Intel response to NIST RFI: Developing a Federal AI Standards Engagement Plan

June 10, 2019

Table of Contents

Introduction 2

Part I: Inputs on achieving U.S. AI policy objectives and federal engagement in AI standards 2

(1) Importance of international standards for AI growth and innovation 2

(2) Prioritizing Federal Government Engagement in AI Standardization 4

Part II: Inputs on AI environment and development and research areas related to trustworthiness of AI 5

Key topics areas in AI and related space 5

(1) Foundational technologies and AI standardization 5

(2) Views on technical elements of trustworthiness 7

(3) Understanding the attacks on AI environments 8

(4) Mitigations for threats in AI systems hardware 9

(5) Privacy aspects of AI workloads 11

(6) Data-related guidelines and best practices 13

(7) Societal issues and standardization areas 14

(8) Use Cases 16

Summary of Recommendations 17

Introduction

Intel Corporation appreciates the opportunity to provide inputs to the NIST RFI on development of a plan for Federal engagement in technical standards and related tools in support of reliable, robust, and trustworthy systems that use AI technologies.

Intel is a world leader in computing and technology innovation. The company designs and builds essential technologies that serve as the foundation for consumer products, commercial systems and infrastructure equipment. Intel also invests in the development and adoption of global standards which have enabled advancements and interoperability of products and systems worldwide.

Intel supports NIST's direction for the plan to follow U.S. policies which emphasize voluntary, private sector-led consensus standardization (OMB Circular A-119 "Federal Participation in the Development and Use of Voluntary Consensus Standards and in Conformity Assessment Activities") **and promote innovation and competition. We further recommend that the plan emphasize the development and adoption of international standards over unique national standards** as the best approach to achieve U.S. AI policy objectives and to model U.S. commitments to the WTO Technical Barriers to Trade agreement.

While AI covers a wide range of technology domains where many existing technical standards can be leveraged, advancements in AI technologies and applications are still in early stages of development and research. It is therefore important to have ongoing dialogue among U.S. stakeholders on the appropriateness and timeliness for developing AI-specific technical standards and tools in support of trustworthy advanced systems.

For this input, we also offer a high-level view of AI environment and descriptions of specific development, research and standardization areas related to AI trustworthiness.

A summary of recommendations and considerations for the development of the plan is included at the end.

Intel welcomes the opportunity to further discuss these inputs as NIST develops the draft plan.

Part I: Inputs on achieving U.S. AI policy objectives and federal engagement in AI standards

(1) Importance of international standards for AI growth and innovation

International standards will continue to support U.S. innovation and technology leadership in contributing to the global evolution of AI. They enable global market access for industry, interoperability among products and services, global supply chains and enhanced consumer welfare by increasing economies of scale and competition. International standards are

especially important for addressing areas that benefit from consistent or harmonized global approach such as areas related to technical interoperability, reliability, safety and trustworthiness. Therefore, **we recommend that the development and adoption of international standards be the focus of the plan to meet the needs of U.S stakeholders in the adoption of AI to existing industries and creation of new AI-related industries**.

The U.S. has long embraced a decentralized, voluntary, and market-driven standards system for developing international standards. The system is diverse, including a great variety of organizations that are open to global participants and consensus-based, ranging from established formal standards development organizations to consortia that focus on specific technical areas with faster specification development time-frames. **We recommend that the plan recognize this diversity as being essential to support the different cross-sector and application requirements for developing appropriate, timely, AI technical standard**s.

One of the ways to encourage U.S. technical leadership in AI, including global standards setting, is to support healthy competition among industry stakeholders and the market-driven standards system. This enables a diversity of innovation and technical approaches (there is not one common U.S. view for technical approaches) which can compete to meet different market needs. The system's diversity and innovation have led to development of standards that have substantial global impact and broad market adoption. While other countries' standards systems may fundamentally differ (e.g., relying on top-down approaches), many of these countries have increasingly recognized the benefits and adopted aspects of the U.S. voluntary, market-driven technical standards. U.S. effectiveness and leadership in AI-related international standardization is best measured by the extent to which the resulting standards support U.S. interests and needs, and enable the private sector to develop trustworthy AI solutions accepted by markets and governments worldwide. **To support U.S. stakeholders' ability to contribute and discuss technical proposals to meet U.S needs, we recommend that the plan reinforce the importance of ensuring and promoting international standards development processes that are consensus-based and open to all interested participants**.

