IDS WG Meeting Minutes
Jun 22, 2023

This IDS WG Meeting was started at approximately 3:00 pm ET on June 22, 2023.

Attendees

Graydon Dobson	Lexmark
Jeremy Leber	Lexmark
Alan Sukert	
Steve Young	Canon

Agenda Items

1. The topics to be covered during this meeting were:
   - Special Topic on US Artificial Intelligence (AI) Legislation
   - Special Topic on NIST AI 100-1 NIST AI Risk Management Framework (AI RMF 1.0).
2. Meeting began by stating the PWG Anti-Trust Policy which can be found at https://www.pwg.org/chair/membership_docs/pwg-antitrust-policy.pdf and the PWG Intellectual Property Policy which can be found at https://www.pwg.org/chair/membership_docs/pwg-ip-policy.pdf.
3. Al began with a status of the EU AI Act. The latest status of this law is that the European Parliament has approved its negotiating position on the proposed AI Act on Jun 14th. It also amended the list of intrusive and discriminatory uses of AI to include 6 new items (most biometric-related). The next step is for Parliament to negotiate with the EU Council and the European Commission to reach a provisional agreement on a legislative proposal that is acceptable to both the Parliament and the Council, the co-legislators.

   This provisional agreement must then be adopted by each of the institution's formal procedures. Once that happens, the AI Act will be approved and implemented. The hope is that this will happen sometime before the end of 2023.
4. Al presented the first special topic on the US AI legislation. Currently there are four US Government AI-related Laws and Executive Orders (EOs) in effect:
   - **AI In Government Act of 2020**
   - **National Artificial Intelligence Initiative Act of 2020**
   - **Executive Order 13859 Maintaining American Leadership in Artificial Intelligence**
   - **Executive Order 13960 Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government**

   The slides Al used to discuss these four laws/EOs can be found at https://ftp.pwg.org/pub/pwg/ids/Presentation/US AI Legislation.pdf.

   a. **AI In Government Act of 2020**
      - This was not a standalone law; it was actually Division U of the Consolidated Appropriations Act, 2021. Its purpose was to create the AI Center of Excellence (AI CoE) to:
        - Facilitate the adoption of artificial intelligence technologies in the Federal Government; and
        - Improve cohesion and competency in the adoption and use of artificial intelligence within the Federal Government
      - The duties of the AI CoE were to be:
        - Regularly convening individuals from agencies, industry, Federal laboratories, nonprofit organizations, institutions of higher education, and other entities to discuss recent developments in artificial intelligence;

  - Collecting, aggregating, and publishing on a publicly available website information regarding programs, pilots, and other initiatives led by other agencies and any other information determined appropriate by the Administrator;
  - Advising the Administrator, the Director, and agencies on the acquisition and use of artificial intelligence through technical insight and expertise, as needed;
  - Assist agencies in applying Federal policies regarding the management and use of data in applications of artificial intelligence;
  - Consulting with agencies, including the Department of Defense, the Department of Commerce, the Department of Energy, the Department of Homeland Security, the Office of Management and Budget, the Office of the Director of National Intelligence, and the National Science Foundation, that operate programs, create standards and guidelines, or otherwise fund internal projects or coordinate between the public and private sectors relating to artificial intelligence;
  - Advising the Director on developing policy related to the use of artificial intelligence by agencies
  - Advising the Director of the Office of Science and Technology Policy on developing policy related to research and national investment in artificial intelligence

- The law includes the following requirement with respect to providing guidance to Federal Agency use of AI:

  No later than 270 days after enactment of this act the Director of the Office of Management and Budget (OMB) shall issue a memorandum to the head of each agency that shall—

  - Inform the development of policies regarding Federal acquisition and use by agencies regarding technologies that are empowered or enabled by artificial intelligence, including an identification of the responsibilities of agency officials managing the use of such technology;
  - Recommend approaches to remove barriers for use by agencies of artificial intelligence technologies in order to promote the innovative application of those technologies while protecting civil liberties, civil rights, and economic and national security;
  - Identify best practices for identifying, assessing, and mitigating any discriminatory impact or bias on the basis of any classification protected under Federal nondiscrimination laws, or any unintended consequence of the use of artificial intelligence, including policies to identify data used to train artificial intelligence algorithms as well as the data analyzed by artificial intelligence used by the agencies; and
  - Provide a template of the required contents of the agency plans

- Another requirement in the law was that not later than 180 days after the date on which the Director of the OMB issues the memorandum required under subsection (a) or an update to the memorandum required under subsection (d), the head of each agency shall submit to the Director and post on a publicly available page on the website of the agency:
  - (1) a plan to achieve consistency with the memorandum; or
  - (2) a written determination that the agency does not use and does not anticipate using artificial intelligence.
  - UPDATES.—Not later than 2 years after the date on which the Director of the OMB issues the memorandum required under subsection (a), and every 2 years thereafter for 10 years, the Director of the OMB shall issue updates to the memorandum

