June 6, 2019

To:
Elham Tabassi,
Acting Chief of Staff, Information Technology Laboratory,
National Institute of Standards and Technology
100 Bureau Drive, Stop 200
Gaithersburg, MD 20899

RE: RFI: Developing a Federal AI Standards Engagement Plan

Dear Ms. Tabassi,

On behalf of the Center for the Governance of AI, the Future of Life Institute, the Center for Long-Term Cybersecurity, and certain researchers at the Leverhulme Centre for the Future of Intelligence, we are pleased to submit comments in response to NIST's request for information on the important topic of artificial intelligence (AI) standards. Our organizations have collaborated on this response in order to leverage diverse expertise and to highlight the consensus of our remarks.

- The Center for the Governance of AI, housed at the Future of Humanity Institute, University of Oxford, pursues interdisciplinary research and policy engagement to reduce global risks in the development of AI.
- The Future of Life Institute, based in Cambridge, Massachusetts, is a non-profit organization whose mission is to catalyze and support research and initiatives for safeguarding life and developing optimistic visions of the future, including positive ways for humanity to steer its own course considering new technologies and challenges.
- The Center for Long-Term Cybersecurity is a research and collaboration hub at the University of California, Berkeley helping people and organizations to anticipate and address tomorrow's information security challenges, in order to amplify and extend the upside of the digital revolution.
- The Leverhulme Centre for the Future of Intelligence is a research group across the Universities of Cambridge, Oxford, Imperial, and Berkeley, that aims to build an interdisciplinary community of researchers working together to ensure that we make the best of the opportunities of artificial intelligence as it develops over the coming decades. Contributing scholars to this effort are Seán Ó hÉigeartaigh and Hadyn Belfield.

Our submission is organized into three sections, to correspond to the three high-level categories that NIST is seeking to better understand: AI Technical Standards Development: Status and Plans; Defining and Achieving U.S. AI Technical Standards Leadership; and Prioritizing Federal Government Engagement in AI Standardization. Relevant question numbers are referenced in each of the three sections. In addition, our reference list includes resources that NIST may find useful in drafting the Federal AI Standards Engagement Plan.

This submission defines “standards” in accordance with OMB Circular No. A-119: “Common and repeated use of rules, conditions, guidelines or characteristics for products or related processes and production methods, and related management systems practices.” Thus, this includes both “performance-based or design-specific technical specifications and related management systems practices” (P.L. 104-113 § 12(5)) and industry guidelines of best practice that are being developed by, e.g., the Partnership on AI.

We agree that AI technical standards will play a crucial role in the research, development, deployment and use of trustworthy AI technologies. We further agree that NIST can beneficially lead engagement to support the development of AI technical standards to best support safe, reliable, and robust AI technologies. We remain at your disposal to provide any further information or clarification.

1. AI Technical Standards Development: Status and Plans

Summary:

- There are several efforts underway to develop industry and cross-sector international standards. The U.S. federal government, led by NIST, should actively engage in those processes.
- Underinvestment in AI safety standards and research due to uncertain market demand is a key challenge in determining the need for AI standards.
- NIST should evaluate five standards ideas for possible inclusion in its guidance for the Federal AI Standards Engagement plan: (1) adversarially robust training certificates, (2) standardized explainability levels, (3) explainability with domain-specific language, (4) explainability with domain-specific language, and (5) machine-readable declarations to advance trustworthiness.

1A. Existing standards and development landscape

Question 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;

There are relevant private consensus-based standards under development today at ISO/IEC JTC 1 SC 42, other JTC 1 subcommittees, the IEEE Standards Association and elsewhere. Industry best practices are under development at the Partnership on AI, and Underwriters

Laboratories is developing sector-specific standards for autonomous vehicles. More information on these efforts is available in Appendix 1.

1B. Needs and challenges for further AI standards

Question 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;

As indicated by industry and policymakers, there is an observed need for AI standards to support the safe and trustworthy research, development, deployment, and use of AI technologies across sectors (Google 2019; Finkel 2018; CESI 2018). In determining this need, there exists a challenge in relying predominantly on the concerns of private-sector firms. Market incentives motivate private development and use of standards via, e.g., customer demand (consumer and/or business), reputational benefits (regulator, consumer, and/or business), liability protection (judicial deference and insurance). Such incentives will be stronger for standards that are used in and related to market transactions. Thus, the observed level of market need for AI standards likely underrepresents the need for standards areas of AI safety research and development that may not be directly incentivized by market transactions.