Since AI systems cover a variety cross-domains and technical areas, many existing international standards (including a majority of ICT technical standards) can be adopted or extended. AI specific international standardization is in early stages, with notable activities under ISO/IEC JTC 1/ SC 42 AI (including foundational work for terminology and trustworthiness) in cooperation with other international committees such as for governance, security, privacy, data-related areas, safety and sector-specific aspects, and under IEEE-SA (P7000 series addressing ethics and other considerations for system design, and an Ethics Certification Program for Autonomous and Intelligent Systems). Technical areas where advancements in AI technologies and applications are in early stages of development and research, or evolving rapidly, may be premature or inappropriate for standardization (including development of conformance testing that need to be based on standards).

When there may be limited needs for the development of national AI standards (such as for public infrastructure), we recommend that the plan provides guidance that the development of these standards should be based on international standards to the extent possible.

In addition to standards, other tools and solutions are also important for the development of robust, trustworthy AI technologies and systems and for accelerating adoption. These include community-based Open Source Software (OSS) Projects (such as OSS tools for machine learning), industry platform specific OSS, consensus-based best practices and guidelines for specific aspects of trustworthiness (e.g. Partnership on AI), machine learning benchmark suites (e.g. MLPerf), public-private-partnership testbeds and reference designs, and proprietary solutions. **We recommend that the plan recognize the importance of ongoing dialogue among U.S stakeholders on the evolving requirements for the development and adoption of AI-specific technical standards, tools and solutions.**

(2) Prioritizing Federal Government Engagement in AI Standardization

Intel recommends that the plan focus U.S. government engagement in industry-led consensus standards bodies, existing or new, that develop international AI standards that are applicable across-sectors or are application specific. **We encourage the plan to recognize the importance for government experts, especially technical standards experts from NIST, to participate regularly and consistently in international standards development and to allocate sufficient resources to support their effective participation**. In this way, government experts can partner with private sector and other stakeholders to efficiently develop international standards which meet U.S market and government needs.

We recognize NIST's important role as a coordinator for U.S. government engagement in bringing agency requirements and proposals to the international standards discussions. We also recognize NIST's expertise in convening public-private sector initiatives and guiding the development of frameworks for increasing trustworthiness of systems and processes. As NIST considers proposals for convening initiatives to address trust in AI, it is important not to duplicate the ongoing international standardization work. Therefore**, we recommend that federal engagement and coordination with private sector be prioritized on direct participation in standards bodies for subject areas which are covered by their standardization work**. **Initiatives convened by NIST in pre-standardization areas can be designed as research and sense-making work with the private-sector. Results can form future proposals to international standards when appropriate.**

Part II: Inputs on AI environment and development and research areas related to trustworthiness of AI

Scope

This part provides brief descriptions of important areas to obtain a consistent big picture for Artificial Intelligence (AI), with the emphasis on the topics that we consider definitional for AI trustworthiness. For each topic, we describe example areas of development and research and relevant standardization to enhance international harmonization in the field, and accelerate the adoption of AI technologies for societal benefits.

References to existing standardization work are focused on known technical areas of standardization under global SDOs and leading industry consortia, and are not elaborated in detail.

For research and development, we describe the areas in AI that yield themselves to standardization or can benefit from pre-standardization activities to determine whether standardization could be timely and beneficial.

The portfolio of topics is not exhaustive, and they are described at a high level for this environmental scan and focus on broadly understood areas related to trustworthiness.

Key topics areas in AI and related space

(1) Foundational technologies and AI standardization

Summary: Current success of machine-learning based AI is predicated on significant advancements in foundational technologies, including computing power, storage, networking bandwidth, battery power, software and hardware architectures, and many others. Deployment of these foundational technologies is enabled a wide range of international technical standards which can be adopted and adapted for AI systems and environments.

AI technologies focusing on machine learning (ML) are in a phase of rapid development and wide adoption. While the concept of AI has existed for over sixty years, real-world applications have accelerated in the last decade with the emergence of more advanced algorithms, increases in networked computing power and proliferation of technologies to capture and store massive amounts of data. At present time, the field of AI is developing computing systems, often probabilistic in nature, that are capable of distilling, storing and processing information in a way that mimics human reasoning.

The current stage of AI would not have been possible without breakthroughs in computing power, networks, storage and battery technologies, energy efficiencies, and many other parts of the computing and networking infrastructure. ML-based AI is data dependent, requiring

massive data sets, frequently with millions of data points, to define viable models and algorithms. These advances in foundational infrastructure-related technologies are the main catalysts for the explosive growth of AI. **Since AI systems are built upon foundational technologies - general purpose technical approaches, components, and protocols – the large body of international technical standards for interoperability and trustworthiness aspects can be leveraged. Existing technical standards and specifications can be adopted, adapted, or extended to support evolving requirements for AI, including emerging standards that address governance and other aspects for trustworthy AI systems.** Examples of foundational international standards areas relevant to modern AI include cryptography, cryptographic protocols, telecommunications, data storage and firmware architectures, governance, product and system development processes, risk assessment. These standards are developed under a range of organizations including ISO and IEC, JTC 1, IEEE, W3C, ETSI and many more.