- Lastly, with respect to training the law included the requirement that not later than 18 months after the date of enactment of this Act, and in accordance with chapter 51 of title 5, United States Code, the Director of the Office of Personnel Management shall —
  - (1) Identify key skills and competencies needed for positions related to artificial intelligence;

- (2) Establish an occupational series, or update and improve an existing occupational job series, to include positions the primary duties of which relate to artificial intelligence;
- (3) To the extent appropriate, establish an estimate of the number of Federal employees in positions related to artificial intelligence, by each agency; and
- (4) Using the estimate established in paragraph (3), prepare a 2-year and 5-year forecast of the number of Federal employees in positions related to artificial intelligence that each agency will need to employ
- PLAN.—Not later than 120 days after the date of enactment of this Act, the Director of the Office of Personnel Management shall submit to the Committee on Homeland Security and Governmental Affairs of the Senate and the Committee on Oversight and Reform of the House of Representatives a comprehensive plan with a timeline to complete requirements described above

Al noted that this law applied only to the Federal Government only and these duties had no impact on AI activities or issues for activities performed outside of the Federal Government.

b. **National Artificial Intelligence Initiative Act of 2020**

The National Artificial Intelligence Initiative Act of 2020 was also not a standalone act; it was Division E, Section 5001 of the “WILLIAM M. (MAC) THORNBERRY NATIONAL DEFENSE AUTHORIZATION ACT FOR FISCAL YEAR 2021”.

One of the interesting things in this law was how it “defined” Artificial Intelligence - ‘‘artificial intelligence’’ means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to— (A) perceive real and virtual environments; (B) abstract such perceptions into models through analysis in an automated manner; and (C) use model inference to formulate options for information or action. It is definitely not the definition we are used to seeing for AI.

The stated purposes of this law were to:

- Ensure continued United States leadership in artificial intelligence research and development;
- Lead the world in the development and use of trustworthy artificial intelligence systems in the public and private sectors;
- Prepare the present and future United States workforce for the integration of artificial intelligence systems across all sectors of the economy and society; and
- Coordinate ongoing artificial intelligence research, development, and demonstration activities among the civilian agencies, the Department of Defense and the Intelligence Community to ensure that each informs the work of the others

The National Artificial Intelligence Initiative that this law created was to perform the following activities:

- Sustained and consistent support for artificial intelligence research and development through grants, cooperative agreements, testbeds, and access to data and computing resources
- Support for K-12 education and postsecondary educational programs, including workforce training and career and technical education programs, and informal education programs
- Support for interdisciplinary research, education, and workforce training programs for students and researchers that promote learning in the methods and systems used in artificial intelligence and foster interdisciplinary perspectives and collaborations among subject matter experts in relevant fields
- Interagency planning and coordination of Federal artificial intelligence research, development, demonstration, standards engagement, and other activities under the Initiative, as appropriate
- Outreach to diverse stakeholders, including citizen groups, industry, and civil rights and disability rights organizations

- Leveraging existing Federal investments to advance objectives of the Initiative
- Support for a network of interdisciplinary artificial intelligence research institutes
- Support opportunities for international cooperation with strategic allies, as appropriate, on the research and development, assessment, and resources for trustworthy artificial intelligence systems

Like any US law it created a whole bureaucracy around the establishment of the National Artificial Intelligence Initiative by creating the following committees or groups to administer the initiative (Note – this is a quick summary of these committees and groups; more details are in the slides available in the link shown above):

- A ‘‘National Artificial Intelligence Initiative Office’’ to (1) Provide technical and administrative support to the Interagency Committee and the Advisory Committee and (2) Serve as the point of contact on Federal artificial intelligence activities for Federal departments and agencies, industry, academia, nonprofit organizations, professional societies, State governments, and such other persons as the Initiative Office considers appropriate to exchange technical and programmatic information
- An Interagency Committee to provide for interagency coordination of Federal artificial intelligence research, development, and demonstration activities and education and workforce training activities and programs of Federal departments and agencies undertaken pursuant to the Initiative
- A National Intelligence Advisory Committee to advise the President and the Initiative Office on matters related to the Initiative, including recommendations related to areas like (1) the current state of United States competitiveness and leadership in artificial intelligence; (2) the progress made in implementing the Initiative; and (3) the state of the science around artificial intelligence
- A National AI Research Task Force to investigate the feasibility and advisability of establishing and sustaining a National Artificial Intelligence Research Resource; and to propose a roadmap detailing how such resource should be established and sustained

Finally, the law established roles for three National AI Research Institutes – the National Institute of Standards and Technology (NIST), the National Oceanic and Atmospheric Administration Artificial Intelligence Center and the National Science Foundation (NSF).