Fundamental research, in particular, presents needs for AI standards that are not reflected in the market today (Cihon 2019). As the development of the technology continues, increasingly capable systems could pose serious risks. The field of AI Safety seeks to address these and related concerns in model specification, oversight, and robustness (See Amodei et al. 2016; Ortega et al. 2018). AI Safety requires further fundamental research, measurement, and standards development. Without greater government involvement, this need will not be met.

Question 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.

Technical standards and guidance ought to address the research, development, deployment, and use of AI technologies. To that end, we advocate that such technical standards help realize the principles set forth in the "Recommendation of the Council on Artificial Intelligence" produced by the OECD and adopted by the United States, most especially in the principles described in "Transparency and explainability" and on "Robustness, security, and safety." For example, these principles call upon all OECD nations to ensure that AI systems are "robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk."

Another useful starting place for scoping the questions at stake in standards development is the draft “Trustworthy AI Assessment List” found in the report, "Ethics Guidelines for Trustworthy AI," published by the High-Level Expert Group on Artificial Intelligence, an independent advisory body to the European Commission. The Assessment List covers issues including technical robustness and safety; privacy and data governance; transparency; accountability; human agency and oversight; diversity, non-discrimination, and fairness; and societal and environmental well-being. Questions on each topic are intended to probe the preparation and robustness of processes in place by AI developers and practitioners. For example, questions include, “Who is the “human in control,” and what are the moments or tools for human intervention?” and “Did you assess potential forms of attacks to which the AI system could be vulnerable?” This list can inform the scoping of topic areas where standards and guidance is currently lacking.

In addition to thinking about relevant topic areas and questions to consider, there are several standards that may help guide trustworthy AI development and deployment. For example, we highlight the following five ideas for the federal government’s consideration in drafting and executing the Federal AI Standards Engagement Plan:

1. **Adversarially Robust Training Certificates.** Machine learning models are susceptible to overconfidence about inputs that are qualitatively different or separate from the types of data they were trained on, whether these novel inputs are naturally out of distribution or they are actually adversarial. By using techniques and concepts from e.g. Roth et al. (2018) and Sinha et al. (2017), procedures can be specified to train the system in an adversarially robust manner, and furthermore to do so with specialized tests that provide robustness guarantees. The technical certificates in this literature would be extended to, and wrapped with, standardized process certifications. If properly developed, these certificates could provide some assurance with respect to aspects of the AI system’s accuracy, security, and robustness. Such a certificate would support reliable deployment in real-world situations that can include situations far from what has been trained on, including adversarial situations.

2. **Standardized Explainability Levels.** A framework standard could develop a typology of explainability levels, possibly through a tiered definition or scoring system. For example, these standardized levels could take into account factors to whom the AI system is explainable (e.g., the end user, any trained machine learning (ML) researcher, or only the AI’s original developer), the completeness of what can be accounted for by the explanation, the ease at which an explanation can be explained (e.g., by simple query or by complicated reverse engineering), and whether the transparency uncovers what the algorithm actually did or provides a post hoc rationalization for human review. These standardized explainability levels would support transparency and general trustworthiness of AI. This standard could also facilitate the integrated use of disparate AI systems while ensuring consistent explainability and transparency.

3. **Explainability with Domain-Specific Language.** Standards bodies could produce a standardized mechanism, via a purpose-designed ontology, for explaining models' behaviors with user-understandable vocabulary. More specifically, the creation of a connective ontology could subsequently be imported and extended per application or vertical. Examples of domain-specific ontologies this would enable AI explanations to be expressed in include, e.g., the Financial Industry Business Ontology in financial applications and NIH's UMLS for healthcare applications. Such ontological connectives enable Domain Specific Language support for explainability (See Walter et al. 2009). This would support transparency and general trustworthiness of AI.