With AI workloads running on variety of general purpose as well as specialized hardware[1] including CPU, GPUs, ASICs, FPGAs etc., opportunities of standardization could be explored around common APIs to deal with heterogeneous computing[2] environments and diverse algorithms and frameworks to support the large number of increasing use cases with varying needs of accuracy.

Foundational and Sectoral Influences

An essential characteristic of ML algorithms, one of the foundations of modern AI, is the ability to distill information from input data through a learning process, generalize the information learned, and support a wide range of diverse cognition functions. AI algorithms are driven by and hosted on a variety of complex platforms ranging from cloud computing systems to edge devices, enabled by scalable computing systems, vast amounts of data and network connectivity. The boundaries of AI technologies are somewhat fuzzy, and while AI discussions today frequently focus on ML-based systems, the traditional areas of ontology and reasoning, as well as, expert and decision support systems continue to be relevant. **Therefore, since AI is a broad field, it is useful to develop common definitions and terminology, and agreed upon scope of areas for standardization. Discussions on AI specific foundational standards have started under international standards bodies, and further work is needed to define appropriate fields of study and applications.**

AI can greatly increase customer service, productivity, product quality, or response time because it expands our potential for identifying and understanding patterns in vast amounts of data and then predicting their implications and offering solutions and suggestions. But because AI can automate some aspects of decision making, concerns arise with regard to trustworthiness, assurance, security and privacy especially for applications that require sensing and reaction in nearly real-time.

---

[1] https://en.wikipedia.org/wiki/AI_accelerator

[2] https://www.intel.ai/heterogenous-computing-ai-hardware-designed-for-specific-tasks/#gs.hbw19x

The wide range of current and potential applications of AI makes it necessary to focus on sectoral requirements. In this respect, AI as a field is not different from other technology development areas. However, the wide applicability of AI techniques makes the task of incorporating, adjusting, and harmonizing the existing foundational and sectoral standards and specifications more daunting than usual. **Thus, standards, specifications, and best practices, developed in specific areas, from electronic banking to industrial control systems, will need to be evaluated for appropriate incorporation to support AI capabilities. These sectoral standards will also need to be considered in the development of AI specific standards, specifications, frameworks, and guidelines that address cross-sectoral common areas including foundational concepts and aspects for trustworthiness of AI.**

AI technology areas which are in early stages of development or evolving rapidly need to reach a certain level of technology maturity before considering appropriate standardization. However, pathfinding activities can be beneficial in determining potential approaches for future standards and contribute to their efficient and timely development.

Beyond Technology

In most complex multi-disciplinary fields, standards and specifications go beyond purely technical areas. Standards and specifications associated with AI need to cover all aspects of a complex lifecycle, including the design, development, deployment, and use of the technologies as well as what happens with the data created, used, shared and destroyed during the system lifecycle**. Standards and specifications are needed to help organizations develop processes for optimal governance and risk management associated with the use of AI systems and data, on which they rely. AI brings new risks and responsibilities that must be addressed by organizations, thus pointing to the need for non-technical standards, guidelines, and best practices. However, in a broad field like AI, one size does not fit all, and contextualization is of paramount importance for developing the standards and guidelines.**

(2) Views on technical elements of trustworthiness

Summary: **Modern complex computing environments, such as AI, require integrated views on privacy, security, and safety (where applicable). Requirements for privacy, security and safety may include conflicting components, and therefore need to be considered in an integrated fashion to ensure alignment.**

The concept of "trustworthiness" has become popular in both research and standardization. The complexity of the ecosystem means that considering security or privacy as separate sets of requirements, may not be viable because some requirements may be in conflict with each other, and an integrated view is necessary. Several approaches have been taken to address trustworthiness. For instance, the NIST CPS Framework creates an integrated model of trustworthiness using the following definition:

> **"Trustworthiness**: demonstrable likelihood that the system performs according to designed behavior under a typical set of conditions as evidenced by its characteristics, such as safety, security, privacy, reliability and resilience" (from NIST CPS Framework v1.0).

In international standardization, for example, several committees under ISO/IEC JTC 1, are approaching trustworthiness as an additional risk-based framework.

In both cases, the more complex models or frameworks are based on the realization that it is impossible to build trustworthy systems if fundamental aspects such as privacy and security are considered in isolation. Other approaches to trustworthiness may focus on specific aspects such as transparency or accountability.

With regards to specific applications in AI, trustworthiness is frequently a concept in the context of the autonomous and intelligence capabilities that AI promises to bring into systems.