- For NIST, there were several key roles mentioned in the law (see the details in the slides available in the link shown above), but the two that are most relevant are:
    - support and strategically engage in the development of voluntary consensus standards, including international standards, through open, transparent, and consensus-based processes
    - RISK MANAGEMENT FRAMEWORK.—Not later than 2 years after the date of the enactment of this Act, the Director shall work to develop, and periodically update, in collaboration with other public and private sector organizations, including the National Science Foundation and the Department of Energy, a voluntary risk management framework for trustworthy artificial intelligence systems

    This law explicitly required NIST to create the AI Risk Management Framework that Al was to discuss in the second special topic later in the meeting.

- The National Oceanic and Atmospheric Administration Artificial Intelligence Center is under the National Oceanic and Atmospheric Administration (NAOA) and is responsible for tasks such as (1) coordinate and facilitate artificial intelligence research and innovation, tools, systems, and capabilities across the National Oceanic and Atmospheric Administration; (2) establish data standards and develop and maintain a central repository for agency-wide artificial intelligence applications; and (3) accelerate the transition of artificial intelligence research to applications in support of the mission of the National Oceanic and Atmospheric Administration

Al noted he was surprised NAOA was involved in AI research and development.

- Some of the AI-related tasks the NSF is responsible for are:
  - (1) support research, including interdisciplinary research, on artificial intelligence systems and related areas;
  - (2) use the existing programs of the National Science Foundation, in collaboration with other Federal departments and agencies, as appropriate to — (A) improve the teaching and learning of topics related to artificial intelligence systems in K-12 education and postsecondary educational programs; and (B) increase participation in artificial intelligence related fields;
  - (3) support partnerships among institutions of higher education, Federal laboratories, nonprofit organizations, State, local, and Tribal governments, industry, and potential users of artificial intelligence systems that facilitate collaborative research, personnel exchanges, and workforce development;
  - (4) ensure adequate access to research and education infrastructure with respect to artificial intelligence systems;

c. **Executive Order 13859 Maintaining American Leadership in Artificial Intelligence**

This Executive Order (EO) was issued February 11, 2019.

A couple of definitions that were included in this EO were:

- 'artificial intelligence'' means the full extent of Federal investments in AI, to include: R&D of core AI techniques and technologies; AI prototype systems; application and adaptation of AI techniques; architectural and systems support for AI; and cyberinfrastructure, data sets, and standards for AI; and
- ''open data'' shall, in accordance with OMB Circular A– 130 and memorandum M–13–13, mean ''publicly available data structured in a way that enables the data to be fully discoverable and usable by end users.

Note that the artificial intelligence definition is really not a definition.

The principle of this EO were to:

- Drive development of appropriate technical standards and reduce barriers to the safe testing and deployment of AI technologies
- Train current and future generations of American workers with the skills to develop and apply AI technologies
- Foster public trust and confidence in AI technologies and protect civil liberties, privacy, and American values in their application
- Promote an international environment that supports American AI research and innovation and opens markets for American AI industries, while protecting our technological advantage in AI and protecting our critical AI technologies from acquisition by strategic competitors and adversarial nations

The objectives of this EO include:

- Promote sustained investment in AI R&D in collaboration with industry, academia, international partners and allies, and other non-Federal entities
- Enhance access to high-quality and fully traceable Federal data, models, and computing resources to increase the value of such resources for AI R&D, while maintaining safety, security, privacy, and confidentiality protections consistent with applicable laws and policies
- Reduce barriers to the use of AI technologies to promote their innovative application while protecting American technology, economic and national security, civil liberties, privacy, and values

- Ensure that technical standards minimize vulnerability to attacks from malicious actors and reflect Federal priorities for innovation, public trust, and public confidence in systems that use AI technologies; and develop international standards to promote and protect those priorities
- Train the next generation of American AI researchers and users through apprenticeships; skills programs; and education in science, technology, engineering, and mathematics (STEM), with an emphasis on computer science, to ensure that American workers, including Federal workers, are capable of taking full advantage of the opportunities of AI
- Develop and implement an action plan to protect the advantage of the United States in AI and technology critical to United States economic and national security interests against strategic competitors and foreign adversaries

The Initiative was to be coordinated through the National Science and Technology Council (NSTC) Select Committee on Artificial Intelligence (Select Committee); any actions were to be implemented by agencies that conduct foundational AI R&D, develop and deploy applications of AI technologies, provide educational grants, and regulate and provide guidance for applications of AI technologies