4. **Safety Development Process Standard.** Existing standards development processes for safety focus on fail safe design, inspired by other industries, e.g., aviation. AI poses novel safety concerns (See, e.g., Amodei et al. 2016), however, and these can manifest across the research, development, and deployment of the technology. Thus, consideration should be given to process standards for the safe research and development of the technology, not simply when a product ships to market. One such standard would be a checklist for researchers to record a precise specification, measures taken to ensure robustness, and methods of assurance before implementing a system (See Ortega et al. 2018). Another approach would be to define high risk projects or a risk typology of multiple categories, with subsequent standards specifying best practices and mitigation strategies to be followed at each risk level.

5. **Machine-Readable Declarations to Advance Trustworthiness.** Trust in AI systems can be improved through the development of a standard for machine-readable declarations that enable AI system developers, or the AI system itself, to consistently attest to or explain certain conditions. Standardized machine-readable declarations could support compliance with public policy or regulatory requirements as they may emerge over time. For example, AI systems that support human decision-making should have standardized declarations that notify end-users of potential "conflicts of interest" to promote trustworthiness, e.g., a mapping service should declare if recommended routes are influenced by sponsorships. Machine-readable declarations also serve the implementation of other technical standards in practice, including:

    a. **Declaration of Side Effect Consideration.** AI agents acting in some environment, by both exploration within the environment and by exploitation of learned dynamics and behaviors, may cause side effects relative to their intended goals. By using techniques and concepts from Krakovna et al. (2019), designers of such systems can characterize, quantify, and potentially mitigate many such side effects. This proposed declaration, certifying both the types of side effects expected and mitigation design options employed, would support robustness and provide another means of establishing trust in an AI system or model. This declaration could follow a standardized format for communicating which classes of side effects of the system have been considered in an agent's

architecture and also the incentive structuring choices resulting from those classes.

b. **Safety Declaration.** Once safety standards are developed, standardized machine-readable declarations would offer a way for developers to document compliance with respect to which safety standards are implemented in the given system. Having these supporting processes for such a safety declaration will enable more reliable deployment of AI systems.

2. Defining and Achieving U.S. AI Technical Standards Leadership

Summary:

- The federal government can lead AI technical standards development if it properly engages with and leverages the expertise of world-leading U.S. private sector firms and academic research community.
- The federal government can support U.S. standards leadership through engagement in international standards efforts to support a global market for AI technologies. While every effort should be made to improve and adopt international standards, the federal government should carefully evaluate circumstances when such a standard would be "impractical" for use by federal agencies in accordance with Circular A-119.
- U.S. engagement in the development of international standards can support global trust in AI systems that can reduce possible dangers in the long-term development of the technology.
- The federal government should also support U.S. standards leadership through prioritizing funding for standards-essential fundamental research.

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

Today, the U.S. is a world leader in AI research and development. We define leadership in AI standards as a commitment to technically sound standards that support a global market in trustworthy AI. Leadership is not simply limited to maintaining technological superiority, but also engaging in international standards to support the development and deployment of safe and accountable AI systems. Despite its prowess in AI, U.S. leadership in technical standards for AI is limited. NIST has played an important role in hosting challenges and testing benchmarks, particularly for biometrics. But the limited U.S. private sector engagement in developing AI standards belies its technical expertise and market dominance globally today. Furthermore, the absence of a Federal strategy for AI standards contrasts with other national governments (See Ding et al. 2018; Dutton 2018; CESI 2018). The current development of the U.S. Plan for

Federal Engagement in Artificial Intelligence Standards will be a welcome chance to regain U.S. leadership in AI standards.

U.S. leadership demands engagement and alignment with standards development by international standards bodies. The Federal AI Standards Engagement Plan should explicitly include renewed efforts to engage in international initiatives underway such as at the ISO/IEC JTC 1 and IEEE, as well as other international partners. This engagement will comply with the intentions of the National Technology Transfer and Advancement Act, as further implemented by OMB Circular A-119, and support overall U.S. leadership.

International standards will shape the global market for AI systems, a market in which the U.S. private sector is the most competitive today. International standards can support fair competition for U.S. private-sector firms in foreign markets. U.S. leadership should evaluate the use of the World Trade Organization dispute settlement system to challenge violations of the Technical Barriers to Trade Agreement (TBT). Greater enforcement of TBT will support international standards and the global success of the U.S. private sector. In addition to supporting U.S. leadership, international standards can support global trust in AI systems and their development that can reduce possible dangers in the long-term development of the technology (Cihon 2019).