Many elements are already considered in standardization with regard to trustworthiness in foundational and sectoral technology areas that are relevant to AI. They include, but are not limited to, integrated models of trustworthiness, novel risk assessment models, related ontologies and reasoning algorithms, and assurance techniques. In addition to these elements, AI specific technical standardization can also provide a means to understand technical models in addressing explainability requirements and algorithmic bias (including mitigations).

Non-technical elements of trustworthiness, such as non-technical bias or fairness, are discussed in a separate section.

(3) Understanding the attacks on AI environments

Summary: Typical attacks on AI environments can inform the views on standardization. While this is an area of research rather than standardization, understanding the typology of attacks is a necessary step towards building viable security and privacy features in AI systems.

This section provides examples of cybersecurity attacks relevant to the AI environments, while the following section outlines potential areas for collecting best practices to mitigate these threats. The objective is to provide a flavor of this subfield.

Attack surface of AI workloads

AI workloads are software programs which are subject to similar attacks as other technical applications. Typical software and hardware attacks on ML workloads are *digital attacks* affecting protection of the data and integrity of the data and computation. Other forms of attacks can lead to denial of service (loss of availability), cause information leakage, or lead to invalid computation. We must also consider the new forms of attacks using atypical methods, such as analog or physical attacks called *adversarial ML attacks* (described in more detail below).

Research on the distinctive nature of AI attack forms and approaches for understanding the foundations for attack resiliency are necessary precursors to forming a direction for

international standardization or the adaptation of existing standards and guidelines to improve resilience of AI systems.

Adversarial attacks on AI systems

The well-defined models for attack and defense in classical computing security do not always transfer to adversarial ML attacks. Minor alterations to the input data can manipulate or poison the ML outcomes. Other attacks attempt to steal the ML model by exfiltrating or reverse engineering the model. The goal of these attacks is often to replicate a service based on such models.

A growing number of studies have shown that neural networks are vulnerable to the presence of subliminal signals, which are capable of causing serious harm by influencing the neural network cognitive functions. Such signals pass unnoticed by the human and may be capable of causing neural networks to misclassify their inputs or learn the wrong things from the environment, resulting in potentially harmful outputs.

While the study of adversarial attacks may not lead to direct standardization efforts, making sense of the new types of threats relevant for AI can provide insights into the foundations of trustworthiness relevant to standards development. Sense-making efforts such as collecting cross-cutting issues and evaluating current solutions in a new or complex field such as AI are an important precursors for considering appropriate standardization.

(4) Mitigations for threats in AI systems hardware

Summary: As AI is built on general purpose foundational technologies, it can also adopt and adapt existing mechanisms to protect security and privacy in AI environments and take advantage of existing international standards in this area.

This section provides examples of the hardware mitigations to common threats to AI. This is not an exhaustive list, and we provide these examples in order to draw attention to relevant hardware capabilities and relevant standards.

Reduction of attack surface via access-control

Access-control mechanisms can be applied via hardware to isolate access to AI assets such as code and data used during the training and inference process. Hardware access-control mechanisms that are non-malleable are preferred over software/firmware mechanisms. As with any access-control scheme, it is critical to ensure privilege separation of the policy owner from the target of the access-control policies. Hardware mechanisms for privilege-levels can also be leveraged by software for enforcement of access-control software mechanisms. **There is a large body of diverse standards for the access-control field which need to be adopted and adapted to AI use cases.** They are developed in a number of organizations including JTC 1, OASIS and industry consortia (e.g. FIDO Alliance, PCI-SIG).

Reduction of memory attack surface via cryptography

Memory encryption is a general technique to support data protection. It is intended to provide confidentiality, integrity and replay-protection of content exposed to external memory buses and memory modules. Memory encryption should be enabled in concert with appropriate access-control modules. Protection should rely on using community-evaluated, standard encryption algorithms with sufficient robustness and key length. If memory is persistent, the encryption strength should be sufficient to address offline attacks as well. **There are relevant technical standards for memory protections mechanisms that need to be adopted and adapted to AI use cases.** Memory encryption relies on international cryptographic standards and related standards for media encryption developed in a number of standards bodies from JTC 1 to IEEE.

Trusted Execution Environments

Trusted Execution Environments (TEEs) are used to protect selected code and data from disclosure or modification. Developers can partition their application into hardware-isolated programs or hardware-protected areas of execution to increase security, even on compromised platforms. Using trusted execution environments, developers can protect ML training programs as well as ML models in use for inference, effectively treating the model as secret data. TEEs can enforce confidentiality and integrity of workload memory (typically using both access-control and cryptography mechanisms) even in the presence of privileged malware at the system software layers.

Trusted execution in different forms has been subject of standardization for two decades. Existing standards need to be adopted and adapted to AI use cases. Standardization work in this area is done in various standards bodies and consortia including the Trusted Computing Group, GlobalPlatform, The Open Group, DMTF and JTC 1.