Heads of all agencies were to review their Federal data and models to identify opportunities to increase access and use by the greater non-Federal AI research community in a manner that benefits that community, while protecting safety, security, privacy, and confidentiality. In identifying data and models for consideration for increased public access, agencies were to consider issues such as:

- Privacy and civil liberty protections for individuals who may be affected by increased access and use, as well as confidentiality protections for individuals and other data providers; and
- Safety and security concerns, including those related to the association or compilation of data and models

Finally, Heads of implementing agencies that also provide educational grants were to, to the extent consistent with applicable law, consider AI as a priority area within existing Federal fellowship and service programs. Eligible programs for prioritization were to give preference to American citizens, to the extent permitted by law, and were to include:

- High school, undergraduate, and graduate fellowship; alternative education; and training programs;
- Programs to recognize and fund early-career university faculty who conduct AI R&D, including through Presidential awards and recognitions;
- Scholarship for service programs;
- Direct commissioning programs of the United States Armed Forces; and
- Programs that support the development of instructional programs and curricula that encourage the integration of AI technologies into courses in order to facilitate personalized and adaptive learning experiences for formal and informal education and training

d. **Executive Order 13960 Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government**

This EO was issued on December 3, 2020, and was intended to::

- Promote the innovation and use of AI, where appropriate, to improve Government operations and services in a manner that fosters public trust, builds confidence in AI, protects our Nation’s values, and remains consistent with all applicable laws, including those related to privacy, civil rights, and civil liberties
- Ensure that responsible agencies shall, when considering the design, development, acquisition, and use of AI in Government, be guided by the common set of Principles which are designed to foster public trust and confidence in the use of AI, protect our Nation’s values, and ensure that the use of AI remains consistent with all applicable laws, including those related to privacy, civil rights, and civil liberties

The main objective of this EO was that when designing, developing, acquiring, and using AI in the Federal Government, agencies shall adhere to the following Principles:

- Lawful and respectful of our Nation's values (including those addressing privacy, civil rights, and civil liberties)
- Purposeful and performance-driven (where the benefits of developing AI significantly outweigh the risks, and the risks can be assessed and managed)
- Accurate, reliable, and effective (the application of AI is consistent with the use cases for which that AI was trained, and such use is accurate, reliable, and effective)
- Safe, secure, and resilient
- Understandable (the operations and outcomes of their AI applications are sufficiently understandable by subject matter experts, users, and others, as appropriate)
- Responsible and traceable (that human roles and responsibilities are clearly defined, understood, and appropriately assigned for the design, development, acquisition, and use of AI)
- Regularly monitored (AI applications are regularly tested against the Principles of this EO)
- Transparent (Agencies shall be transparent in disclosing relevant information regarding their use of AI to appropriate stakeholders)
- Accountable (Agencies shall be accountable for implementing and enforcing appropriate safeguards for the proper use and functioning of their applications of AI)

In terms of the scope of this EO:

- The Principles and implementation guidance in this order shall apply to AI designed, developed, acquired, or used specifically to advance the execution of agencies' missions, enhance decision making, or provide the public with a specified benefit
- This order applies to both existing and new uses of AI; both standalone AI and AI embedded within other systems or applications; AI developed both by the agency or by third parties on behalf of agencies for the fulfilment of specific agency missions, including relevant data inputs used to train AI and outputs used in support of decision making; and agencies' procurement of AI applications
- This order does not apply to:
    - AI used in defense or national security systems, in whole or in part, although agencies shall adhere to other applicable guidelines and principles for defense and national security purposes, such as those adopted by the Department of Defense and the Office of the Director of National Intelligence;
    - AI embedded within common commercial products, such as word processors or map navigation systems, while noting that Government use of such products must nevertheless comply with applicable law and policy to assure the protection of safety, security, privacy, civil rights, civil liberties, and American values; and
    - AI research and development (R&D) activities, although the Principles and OMB implementation guidance should inform any R&D directed at potential future applications of AI in the Federal Government

5. Al then presented the second special topic on the NIST AI Risk Management Framework. https://ftp.pwg.org/pub/pwg/ids/Presentation/NIST AI Risk Management Framework v2.pdf This topic had actually been previously discussed at the January 26, 2023 IDS WG Meeting, but that discussion was based on a draft version of the NIST AI Risk Management Framework.

   Al decided to discuss the framework again because the framework had been published in January 2023 as NIST AI 100-1 NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0). Al wanted to see what had been changed between the draft version and the published version of the framework. For the purposes of the rest of these minutes, only the changes between the draft and published versions will be discussed; see the slides in the link above for the full discussion of the NIST AI Risk Management Framework.

