PARTNERSHIP ON AI

AI Risk Management Framework c/o Mark Przybocki
U.S. National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899
AIframework@nist.gov

August 19, 2021

The Partnership on AI Response to the National Institute of Standards and Technology Request for Information - AI Risk Management Framework

Dear Mr. Przybocki:

The Partnership on AI (PAI) is a non-profit partnership of academic, civil society, industry, and media organizations creating solutions so that AI advances positive outcomes for people and society. The Partnership was established to study and formulate sociotechnical approaches to the responsible development of AI technologies, to advance the public's understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society. Today, PAI convenes nearly 100 partner organizations from around the world to be a uniting force for the responsible development and fielding of AI technologies. Partnership staff composed this response based on some of PAI's recent work, much of it inspired and informed collectively by consulting over several years with our international group of multi-stakeholder Partner organizations. The information in this document is provided by PAI and is not intended to reflect the view of any particular Partner organization of PAI.

PAI develops tools, recommendations, and other resources by inviting diverse voices from across the AI community and beyond to share insights that can be synthesized into actionable guidance. We then work to drive adoption in practice, inform public policy, and advance public understanding. Through dialogue, research, and education, PAI is addressing some of the most important and difficult questions concerning the future of AI. PAI and its Partners inform the development of norms for the responsible development of AI technologies in four current Program areas: (1) AI & Media Integrity, (2) AI, Labor and the Economy, (3) Fairness, Transparency and Accountability, and (4) Safety Critical AI.

Response to NIST - AI Risk Management Framework RFI Points 5 & 7

With regard to the specific request for information on NIST's AI Risk Management Framework in points 5 and 7 (referenced below), PAI is pleased to submit two brief examples of our related work, and to provide a more detailed focus on the AI Incident Database.

> 5. "Standards, frameworks, models, methodologies, tools, guidelines and best practices, and principles to identify, assess, prioritize, mitigate, or communicate AI risk and whether any currently meet the minimum attributes described above;

7. AI risk management standards, frameworks, models, methodologies, tools, guidelines and best practices, principles, and practices which NIST should consider to ensure that the AI RMF aligns with and supports other efforts;"

One example of PAI's work in this area is a recent (2021) Safety Critical AI report on "Managing the Risks of AI Research: Six Recommendations for Responsible Publication," which addresses some of the potential risks of AI research and makes recommendations regarding research publication and dissemination practices in order to minimize its misuse.

A second example is PAI's Annotating and Benchmarking on Understanding and Transparency in Machine learning Lifecycles (ABOUT ML), which brings together a wide range of stakeholders to advance public discussion, and promulgate best practices into new norms for greater transparency in the use of ML in industry, government, and civil society. Current research seeks to address the organizational, technological, or other challenges for implementing documentation in key phases throughout the ML system lifecycle, from design to deployment, including annotations of data, algorithms, performance, and maintenance requirements.

The third example that we would like to focus on is the AI Incident Database[1] (AIID), which is supported by PAI. The database is a tool to identify, assess, manage, and communicate AI risk and harm. Currently, the Database is the only collection of AI deployment harms or near harms across all disciplines, geographies, and use cases. The AI Incident Database was created as an open, collective record of AI harms to inform the beneficial development of AI technologies moving forward. Leading the development and management of the project are Sean McGregor, PhD, a machine learning researcher and technical lead for the IBM Watson AI XPRIZE at the XPRIZE Foundation (a PAI partner), and Cristine Custis, PhD, Head of ABOUT ML and Fairness, Transparency, and Accountability at PAI. Additional PAI Partners provided important input into the development of the database.

The database is a constantly evolving data product. Current and intended users include system architects, industrial product developers, academics, researchers, public relations managers, standards organizations, and policy makers. These users are invited to use the Discover application to proactively explore how recently deployed intelligent systems have produced unexpected outcomes in the real world. In doing so, they may avoid making similar mistakes in their development.

AI Risk Management Framework Attributes

The AI Incident Database involves a number of the minimum attributes in NIST's AI Risk Management Framework.

- **"Be consensus-driven and developed and regularly updated through an open, transparent process" (Attribute 1); "Be risk-based, outcome-focused, voluntary, and non-prescriptive" (Attribute 5)**

[1] https://incidentdatabase.ai/

The development of the incident database is managed in a participatory manner by individuals and organizations contributing code and incidents over time. They are invited to submit reports to the database, whereupon incidents are indexed and made discoverable for those people developing and deploying both contemporary and next generation AI technologies.