Specialized hardware

Complex device models may be supported by specialized accelerators to enhance ML workload performance. In many cases, these accelerators or devices may be para-virtualized or emulated, and in some cases certain workloads may benefit from using devices which are directly assigned to them (for higher efficiency). However, these devices should be verified (via attestation) to ensure that the device is capable of upholding the privacy and security requirements of the AI workload. Hardware IO memory management capabilities should be used to securely bind devices to workloads including direct memory access into protected memory. Attack vectors that must be addressed in this domain include device spoofing, runtime memory remapping attacks, and man-in-the-middle attacks.

Technical approaches for device verification (attestation) as specified in existing industry standards and practices need to be extended to support the complex device models required to execute ML workloads.

(5) Privacy aspects of AI workloads

Data processing and analytics occur across the infrastructure - at the edge, on the network and in the data centre. Personal information is not only collected from individuals who provide it for particular uses, but also gathered by sensors in connected devices, and can be derived or created through further automated processing. The collected data represents a combination of the elements of personally identifiable information (PII) and machine-to-machine feeds. Increasingly autonomous and ubiquitous technologies take advantage of large datasets and data from multiple sources to make autonomous determinations in near-real time. In some cases (such as banking, human resources, transportation), these AI-enabled decisions may affect an individual's private life, physical safety, position in society, or interaction with others.

Due to the massive amounts of data containing PII and non-PII elements, pathfinding and sense-making efforts may explore machine readable elements that can be standardized, if needed, such as associated metadata or other elements that could assist in achieving the privacy objectives such as accountability, transparency, and user control. In some AI application fields, technologies and practices to separate PII from machine-to-machine data will be of great importance.

Increased automation should not result in less privacy protection. Privacy protection aims to prevent unauthorised access, modification and loss of personal data. AI techniques with potential to create data such as images, videos, and sounds are moving discussions beyond the pure risk of identification or unauthorized processing of PII. AI is increasing the risks to individuals of potentially creating false information that may manipulate people and impact their perception of reality.

Privacy protection is best addressed by a combination of regulatory and technical means. Existing regulatory frameworks and international technical standards for data and privacy protection need to be considered for appropriate adaption to support evolving AI and machine learning environments.

Technical areas of research and development

Data security is a space that attracts a lot of attention from regulators and researchers, and has created commonly used principles and approaches (for example, those adopted in the EU General Data Protection Regulation). Data security relies on methods to keep user data *confidential*. ML-based AI poses new problems with regards to traditional methods used to protect data from re-identification and creates new challenges for processing data in ways that keep data encrypted or otherwise limit the ability to read the data being processed. ML-based AI is an intrinsically multi-party computation, with multiple stakeholders: training data owners, model owners, inference data owners, ML service providers, and infrastructure providers, to name a few. The nature of digital data thus requires a highly complex web of trust between all of these parties—a web that quickly becomes difficult, but critical, to manage.

For certain use cases, by enabling data interactions between untrusted parties, previously

impossible ML use cases across broad industry segments can be enabled. For example, rival banks could choose to create a fraud model available to both based on their joint private data, thereby reducing their risk without divulging core business secrets. Or a hospital could use cloud-based analytics on patient data while keeping it encrypted at all times.

The new techniques described below are broadly applicable on diverse datasets with multiple ownership. While there are some existing standards which can support privacy while preserving data processing in these contexts, new standards, best practices and guidelines will need to be created. This is an important area of pre-standardization research and development.

Federated Learning: This technique allows multiple owners of private data sets to jointly train a model based on the union of their data without sharing their data with other parties. First, a central server shares an initial version of the model with the data owners. Next, each data owner uses its private data to adjust and improve the model; these adjustments are aggregated at the central server, which then sends the improved model to the data owners for further rounds of adjustments.

Differential Privacy (DP): Ideally, processes for aggregating data should hide PII of its individual contributors. In practice, however, under certain circumstances, it is possible to re-identify individual information in an aggregated data set. DP is a definition of privacy that provides a probabilistic bound on any single contributor's impact on a given statistic—for example, how much the distribution of sample means would change by adding or removing a sample. A typical strategy for reducing the effect of any one entry in a dataset is to add randomness to each user's data when training a model. It can be challenging to achieve the right balance between privacy and utility, since increasing the magnitude of added noise will come at the expense of the model's accuracy, in the limit.

Homomorphic Encryption (HE): This technique allows ML algorithms to operate on data while it is still encrypted, that is, without access to the underlying sensitive data. Using HE, a hospital could lock sensitive medical data; send it for analysis on a remote, untrusted system; receive back encrypted results; then use its key to decrypt the results—all without ever revealing the underlying data. Using HE also provides ongoing protection where a solution like DP would not - even if a machine stores data without authorization, for example, the data remains protected by encryption. Operating on homomorphically encrypted data require significantly more computation compared to the equivalent operations on cleartext, although this computation gap has narrowed rapidly by several orders of magnitude in the last three years with the advancements in computing technologies.