- The introductory information in the Framework document was significantly changed in the published version. The general purposes of the Framework, as stated in the published version are that it:
  - Offers a resource to the organizations designing, developing, deploying, or using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems.
  - Is intended to be:
    - ***Voluntary***, rights-preserving, non-sector-specific, and use-case agnostic, providing flexibility to organizations of all sizes and in all sectors and throughout society to implement the approaches in the Framework
    - Practical, to adapt to the AI landscape as AI technologies continue to develop, and to be operationalized by organizations in varying degrees and capacities so society can benefit from AI while also being protected from its potential harms
    - Flexible and to augment existing risk practices which should align with applicable laws, regulations, and norms
  - Is designed to equip organizations and individuals – referred to here as *AI actors* – with approaches that increase the trustworthiness of AI systems, and to help foster the responsible design, development, deployment, and use of AI systems over time
  - Offers approaches to minimize anticipated negative impacts of AI systems *and* identify opportunities to maximize positive impacts
  - Designed to address new risks as they emerge
- The attributes that the Framework were developed around were:
  - Be risk-based, resource-efficient, pro-innovation, and voluntary
  - Be Consensus-driven and developed and regularly updated through an open, transparent process
  - Uses clear and plain language that is understandable by a broad audience, including senior executives, government officials, non-governmental organization leadership, and those who are not AI professionals – while still of sufficient technical depth to be useful to practitioners
  - Provide common language and understanding to manage AI risks
  - Be easily usable and fit well with other aspects of risk management
  - Be useful to a wide range of perspectives, sectors, and technology domains
  - Be outcome-focused and non-prescriptive
  - Take advantage of and foster greater awareness of existing standards, guidelines, best practices, methodologies, and tools for managing AI risks – as well as illustrate the need for additional, improved resources
  - Be law- and regulation-agnostic
  - Be a living document
- Some key definitions that were extracted from the published version of the Framework are:
  - **AI actors**: Those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI (OECD[1] (2019) Artificial Intelligence in Society—OECD iLibrary)
  - **Artificial Intelligence (AI) System**: An engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying

[1] OECD - Organisation for Economic Co-operation and Development

levels of autonomy (Adapted from: OECD Recommendation on AI:2019; ISO/IEC 22989:2022)

- **Risk**: The composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats (Adapted from: ISO 31000:2018)
- **Risk Management**: Coordinated activities to direct and control an organization with regard to risk (Source: ISO 31000:2018)
- **Risk Tolerance**: The organization's or stakeholder's readiness or appetite to bear the risk in order to achieve its objectives
- **Accuracy**: *C*loseness of results of observations, computations, or estimates to the true values or the values accepted as being true (ISO/IEC TS 5723:2022)
- **Robustness or generalizability**: Ability of an AI system to maintain its level of performance under a variety of circumstances
- **Social responsibility**: The organization's responsibility "for the impacts of its decisions and activities on society and the environment through transparent and ethical behavior" (ISO 26000:2010)
- **Sustainability**: The "state of the global system, including environmental, social, and economic aspects, in which the needs of the present are met without compromising the ability of future generations to meet their own needs" (ISO/IEC TR 24368:2022)
- "**Professional Responsibility**": An approach that "aims to ensure that professionals who design, develop, or deploy AI systems and applications or AI-based products or systems, recognize their unique position to exert influence on people, society, and the future of AI" (ISO/IEC TR 24368:2022)
- **TEVV**: Test, Evaluation, Verification and Validation
- **Validation**: The "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled" (Source: ISO 9000:2015)
- **Reliability**: The "ability of an item to perform as required, without failure, for a given time interval, under given conditions" (Source: ISO/IEC TS 5723:2022)

AI indicated the term 'TEVV" is used a lot in the Framework; also, to pay attention to the term 'AI Actor" because it replaced the term "Stakeholder" in the draft version. Finally, the definition of AI System is an interesting one; not sure how it matches up with other definitions of this term.

- There was an interesting new chart on Slide 6 of the presentation that shows an example of the types of harm that AI can cause to people, organizations and the ecosystem. For example, harms to a person's civil liberties, rights or physical safety; harm to an organization via security breaches or to its reputation; and harm to the environment and the planet.
- Another new diagram on Slide 8 shows the Life Cycle and Key Dimensions of an AI System. The AI Key Dimensions – Data and Input, AI Model, Task and Output, Application Context, and People and Planet – form the inner two rings while the AI Life Cycle Phases – Plan and Design, Collect and process data, Build and use model, Deploy and use, and Operate and Monitor – form the outer ring.
- A companion new chart to the above diagram on Slide 9 shows the AI Actors associated with the appropriate AI Life Cycle phase and AI Key Dimension. For example, the Data and Input Key Dimension would be associated with the Collect and process data Life Cycle phase, and would involve AI Actors such as data scientists, data engineers and domain expert doing activities such as gathering and validating data.
- The concept of "Trustworthiness" is central to the AI Risk Management Framework. Trustworthiness AI is defined as "**valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and *Fair with Harmful***

Bias Managed (Note: the last attribute was changed in the published version; it was "**fair and bias is managed** in the draft version). Note that the definitions of these 7 "Characteristics of Trustworthiness" shown in Slide 11 did not change in the published version from the draft version.

