and make predictions on input data, and at times the training data, input data, or both can be sensitive in nature. Although there may be enormous benefits to building a model that operates on sensitive data (e.g., a cancer detector trained on a dataset of biopsy images and deployed on individual patient scans), it is essential to consider the potential privacy implications in using sensitive data. What safeguards need to be put in place to ensure the privacy of individuals if an ML model is intended to remember or reveal aspects of the data it has been exposed to? What steps are needed to ensure users have adequate transparency and control of their data?

Fortunately, the possibility that ML models reveal underlying data can be minimized by appropriately applying various techniques in a precise, principled fashion. Google is constantly developing such techniques to protect privacy in AI systems. This is an ongoing area of research in the ML community with significant room for growth. Below we share the lessons we have learned so far.

Recommended Practices for Privacy

Just as there is no single "correct" model for all AI tasks, there is no single correct approach to AI privacy protection across all scenarios. In practice, researchers and developers must iterate to find an approach that appropriately balances privacy and utility for the task at hand; for this process to succeed, a clear definition of privacy is needed, which can be both intuitive and formally precise.[12]

[12] https://arxiv.org/abs/1802.08908

1. Collect and handle data responsibly:
    - Identify whether an AI model can be trained without the use of sensitive data, e.g., by utilizing non-sensitive data collection or an existing public data source.
    - If it is essential to process sensitive training data, strive to minimize the use of such data. Handle any sensitive data with care: e.g., comply with relevant laws and standards, provide users with clear notice and give them any necessary controls over data use where applicable, and consider best practices such as encryption in transit and rest.
    - Anonymize and aggregate incoming data using best practice data-scrubbing pipelines: e.g., consider removing personally identifiable information (PII) and outlier or metadata values that might allow de-anonymization (including implicit metadata such as arrival order, removable by random shuffling, as in Prochlo[13]).

2. Leverage on-device processing where appropriate:
    - If the goal is to learn statistics of individual interactions (e.g., how often certain UI elements are used), consider collecting only statistics that have been computed locally, on-device, rather than raw interaction data, which can include sensitive information.
    - Consider whether techniques like federated learning,[14] where a fleet of devices coordinates to train a shared global model from locally-stored training data, can improve privacy.
    - When feasible, apply aggregation, randomization, and scrubbing operations on-device (e.g., Secure aggregation, RAPPOR, and Prochlo's encode step). Note that these operations may only provide pragmatic, best-effort privacy unless the techniques employed are accompanied by proofs.

3. Appropriately safeguard the privacy of AI models.
    - If AI models may expose details about their training data via both their internal parameters as well as their externally-visible behavior, consider the privacy impact of how the models were constructed and may be accessed.
    - Estimate whether a model is unintentionally memorizing or exposing sensitive data using test based on "exposure" measurements[15] or membership inference assessment.[16] These metrics can additionally be used for regression tests during model maintenance.
    - Experiment with parameters for data minimization (e.g., aggregation, outlier thresholds, and randomization factors) to understand tradeoffs and identify optimal settings for a model.
    - Train models using techniques that establish mathematical guarantees for privacy. Note that these analytic guarantees are not guarantees about the complete operational system.

---

[13] https://arxiv.org/abs/1710.00901
[14] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
[15] https://arxiv.org/abs/1802.08232
[16] https://arxiv.org/abs/1610.05820

- Follow best-practice processes established for cryptographic and security-critical software, e.g., the use of principled and provable approaches, peer-reviewed publication of new ideas, open-sourcing of critical software components, and the enlistment of experts for review at all stages of design and development.

Security

Safety and security entails ensuring AI systems behave as intended, regardless of how attackers try to interfere. It is essential to consider and address the security of an AI system before it is widely relied upon in safety-critical applications. There are many challenges unique to the security of AI systems. For example, it is hard to predict all scenarios ahead of time, especially when AI is applied to problems that are difficult for humans to solve. It is also hard to build systems that provide both the necessary restrictions for security as well as the necessary flexibility to generate creative solutions or adapt to unusual inputs. As AI technology develops, attackers will surely find new means of attack, and new solutions will need to be developed in tandem.

Recommended Practices for Security

Security research in AI spans a wide range of threats, including training data poisoning, recovery of sensitive training data, model theft and adversarial examples. Google invests in research related to all of these areas, and some of this work is related to practices in AI and privacy. One key focus area of security research at Google has been adversarial learning—the use of one neural network to generate adversarial examples that can fool a system, coupled with a second network to try to detect the fraud.

Currently, the best defenses against adversarial examples are not yet reliable enough for use in a production environment. It is an ongoing,[17] extremely[18] active[19] research area. Because there is not yet an effective defense, developers should think about whether their system is likely to come under attack, consider the likely consequences of a successful attack and in most cases should simply not build systems where such attacks are likely to have significant negative impact.