Multi-party computation (MPC): This cryptographic technique enables two or more parties to compute an output that depends on inputs that each party would like to keep secret, in a way that the parties learn the output but nothing about the secret inputs. There are many MPC protocols, but they all tend to be communication intensive, that is, rate-limited not by

computation speed but by communication bandwidth between the parties.

In these areas, there is significant reliance on existing areas of standardization such as cryptography. Work related to anonymization and obfuscation of different kinds have been done in a number of standards bodies including JTC 1, ISO and the Trusted Computing Group. There are also early standardization and pre-standardization efforts in specific emerging areas, for instance homomorphic encryption under JTC 1, and the NIST differential privacy synthetic data challenge[3]

(6) Data-related guidelines and best practices

Data-related areas and topics represent a complex field, since they are subject to regulatory regimes with regards to privacy protection, data sovereignty, localization, and cross-border transfers. The development of international technical standards addressing these data-related areas can be beneficial to harmonize common regulatory approaches which are necessary for the success of global AI applications. Additionally, standardization enables regulations to focus on high-level, longer-term regulatory requirements and rely on voluntary standards to addressing evolving technical requirements for AI data use cases.

In similar areas, highly affected by privacy regulations, international standards have been successful in defining useful mechanisms to support the regulatory objectives in a harmonized manner via technology means where appropriate. Examples of such standards include the Do Not Track standard (under W3C) and anonymous signatures and authentication standards (under ISO/IEC JTC 1 SC 27).

Access to datasets

Access to large and reliable datasets is essential to the development and deployment of robust and trustworthy AI systems. It will be beneficial for a number of reasons, including the mitigation of algorithm and data bias and improving the quality of algorithms. **Standards and guidelines can play an important role in developing approaches for access to AI datasets, but need to be carefully defined based on different use case contexts and consider common privacy regulations and legal obligations associated with data sets**. Data-related standards, such as metadata and format interoperability standards, can facilitate the following activities:

1. Making available public sources of information in structured and accessible databases (open government data).
2. Creation of reliable datasets (with techniques to protect personal information of individuals), which could be used by all AI developers to test automated solutions and benchmark the quality of their algorithms.
3. Fostering incentives for data sharing between the public and private sector and among industry players.
4. Promoting diversity in datasets.

[3] https://www.topcoder.com/community/data-science/Differential-Privacy-Synthetic-Data-Challenge

(7) Societal issues and standardization areas

Societal issues is another area at the intersection of policies, regulations and standards activities which is critical to the adoption and trust of AI technologies. A number of organizations have developed "AI Principles" or "AI Ethics Principles" to address a set of closely interrelated issues expected to meaningfully affect the lives of individuals and communities[4] and also have an impact on how AI systems are developed and how AI related data sets are handled. Falling under the broad heading of "ethics," these social issues (including bias and inclusion, safety, fairness, privacy, security, and the future of work) often garner attention following high-profile events or well-publicized narratives, such as facial recognitions systems failing to detect faces not associated with predominant ethnic groups in a region (bias),[5] autonomous vehicles harming drivers and pedestrians (safety),[6] and automated decision systems denying access to housing, credit, education, or freedom from incarceration (fairness).[7]

In these and other cases, there is a strong regulatory and public demand to understand the source and nature of the problematic events in order to prevent similar future social harms. The nature of the issues are complex, for example, algorithm and data bias is not always related to discrimination. For instance, crowd analysis systems in Asia have problems assigning age correctly to non-Asians because of the differences in parametrization. And fraud detection systems frequently fail to make a distinction between bots and humans using privacy protection techniques.

Potential areas to define technical standards that can minimize societal issues should be explored. For example, reducing algorithm and data bias by using techniques developed to support context discrimination and error control. Where specific standardization direction may not be clear, pre-standardization activities can be helpful, examining the space for potential areas that can benefit from standardization.

The same approach may be recommended for automated decision making, specifically with regards to explainability. There is a perception that algorithmic decision-making occurs within a "black box" that renders meaningful human explanations inscrutable or even technically impossible. For this reason, explainable AI (xAI), the ability to describe why algorithms make particular decisions, has become a cross-cutting ethical issue in its own right, pursued simultaneously by engineers, regulators and other watchdogs, and members of the public. **Pre-standardization examination would also be beneficial, along the lines of the four areas**

---

[4] E.g., IEEE's Ethically Aligned Design, World Economic Forum, AI Now Institute, UN Guiding Principles

[5] When Joy Buolamwini was a computer science graduate student at MIT, for example, she had to wear a white mask to work on her thesis project, because the face recognition algorithm she was using didn't recognize her brown-skinned face as a human face.
https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms?language=en

[6] E.g., https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html

[7] https://medium.com/@AINowInstitute/taking-algorithms-to-court-7b90f82ffcc9

briefly described below.[8] We provide additional considerations in addressing the areas with regards to standardization or purely regulatory approaches.