- The published version provided the following set of benefits of the Framework to users:
  - Enhanced processes for governing, mapping, measuring, and managing AI risk, and clearly documenting outcomes;
  - Improved awareness of the relationships and tradeoffs among trustworthiness characteristics, socio-technical approaches, and AI risks;
  - Explicit processes for making go/no-go system commissioning and deployment decisions;
  - Established policies, processes, practices, and procedures for improving organizational accountability efforts related to AI system risks;
  - Enhanced organizational culture which prioritizes the identification and management of AI system risks and potential impacts to individuals, communities, organizations, and society;
  - Better information sharing within and across organizations about risks, decision-making processes, responsibilities, common pitfalls, TEVV practices, and approaches for continuous improvement;
  - Greater contextual knowledge for increased awareness of downstream risks;
  - Strengthened engagement with interested parties and relevant AI actors; and
  - Augmented capacity for TEVV of AI systems and associated risks
- The core steps (now denoted as 'Functions' in the published version) remain the same:
  - **Govern**: Cultivate and implement a culture of risk management within organizations developing, deploying, or acquiring AI systems
  - **Map**: Establish the context to frame risks related to an AI system
  - **Measure**: Employ quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts
  - **Manage**: Entails allocating risk management resources to mapped and measured risks on a regular basis and as defined by the Govern function

Note: The rest of the discussion of the Framework will be discussing the various categories and subcategories of each core function, but only discussion those that have changed from the draft version.

a. **Govern** function

The **Govern** function provides the following:

- Cultivates and implements a culture of risk management within organizations designing, developing, deploying, evaluating, or acquiring AI systems;
- Outlines processes, documents, and organizational schemes that anticipate, identify, and manage the risks a system can pose, including to users and others across society – and procedures to achieve those outcomes;
- Incorporates processes to assess potential impacts;
- Provides a structure by which AI risk management functions can align with organizational principles, policies, and strategic priorities;
- Connects technical aspects of AI system design and development to organizational values and principles, and enables organizational practices and competencies for the individuals involved in acquiring, training, deploying, and monitoring such systems; and
- Addresses full product lifecycle and associated processes, including legal and other issues concerning use of third-party software or hardware systems and data

Categories/Subcategories

- **GOVERN 1: Policies, processes, procedures, and practices across the organization related to the mapping, measuring, and managing of AI risks are in place, transparent, and implemented effectively**

  Added the following new Subcategories in the published version:

  - GOVERN 1.3: Processes, procedures, and practices are in place to determine the needed level of risk management activities based on the organization's risk tolerance.
  - GOVERN 1.6: Mechanisms are in place to inventory AI systems and are resourced according to organizational risk priorities.
  - GOVERN 1.7: Processes and procedures are in place for decommissioning and phasing out AI systems safely and in a manner that does not increase risks or decrease the organization's trustworthiness.

  Renumbered the old GOVERN 1.3 and GOVERN 1.4 in the published version to GOVERN 1.4 and GOVERN 1,5, respectively

  Modified the new GOVERN 1.4 to read
  GOVERN 1.4 The risk management process and its outcomes are established through transparent policies procedures, and other controls based on organizational risk priorities.

- **GOVERN 2: Accountability structures are in place so that the appropriate teams and individuals are empowered, responsible, and trained for mapping, measuring, and managing AI risks**

  Added the following new Subcategory in the published version - GOVERN 2.3**:** Executive leadership of the organization takes responsibility for decisions about risks associated with AI system development and deployment.

- **GOVERN 3: Workforce diversity, equity, inclusion, and accessibility processes are prioritized in the mapping, measuring, and managing of AI risks throughout the lifecycle**

  Revised Subcategory Govern 3.1 in the published version to read:
  GOVERN 3.1: Decision-making related to mapping, measuring, and managing AI risks throughout the lifecycle is informed by a diverse team (e.g., diversity of demographics, disciplines, experience, expertise, and backgrounds).

  Added the new Subcategory in the published version:
  GOVERN 3.2: Policies and procedures are in place to define and differentiate roles and responsibilities for human-AI configurations and oversight of AI systems.