Today, the U.S. is home to leading AI research labs and/or their corporate parents, and these companies should proactively engage in the development of international standards. The federal government can further national leadership by encouraging and supporting its leading private sector organizations to engage in these fora. Support can take the form of, inter alia, knowledge sharing on standardization processes and funding for academic experts and startups to engage at international fora. Expanded engagement will bring expertise to improve the quality of standards and increase the likelihood that resulting standards will reflect the needs of U.S. industry internationally.

In the context of its engagement, the federal government should carefully evaluate if, how, and when any developed international standard may be “impractical” for use by federal agencies. As defined by Circular A-119, “impractical” includes circumstances when the standard is either “inadequate, ineffectual, inefficient, or inconsistent with agency mission;...” Thus, after robustly engaging in the development of international standards, if NIST or other U.S. agencies assess that those standards are impractical in key regards, the U.S. should be willing to develop additional standards for the safe and ethical use of AI.

Additionally, leadership in AI technical standards development can be furthered by increasing support for standards-essential fundamental research. The U.S. should pursue standards leadership through support for fundamental research in priority standards areas, including safety, transparency, security, and accuracy. Increased research support will also enable more timely delivery of essential standards. Further detail on suggested research interventions is presented below in the answers to Questions 15 and 18.

3. Prioritizing Federal Government Engagement in AI Standardization

Summary:

- The federal government's current approach and policy towards standards development is adequate in that it properly encourages private sector engagement and favors performance standards. However, the Federal AI Standards Engagement Plan should prioritize international standards above national ones.
- The federal government has a unique need for standards that support algorithmic accountability and bias mitigation. These standards will be useful for state and local governments as well as the economy as a whole.
- The federal government should prioritize engagement in cross-sector international standards for safety, transparency, and security. This prioritization should include funding for standards-essential research.
- The federal government can help ensure standards and guidance are useful by: (1) supporting private-sector engagement in their development; (2) signaling the potential use of standards as future requirements in government procurement; (3) developing a National Testbed for AI within NIST; (4) issuing plain language, generalized descriptions of any standards' potential applicability to essential policy matters in their final adoption; and (5) creating technical assistance programs to reduce the burden of standards adoption for small businesses.

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

The Federal government has a unique need for technical standards that support algorithmic accountability and mitigate bias in AI systems used for government business, e.g., in the provision of government services and the administration of the criminal justice system (see, e.g., Partnership on AI 2019). Algorithmic accountability can be defined as the obligation for a decision-maker to "provide its decision-subjects with reasons and explanations for the design and operation of its automated decision-making system" (Binns 2017, 544). These needs can be partially met through standards for transparency and interpretability of AI systems. Although they are of particularly acute need for the Federal government, these standards will also support needs of state and local government. For example, a bill introduced in California this year (AB-459) would require that a possible "AI in State Government Services Commission" recommend standards for government use of AI to ensure accountability, prioritize safety and security, protect privacy, and monitor impacts. The bill also calls for the measurement of reliability and robustness, and minimizing the potential for misuse. Beyond government use,

standards that support algorithmic accountability will support applications across the U.S. economy.

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

The U.S. Federal government should prioritize engagement in cross-sectoral standards for safety, transparency, security, and accuracy that require further fundamental research. This prioritization should carry through to R&D funding, data, and compute allocation in executing the Executive Order on Maintaining American Leadership in Artificial Intelligence. In particular, federal prioritization of safety research and safety standards in fundamental research can address the challenges to standards development that may be underprovided by the private sector, as explained in the response to Question 3 above. Federal engagement in standardization processes for safety, transparency, security, and accuracy can support the translation of needed fundamental research directly to standards.

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

Current Federal policies (National Technology Transfer and Advancement Act; OMB Circular A-119) are adequate for the ongoing development of AI standards. With thoughtful guidance from NIST in place, the U.S. private sector is likely to have the expertise and resources to lead on AI standards development. The Federal government should support and encourage leading U.S. firms to engage in ongoing consensus standardization activities. Support can take the form of, inter alia, knowledge sharing on standardization processes and funding for academic experts and startups to engage at standards bodies.