1. *Claim of confidentiality*: Entities may understand how a particular algorithmic decision was made but withhold explanation in order to protect intellectual property or security interests;
2. *Complexity*: Engineers may understand an algorithm but be unable to formulate a simple, holistic, big-picture summary of algorithmic behavior that would satisfy non-technical audiences;
3. *Non-intuitiveness*: Engineers may be unable to make intuitive sense of a rule that an AI system discovers; and
4. *Lack of justification*: Engineers may understand how an algorithm works and why certain outcomes are being reached, but the explanation itself is not fair or reasonable.

A solution to the first source of difficulty may be found through legal or policy channels. The second may require engineers to work closely with non-technical colleagues to develop human-centric explanations and relevant best practices, guidelines, or specifications. The third requires continued research in AI theoretical and algorithmic space to translate the rules that an AI system discovers into visual and intuitive explanations. Furthermore, it requires research in novel AI algorithms that take into account outputs which are both meaningful and interpretable. The fourth issue, the desire for *justification*, is arguably the most critical today. Understanding the source of algorithmic injustice alerts users of AI systems to their social dangers, curbing a human tendency to regard AI as infallible and to over-apply it, and it also creates space for engineers and regulators to architect technical, legal, and policy guardrails.[9] Research in this area can provide useful inputs into standardization work.

We can also make a distinction between the different roles and responsibilities of AI system stakeholders and related considerations for standardization work.

For creators of AI systems, it will be critical to think about interpretability at the outset, as system requirements are being developed, and iteratively throughout design, implementation, testing, and maintenance phases. Engineers should openly discuss and *document* the purpose of the AI system, the social context in which the system will be deployed, the full complement of intended subjects or users (including their ages, races, and genders), and the relationship between these design considerations and the sources and composition of ML training datasets. Similarly, requirements for bidirectional *traceability* between design requirements and source code will help engineers provide explanations of system behavior. Documentation and traceability will be especially critical when applications will affect financial, medical, or public resource decision-making.

Deployers of AI systems have the responsibility of familiarizing themselves with system assumptions and limitations, ensuring that their applications are consistent with the design of

[8] https://freedom-to-tinker.com/2017/05/31/what-does-it-mean-to-ask-for-an-explainable-algorithm/
[9] E.g., https://arxiv.org/abs/1701.08230

the system, and documenting deviations. For example, a tumor detection system designed to detect prostate cancer in men, may not be optimized for detecting breast cancer in women. An autonomous vehicle system trained in an urban environment with California traffic laws and norms may not perform optimally in a rural area outside of the USA. In these situations, deviation from the environment in which the AI system was trained may affect the accuracy and reliability of algorithmic decision-making, providing insight into sources of errors

Organizations should reinforce the importance of designing and evaluating AI systems with fairness, transparency and accountability as additional metrics to the existing and typical AI system metrics, such as total cost of ownership, accuracy and generalizability etc. Building these additional measurements into AI system development process can potentially encourage new practices and processes to improve trustworthiness of AI.

(8) Use Cases

AI technologies have broadly applicability across segments and industries. While technology is an enabler, successful use cases will help establish the market for AI solutions which will then lead to continued investment bringing about a further advancement of technology completing a circle of exponential growth. The last two periods of growth in AI petered out due to unmet expectations from multiple use cases in the industry leading to what are referred to as the "winters of AI".

A deeper understanding of practical use cases will (i) help define areas where AI specific technical standardization may be beneficial, including addressing aspects of trustworthiness (such as algorithmic bias, data privacy) and societal concerns (ii) prevent "overselling of AI" which has been the cause of disappointment in the past leading to abrupt reductions of investment (iii) bring a broader understanding of where current AI technology successfully solves industrial problems (iv) help evangelize areas where AI is solving human problems, "AI for good[10]", to counter overly negative narratives and (v) help ensure that standardization efforts are "broad enough" to cover cross-sectors and broad application needs.

We want to stress the importance of use cases for the development of timely and useful AI technical standards. Incentives needs to be created to share viable use cases and creation of mechanisms to enable such sharing would be beneficial.