- **GOVERN 4: Organizational teams are committed to a culture that considers and communicates risk.**

  Made minor wording changes in the three subcategories GOVERN 4.1, GOVERN 4.2 and GOVERN 4.3 such as adding 'AI' where appropriate.

- **GOVERN 5: Processes are in place for robust stakeholder engagement.**

  Modified Subcategory GOVERN 5.1 in the published version to read:
  GOVERN 5.1: Organizational policies and practices are in place to collect, consider, prioritize, and integrate feedback from those external to the team that developed or deployed the AI system regarding the potential individual and societal impacts related to AI risks.

  Modified Subcategory GOVERN 5.2 in the published version to read:
  GOVERN 5.2: Mechanisms are established to enable AI actors the team that developed or deployed AI systems to regularly incorporate adjudicated feedback from relevant AI actors into system design and implementation.

  Note: A common change made throughout the Framework was to remove the term "stakeholder" and replace it with AI Actor or some related term.

- **GOVERN 6: Policies and procedures are in place to address AI risks arising from third-party software and data and other supply chain issues.**

  Modified Subcategory GOVERN 6.1 in the published version to read:
  GOVERN 6.1: Policies and procedures are in place that address AI risks associated with third-party entities, including risks of infringement of a third-party's intellectual property or other rights.

b. **Map** Function

The **Map** Function helps organizations proactively prevent negative risks and develop more trustworthy AI systems by:

- Improving their capacity for understanding contexts;
- Checking their assumptions about context of use;
- Enabling recognition of when systems are not functional within or out of their intended context;
- Identifying positive and beneficial uses of their existing AI systems;
- Improving understanding of limitations in AI and ML processes;
- Identifying constraints in real-world applications that may lead to negative impacts;
- Identifying known and foreseeable negative impacts related to intended use of AI systems; and
- Anticipating risks of the use of AI systems beyond intended use

Categories/Subcategories

- **MAP 1: Context is established and understood.**

  Subcategory MAP 1.1 was modified in the published version to read:
  MAP 1.1: Intended purpose, prospective settings in which the AI system will be deployed, the specific set or types of users along with their expectations, and impacts of system use are understood and documented. Considerations include: the specific set or types of users along with their expectations; potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet; assumptions and related limitations about AI system purposes, uses, and risks across the development or product AI lifecycle; and related TEVV and system metrics.

  Subcategories MAP 1.3 and MAP 1.4 for some reason switched designations in the published version, so MAP 1.3 became MAP 1.4 and MAP 1.4 became MAP 1.3. However, the wording of both subcategories didn't change.

  The following Subcategory MAP 1.6 in the draft version was removed in the published version:
  MAP 1.6: Practices and personnel for design activities enable regular engagement with stakeholders, and integrate actionable user and community feedback about unanticipated negative impacts.

  MAP 1.7 in the draft version was changed to MAP 1.6 in the published version, but the wording remained the same.

- **MAP 2: Categorization of the AI system is performed.**

  Subcategory MAP 2.2 was modified in the published version to read:
  MAP 2.2: Information about the AI system's knowledge limits and how system output may be utilized and overseen by humans is documented. Documentation provides sufficient information to assist relevant AI actors when making decisions and taking subsequent actions.

- **MAP 3: AI capabilities, targeted usage, goals, and expected benefits and costs compared with the status quo are understood.**

  Subcategory MAP 3.2 was modified in the published version to read:
  MAP 3.2: Potential costs, including non-monetary costs, which result from expected or realized AI errors or system performance and trustworthiness – as connected to organizational risk tolerance – are examined and documented.

  Subcategory MAP 3.3 was modified in the published version to read:
  MAP 3.3: Targeted application scope is specified, and documented based on the system's capability, established context and AI system classification categorization.

  Added the following new subcategories in the published version:
  - MAP 3.4: Processes for operator and practitioner proficiency with AI system performance and trustworthiness – and relevant technical standards and certifications – are defined, assessed, and documented.
  - MAP 3.5: Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from the **GOVERN** function.

- **MAP 4: Risks and benefits are mapped for *all components of the AI system including* third-party software and data.**

  Subcategory MAP 4.1 was modified in the published version to read:
  MAP 4.1: Approaches for mapping AI and legal risks of its components – including the use of third-party data or software – are in place and documented, as are risks of infringement of a third party's intellectual property or other rights.

  Subcategory MAP 4.2 was modified in the published version to read:
  MAP 4.2: Internal risk controls for components of the AI system, including, third-party AI technologies risks are in place identified and documented.