Current policy also enables a focus on the future development of the technology, which is essential for standards to support continued AI innovation. AI systems will increase in capabilities as greater data, funding, and research talent continue to flow into development of the technology. Standards that are under development today can support the beneficial development of more capable AI systems (See Cave & ÓhÉigeartaigh 2019). In this regard, it is important to emphasize the continued U.S. policy of prioritizing performance standards over prescriptive standards. Engagement in standards processes developing performance standards for safety, transparency, security, and accuracy can avoid stifling innovation with prescriptive system requirements. Once established, performance standards can contribute to the beneficial development of more advanced systems.

Current policy does not create a preference for international over national standards, “However, in the interests of promoting trade and implementing the provisions of international treaty agreements, your agency should consider international standards in procurement and regulatory applications” (OMB Circular A-119, 9). Insofar as the Federal AI Standards Engagement Plan prioritizes international standards, leading private-sector labs will be more likely to engage in these international standards bodies. This would help to improve the quality of international standards and support a greater international market share in AI for the U.S. private sector in the future.

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

The Federal government should consider taking five actions to further support desired AI technical standards to be useful and incorporated into practice. The following actions may be taken by relevant agencies within the Federal government at large, not simply NIST, and may require acts of Congress if additional authority or funding is required:

First, in order to support standards-essential research, the federal government should consider creating a National Testbed for AI within NIST. The National Testbed would serve to integrate relevant expertise now scattered across NIST divisions by drawing on and coordinating among its other Testbed facilities, particularly that of the Robotics Test Facility, the Systems Engineering Group, the Privacy Engineering Program, the Information Access Division, and the National Cybersecurity Center of Excellence. Externally, the Testbed can serve as a locus for collaborative research essential to measurement, benchmarks, and standards for AI systems. The Testbed can provide government resources, i.e., datasets and computing resources, to interested academics and private-sector researchers. Alternatively, the Federal government may consider pursuing a similar model in a National Laboratory.

Second, the government should support and encourage U.S. private-sector AI research organizations to participate in ongoing standardization processes. This encouragement and support, e.g., knowledge sharing on standardization processes and funding for academic experts and startups to engage at international standards bodies, can ensure that standards reflect the world-leading expertise of the U.S. private sector.

Third, the Federal AI Standards Engagement Plan should signal that standards will be used to create requirements in future government procurement contracts. This would incentivize greater private-sector participation in standards development and encourage higher quality standards that can benefit both Federal use and the economy as a whole. Insofar as the Plan emphasizes procurement requirements for AI safety in the process of research and development, as well as deployment and use, the government could partially address the possible market failure in AI safety.

Fourth, in order to increase the usefulness of technical AI standards, the Federal government should openly anticipate and plan for the potential legal and economic ramifications of their issuance. For example, it is predictable that such standards may become a de facto form of liability protection for private sector actors that adequately comply with them, even if applicable laws do not explicitly provide such a safe harbor provision. Further, technical standards, even if performance-based, can result in clear economic "winners" in those that hold intellectual property patents that are potentially critical for meeting the new standards. In anticipation of these and other ramifications, the federal government should consider issuing plain language, generalized descriptions of the standards *potential* applicability to such policy matters in conjunction with their final adoption. These could take the form of "frequently asked questions" or other forms of additional helpful guidance.

Fifth, to be useful, developers of technical standards must anticipate that such standards can serve as a barrier to entry for smaller private sector actors wishing to develop and deploy AI-based technologies. While this anticipated outcome should *never* result in the creation of less effective standards than what are necessary for the public good (especially in the realm of AI safety and data privacy), anticipating this outcome should necessitate planning by the Federal government on how to mitigate the effects of this potentially unhealthy market distortion. This could involve, for instance, the creation of technical assistance programs that coincide with the development and adoption of standards, or federal subsidies for the necessary training and education required to properly institute the standards in practice.