1. Identify potential threats to the system:
   - Consider whether anyone would have an incentive to make the system misbehave. For example, if a developer builds an app that helps a user organize their own photos, it would be easy for users to modify photos to be incorrectly organized, but users would have limited incentive to do so.
   - Identify what unintended consequences would result from the system making a mistake, and assess the likelihood and severity of these consequences.

[17] https://www.youtube.com/watch?v=Zd9kYgUjgSU
[18] https://arxiv.org/abs/1802.00420
[19] https://arxiv.org/abs/1801.09344

- Build a rigorous threat model to understand all possible attack vectors. For example, a system that would allow an attacker to change the input to the ML model may be much more vulnerable than a system that processes metadata collected by the server, like timestamps of actions the user took, since it is much harder for a user to intentionally modify input features collected without their direct participation.

2. Develop an approach to combat threats:
   - Test the performance of systems in the adversarial setting. In some cases this can be done using tools such as CleverHans.[20]
   - Create an internal red team to carry out the testing, or host a contest or bounty program encouraging third parties to adversarially test your system.

3. Keep learning to stay ahead of the curve:
   - Stay up to date on the latest research advances. Research into adversarial machine learning continues to offer improved performance[21] for defenses and some defense techniques are beginning to offer provable guarantees.[22]
   - Beyond interfering with input, it is possible that there may be other vulnerabilities in the AI supply chain.[23] While to our knowledge such an attack has not yet occurred, it is important to consider the possibility and be prepared.

II. Recommendations for the Federal Government

Google believes the Federal Government has an important role to play in the development of responsible practices and standards for AI. As described above, this is an area of active investment and research across industry and academia, but government can help reinforce this work and guide both domestic and international communities toward shared definitions and best practices. Below, we list specific recommendations for NIST and the Federal Government to consider.

Support continued research into responsible practices and standards

The development and use of AI is still in a very nascent stage, and while there is already a significant amount of research happening within academia and industry to uncover best practices and standards, it remains a quickly evolving space. NIST should continue to engage this community and support a broad, multi-stakeholder process to ensure standards development reflect the best and most solid results of these research efforts.

Invest in research and foster public-private partnerships

[20] https://github.com/tensorflow/cleverhans
[21] https://arxiv.org/abs/1803.06373
[22] https://arxiv.org/abs/1711.00851
[23] https://arxiv.org/abs/1708.06733

Government can complement and enhance the research being done in academia and industry in several important ways. We encourage the government, through initiatives like the American AI Initiative, to complement and enhance this work, including by:

- Developing model testing data sets to evaluate different AI models designed to solve a specific problem;
- Investing in efforts to compile robust, clean, and open datasets that can enable broader testing and training; and
- Exploring the development of model cards or data cards that establish standardized ways for communicating the core features of an AI model, information around the origins of the model, and its intended uses.

Additionally, as AI is increasingly adopted by organizations across diverse sectors, government should convene broad-based fora that solicit input from not just AI technology developers and providers but also users and stakeholders from across society.

Coordinate with sector-focused authorities

NIST should consider the ways in which the development of new standards for AI map to existing standards, regulations, and legal requirements in specific domains of application and the work that authorities in specific sectors are doing to incorporate AI-related considerations into their oversight. Examples of this range from the Food and Drug Administration's (FDA) discussion paper on Artificial Intelligence and Machine Learning in Software as a Medical Device[24] to U.S. Department of Treasury Financial Crimes Enforcement Network (FinCEN) and Federal Banking Agencies' joint statement[25] on innovative approaches to anti-money laundering (AML) compliance.

Drive consensus toward common definitions and basic terms

While there is growing consensus around broad areas like fairness, explainability, security, and privacy as important areas to develop best practices and standards for AI systems, awareness of the precise definitions of these terms -- and which definitions are important for which contexts -- remains relatively limited. International standards bodies, such as ISO, have established workstreams to develop consistent concepts and terminology. We encourage NIST to play a harmonizing role and work to achieve consensus on and greater awareness of precise definitions in these areas.

---

24 https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

25 https://www.fincen.gov/news/news-releases/treasurys-fincen-and-federal-banking-agencies-issue-joint-statement-encouraging

Actively engage international standards bodies

The Federal Government and applicable US standards bodies, such as NIST, should continue to contribute to and collaborate with international standards bodies, including ISO, IEEE, and others. In addition to resourcing a robust presence at these fora, the government should advocate for producing standards that are internationally applicable, to minimize the risk of fragmentation across multiple country-specific standards covering the same topics. This promotes economic growth and innovation, facilitates integration and interoperability, and improves the overall efficiency and quality of the technology.

* * *

Thank you for the opportunity to provide comments in response to the RFI regarding Artificial Intelligence Standards. We look forward to continuing to work with NIST on these matters.