[10] Example initiatives: https://www.intel.ai/ai4socialgood

Summary of Recommendations

Achieving U.S. AI policy objectives and priority for federal engagement in standards

1. **Ensure the consistency of the plan with U.S. policies.** Emphasize voluntary private sector-led consensus standardization, promote innovation and competition, and focus on international standards over unique national standards as best approach to meet U.S stakeholder needs and support U.S. technology leadership.
2. **Recognize the diverse market-driven standards system for timely AI technical standardization.** The diversity of organizations developing international standards is essential to support the different cross-sector and application requirements for developing appropriate, timely, AI technical standards.
3. **Reinforce and promote open, consensus-based international standardization processes.** To support U.S. stakeholders' ability to contribute and discuss technical proposals, it is important to ensure and promote international standards development processes that are consensus-based and open to all interested participants.
4. **Recognize the need for ongoing dialogue among U.S. stakeholders as AI develops and evolves**. Advancements in AI technologies and applications will influence evolving requirements for the adoption of existing standards and appropriate timing for development of AI-specific technical standards, tools and solutions.
5. **Prioritize consistent federal engagement in international standards work.** Allocate sufficient resources to support continuous federal engagement and NIST's role as coordinator and expert. Support NIST's expertise in convening public-private sector initiatives that support U.S. stakeholders' participation in ongoing international standardization work, and pre-standardization work in areas of research or possible future AI technical standards proposals.

Foundational technologies and sectoral influence

6. **Make full use of existing general purpose technical standards.** General purpose standards, specifications, and best practices continue to be important for AI systems and environment. They need to be adopted and adapted for AI.
7. **Make full use of technical standards developed for specific contexts or sectors.** Standards, specifications, and best practices developed in specific areas from electronic banking to industrial control systems can be incorporated and adjusted. Greater attention needs to be paid to early adoption areas for AI with broad public impact (e.g, healthcare or transportation).
8. **Consider needs of non-technical areas.** Standards and best practices in non-technical areas are needed to help organizations develop processes for optimal governance and risk management associated with AI systems and data, on which they rely.

General views on Trustworthiness

9. **Work with international standards bodies to develop a unified, but flexible approach to Trustworthiness.** Many elements are already considered in standardization efforts of trustworthiness for foundational and sectoral technology areas and AI-specific aspects. They include (but are not limited to) integrated models of trustworthiness, risk models of trustworthy systems, related ontologies, reasoning algorithms and assurance techniques. Trustworthiness is also frequently connected to privacy, transparency, accountability and explainability.

Security attacks on AI Surfaces

10. **Study specific attacks to understand trustworthiness requirements for AI.** The distinctive nature of some AI attacks forms a viable pre-standardization area. Approaches for understanding the foundations of attacks and improving resiliency in AI need to be studied by the international standardization community in order to form a direction for international standardization or plan for the adaptation of existing standards and guidelines. Adversarial ML attacks present an attractive and informative field for study.

Mitigation techniques

11. **Use existing mitigation techniques, and don't forget hardware.** AI technologists can adopt and adapt existing mechanisms to protect security and privacy in AI environments and take advantage of existing international standards in this area. Areas such as access-control, Trusted Computing, Memory Protection and Trusted Execution are already highly standardized and should be adapted for AI use cases.

Privacy and data security

12. **Examine the best technology and process based approaches for privacy preserving AI before proceeding to international standardization.** Due to the massive amounts of data containing PII and non-PII elements, pathfinding and sense-making efforts may be needed to discover optimal approaches that could assist in achieving privacy objectives including accountability, transparency, and user control. Technologies such as homomorphic encryption, multi-party secure computation, and federated machine learning as well as approaches such as differential privacy should be studied for standardization potential.

Data sharing and data access

13. **Study the area based on lessons learned from similar complex context driven fields**. **Focus pre-standardization efforts on structural characteristics of datasets.** In similar areas highly affected by privacy regulations, international technical standards have been useful in defining mechanisms to streamline regulatory objectives in a harmonized manner. Potential standards for structural harmonization of datasets (e.g., compatible formats and metadata) can be examined, based on existing standards or adaptions.

14. **Consider initiatives to support the following recommendations to improve data access and sharing:**

    a. Make available public sources of information in structured and accessible databases (open government data).

    b. Create reliable datasets (including personal information of individuals), which could be used by all AI developers, by start-ups and more broadly by industry to test automated solutions and benchmark the quality of their algorithms.

    c. Foster incentives for data sharing between the public and private sector and among industry players.

    d. Promote diversity of content in datasets.

Societal Issues

15. **Study societal issues such as explainability, transparency and ethics to understand the requirements for technical standards.** The AI community needs to undertake a formal study of societal issues through research and pre-standardization pathfinding before charting the course for international standardization in areas that need interoperability, harmonization, multiple stakeholders, and will benefit from voluntary technical instruments.

Use Cases

16. **Create a repository of representative, diverse AI use cases to support the development of timely and useful standards.** Access to use cases will be beneficial to for the entire AI community including technologists, designers, users, policy makers and regulators.