Stakeholders seldom reach consensus on many aspects of an incident such as the scale of negative impacts, the cause, etc. As detailed by McGregor (2020) in a AAAI/IAAI paper, the AI Incident Database does not prescribe a single classification scheme. Organizations with expertise in safety, fairness, and sectoral or population-specific interests can each develop and manage their own view into the dataset. Additionally, the intention is not to prescribe a solution to an organization's discovered harm. Instead, the Database is an opportunity to report harms and collect best practices for risk mitigation and future related policy change. Thus, the Database supports multiple perspectives on incidents both by ingesting multiple reports (to date, approximately 1199 authors/sources), and by supporting multiple taxonomies.

- **"Provide common definitions" (Attribute 2); "Use plain language that is understandable by a broad audience, including senior executives and those who are not AI professionals" (Attribute 3);**

  One of the challenges in developing an open sourced, shared risk management resource such as the AI Incident Database is the potentially different definitions deployed by users. The AI Incident Database has approached the definitional challenge by allowing the broader community to converge on a shared definition of "AI Incident" through exploration of the candidate incidents submitted. Generally, an "AI incident" is a situation in which AI systems caused, or nearly caused, real-world harm.

  To begin to develop a taxonomy for incident submissions to date, the database employs a process developed by the Georgetown Center for Security and Emerging Technology (CSET). The CSET taxonomy is developed through a process that involves peer review to make classifications consistent across multiple annotators. To implement this process, CSET develops a large number of classified attributes, including ones pertaining to safety, fairness, industry, geography, timing, and cost. All classifications within the CSET taxonomy are first applied by one CSET annotator and reviewed by another CSET annotator before the classifications are finalized. Additionally, CSET invites public reporting of errors in classification for adjudication by CSET staff. The ability to integrate multiple viewpoint taxonomies through the CSET taxonomy into the system architecture is intended to ensure that the system is not only representative of a single viewpoint of AI incidents.

- **"Be consistent, to the extent possible, with other approaches to managing AI risk" (Attribute 7)**

  The AI Incident Database is inspired by, and combines the strengths of, similar incident databases in aviation and computer security. In aviation, an "accident" is a case where

substantial damage or loss of life occurs. “Incidents” on the other hand are cases where the risk of an accident substantially increases. The FAA aviation database indexes flight log data and subsequent expert investigations into comprehensive examinations of both technological and human factors. In part due to this continual self-examination, air travel is one of the safest forms of travel. Decades of iterative improvements to safety systems and training have decreased fatalities 81 fold since 1970 when normalized for passenger miles.

Another risk management approach inspiring the AI Incident Database is the Common Vulnerabilities and Exposures (CVE) system, which contains more than 159,000 publicly disclosed cybersecurity vulnerabilities and exposures. The CVE site serves as critical security infrastructure across all industries by enabling vulnerabilities to be circulated and referenced with a consistent identifier.

- **“Be adaptable to many different organizations, AI technologies, lifecycle phases, sectors, and uses” (Attribute 4); “Be a living document” (Attribute 8)**

  The AI Incident Database is built on a document database and a collection of serverless browser applications in order to be highly extensible, and to allow an unlimited number of views into the application layer of the database. The intention behind this architecture is to allow many communities to have different views both within and into the data. The Database is intended to be a living and adaptable analytical framework from which we collectively produce better intelligent systems that avoid potential harmful outcomes. Anyone can submit an app that interfaces with the incident data. More details can be found on the GitHub repository. Future uses will evolve through the code contributions of the open source community, including additional database summaries and taxonomies.

Thank you for this opportunity to provide some information about the Partnership on AI’s work related to AI risk management, specifically the AI Incident Database. One of the benefits of multi-stakeholder organizations such as PAI is the opportunity to convene and connect diverse perspectives from across sectors, disciplines, geographies and lived experiences - a critical component to understanding and developing risk frameworks. The Partnership on AI is happy to provide more details or additional information about the research, workshops, convenings, and other activities we conduct to understand risk and develop resources to prevent harms and promote the development of AI that benefits people and society.

Regards,

Rebecca Finlay
Acting Executive Director

Mark Latonero, PhD
Senior Policy Advisor