Conclusion

The U.S. can secure leadership in AI standards and reduce risks by engaging in the development of standards, especially at the international level. The U.S. private sector currently leads in AI research and development, but requires federal support to engage successfully at international standards bodies. Effective standards require further fundamental research. The Federal government can provide essential assistance in this research, especially in neglected areas including AI safety and security. We are confident that an effective Federal AI Standards Engagement Plan can support the beneficial development of the technology. We stand ready to support NIST throughout the critical process of developing this Plan.

For further information or clarification, please contact:

- Peter Cihon, Research Affiliate, Center for the Governance of AI, Future of Humanity Institute, University of Oxford, at petercihon@gmail.com.
- Jared Brown, Senior Adviser for Government Affairs, Future of Life Institute, at jared@futureoflife.org.
- Jessica Cussins Newman, Research Fellow, Center for Long-Term Cybersecurity, University of California, Berkeley, at jessica.cussins@berkeley.edu.

- Haydn Belfield, Associate Fellow, Leverhulme Centre for the Future of Intelligence, University of Cambridge, at hb492@cam.ac.uk.

References

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete problems in AI safety." *arXiv preprint* (2016). https://arxiv.org/abs/1606.06565.

Binns, Reuben. "Algorithmic accountability and public reason." *Philosophy & Technology* 31, no. 4 (2018): 543-556. https://link.springer.com/article/10.1007/s13347-017-0263-5

Cave, Stephen, and Seán S. ÓhÉigeartaigh. "Bridging near-and long-term concerns about AI." *Nature Machine Intelligence* 1, no. 1 (2019): 5. https://www.nature.com/articles/s42256-018-0003-2.

China Electronics Standardization Institute (CESI). "AI Standardization White Paper." (2018). Translation by Jeffrey Ding. https://docs.google.com/document/d/1VqzyN2KINmKmY7mGke_KR77o1XQriwKGsuj9dO4MTDo/.

Cihon, Peter. "Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development." Technical Report, *Future of Humanity Institute, University of Oxford*. (2019). https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.

Ding, Jeffrey, Paul Triolo, and Samm Sacks. "Chinese Interests Take a Big Seat at the AI Governance Table." *New America*. (2018). https://www.newamerica.org/cybersecurity-initiative/digichina/blog/chinese-interests-take-big-seat-ai-governance-table/.

Dutton, Tim. "An Overview of National AI Strategies." *Medium, Politics + AI*. (2018). https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd.

Finkel, Alan. "What will it take for us to trust AI?" *World Economic Forum*: *Agenda*. (2018). https://www.weforum.org/agenda/2018/05/alan-finkel-turing-certificate-ai-trust-robot/.

Google. "Perspectives on Issues in AI Governance." *Google AI*. (2019). https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.

High-Level Expert Group on AI. "Ethics Guidelines for Trustworthy AI." *European Commission*. (2019). https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

International Electrotechnical Commission (IEC). “Artificial intelligence across industries.” *IEC*. (2018). https://basecamp.iec.ch/download/iec-white-paper-artificial-intelligence-across-industries-en/.

Krakovna, Victoria, Ramana Kumar, Laurent Orseau, Alexander Turner. “Designing agent incentives to avoid side effects.” *Medium, DeepMind Safety Research*. (2019). https://medium.com/@deepmindsafetyresearch/designing-agent-incentives-to-avoid-side-effects-e1ac80ea6107.

OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Ortega, Pedro A., Vishal Maini, and the DeepMind safety team. “Building safe artificial intelligence: specification, robustness, and assurance.” *Medium, DeepMind Safety Research*. (2018). https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1.

Partnership on AI. “Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System.” Report, *Partnership on AI*. (2019). https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/.

Roth, Kevin, Aurelien Lucchi, Sebastian Nowozin, and Thomas Hofmann. "Adversarially robust training through structured gradient regularization." *arXiv preprint* (2018). https://arxiv.org/abs/1805.08736.

Sinha, Aman, Hongseok Namkoong, and John Duchi. "Certifiable distributional robustness with principled adversarial training." *stat* 1050 (2017): 29. https://www.researchgate.net/publication/320727277_Certifiable_Distributional_Robustness_with_Principled_Adversarial_Training.

Walter, Tobias, Fernando Silva Parreiras, and Steffen Staab. "OntoDSL: An ontology-based framework for domain-specific languages." In *International Conference on Model Driven Engineering Languages and Systems,* pp. 408-422. Springer, Berlin, Heidelberg, 2009. https://www.researchgate.net/publication/226633047_OntoDSL_An_Ontology-Based_Framework_for_Domain-Specific_Languages.

Appendix 1

There are multiple private cross-sector consensus-based standards under development today at ISO/IEC JTC 1 SC 42 (herein: SC 42) and the IEEE Standards Association. Lists of ongoing activities on AI at both are available here:

- SC 42: https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0
- IEEE P7000 Series: https://ethicsstandards.org/p7000/

Standards currently under development that will be helpful to reduce barriers to the safe testing and deployment of AI systems and that can support reliable, robust, and trustworthy systems that use AI technologies include:

- IEEE P7000 Model Process for Addressing Ethical Concerns During System Design
  (Current project expiration date: December 2020)
- IEEE P7001 Transparency of Autonomous Systems
  (Current project expiration date: December 2020)
- IEEE P7003 Algorithmic Bias Considerations
  (Current project expiration date: December 2021)
- IEEE P7009 Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
  (Current project expiration date: December 2021)

Standards development at SC 42 is in its early stages. Work is ongoing for foundational standards:

- "Concepts and terminology" (Anticipated publication: March 2021) and
- "Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)"
  (Anticipated publication: March 2022)

SC 42 is also developing Technical Reports on topics that will be helpful to reduce barriers to the safe testing and deployment of AI systems, including:

- "Assessment of the robustness of neural networks"
  (Anticipated publication in August 2019)
- "Overview of trustworthiness in Artificial Intelligence"
  (Anticipated publication in August 2019)
- "Bias in AI systems and AI aided decision making"
  (Anticipated publication in January 2021)

Also under development are industry best practices at the Partnership on AI, which focuses on six thematic pillars:

1. Safety-Critical AI
2. Fair, Transparent, and Accountable AI
3. AI, Labor, and the Economy

4. Collaborations Between People and AI Systems
5. Social and Societal Influences of AI
6. AI and Social Good

Thus far, the best practice effort most relevant to technical standards is the "Annotation and Benchmarking on Understanding and Transparency of Machine learning Lifecycles" (ABOUT ML). More information is available here: https://www.partnershiponai.org/the-partnership-on-ai-launches-multistakeholder-initiative-to-enhance-machine-learning-transparency/

The ITU-T is pursuing standardization work for AI in the following working groups:

- Focus Group on Machine Learning for Future Networks including 5G
    - https://www.itu.int/en/ITU-T/focusgroups/ml5g/Pages/default.aspx
- Focus Group on Artificial Intelligence for Health
    - https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx
- Focus Group on Environmental Efficiency for Artificial Intelligence and other Emerging Technologies
    - https://www.itu.int/en/ITU-T/focusgroups/ai4ee/Pages/default.aspx

In addition to AI-specific standards, ISO/IEC JTC 1 and ISO have produced standards on a series of related topics, see each Standards Committee for further information:

- JTC 1 SC 7: Software and systems engineering
- JTC 1 SC 17: Cards and security devices for personal identification
- JTC 1 SC 22: Programming languages, their environments and system software interfaces
- JTC 1 SC 24: Computer graphics, image processing and environmental data representation
- JTC 1 SC 27: Information Security, cybersecurity and privacy protection
- JTC 1 SC 28: Office equipment
- JTC 1 SC 29: Coding of audio, picture, multimedia and hypermedia information
- JTC 1 SC 36: Information technology for learning, education and training
- JTC 1 SC 37: Biometrics
- JTC 1 SC 40: IT Service Management and IT Governance
- JTC 1 SC 41: Internet of Things and related technologies
- ISO TC 184: Automation systems and integration
- ISO TC 199: Safety of machinery
- ISO TC 299: Robotics

For further information on efforts at IEC and ETSI, please see IEC (2018).

The authors are aware of few AI-specific, sector-specific standards. Underwriters Laboratories is developing UL 4600, a safety standard for autonomous vehicles. Its proposed development

timeline would see it published in Q4 2019. More information is available here: https://edge-case-research.com/ul4600/.