- **MAP 5: Impacts to individuals, groups, communities, organizations, and society are characterized.**

  The following Subcategories in the draft version were removed in the published version:
  - MAP 5.1: Potential positive and negative impacts to individuals, groups, communities, organizations, and society are regularly identified and documented.
  - MAP 5.3: Assessments of benefits versus impacts are based on analyses of impact, magnitude, and likelihood of risk.

  Subcategory MAP 5.2 in the draft version was renamed as MAP 5.1 in the published version and modified to read:
  MAP 5.1: Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, stakeholder feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.

  A new Subcategory was added in the published version:
  MAP 5.2: Practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, and unanticipated impacts are in place and documented.

c. **Measure** Function

The **Measure** Function:

- Employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts
- Uses knowledge relevant to AI risks identified in the **MAP** function and informs the **MANAGE** function

- Includes tracking metrics for trustworthy characteristics, social impact, and human-AI configurations
- Should include rigorous software testing and performance assessment methodologies with associated measures of uncertainty, comparisons to performance benchmarks, and formalized reporting and documentation of results

Categories/Subcategories

- **MEASURE 1: Appropriate methods and metrics are identified and applied.**

  Subcategory MEASURE 1.1 was modified in the published version to read:
  MEASURE 1.1: Approaches and metrics for quantitative or qualitative measurement of AI risks enumerated during the Map function, are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot - be measured are properly documented.

  Subcategory MEASURE 1.2 clarified in the published version that the metrics being referred to in the Subcategory were AI metrics.

  Subcategory MEASURE 1.3 was modified in the published version to read:
  MEASURE 1.3: Internal experts who did not serve as front-line developers for the system and/or independent assessors are involved in regular assessments and updates. Domain experts, users, AI actors external to the team that developed or deployed the AI system, and affected communities are consulted in support of assessments as necessary per organizational risk tolerance.

- **MEASURE 2: Systems are evaluated for trustworthy characteristics.**

  Subcategory MEASURE 2.2 was modified in the published version to read:
  MEASURE 2.2: Evaluations involving human subjects meet applicable requirements (including human subject protection); and are representative of the intended relevant population. .

  Subcategory MEASURE 2.3 had a minor change in the published version to indicate it was for an AI system.

  Subcategories MEASURE 2.4, MEASURE 2.5, MEASURE 2.8, MEASURE 2.9 and MEASURE2.10 in the draft version were changed in the published version to be MEASURE 2.5, MEASURE 2.6, MEASURE 2.9, MEASURE 2.10 and MEASURE 2.12, respectively.

  The new Subcategory MEASURE 2.6 (formerly Subcategory MEASURE 2.5) was modified in the published version to read:
  MEASURE 2.6: AI system is evaluated regularly for safety. The AI system to be deployed product is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and it can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics reflect implicate system reliability and robustness, real-time monitoring, and response times for AI system failures.

  Subcategory MEASURE 2.7 was modified in the published version to read:
  MEASURE 2.7: AI system resilience and security – as identified in the **MAP** function – is evaluated regularly and documented.

  The new Subcategory MEASURE 2.9 (formerly Subcategory MEASURE 2.8) was modified in the published version to read:
  MEASURE 2.9: AI model is explained, validated, and documented, and AI system output is interpreted within its context – as identified in the **MAP** function – and to inform responsible use and governance.

  The new Subcategory MEASURE 2.10 (formerly Subcategory MEASURE 2.9) was modified in the published version to read:
  MEASURE 2.10: Privacy risk of the AI system – as identified in the **MAP** function – is examined regularly and documented.

The new Subcategory MEASURE 2.12 (formerly Subcategory MEASURE 2.10) was modified in the published version to read:
MEASURE 2.12: Environmental impact and sustainability of AI model training and management activities – as identified in the **MAP** function – are assessed and documented.

The following new Subcategories were added in the public version:

- *MEASURE 2.4: The functionality and behavior of the AI system and its components – as identified in the **MAP** function – are monitored when in production.*
- *MEASURE 2.8: Risks associated with transparency and accountability – as identified in the **MAP** function – are examined and documented.*
- *MEASURE 2.11: Fairness and bias – as identified in the **MAP** function – are evaluated and results are documented.*
- *MEASURE 2.13: Effectiveness of the employed TEVV metrics and processes in the **MEASURE** function are evaluated and documented.*

- **MEASURE 3: Mechanisms for tracking identified *AI* risks over time are in place.**

  In Subcategory MEASURE 3.1, indicated in the published version that 'unanticipated' AI risks should also be regularly identified and tracked.

  Added the following new Subcategory in the published version:
  MEASURE 3.3: Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics.

- **MEASURE 4: Feedback about efficacy of measurement is gathered and assessed.**

  In Subcategory MEASURE 4.1, it was indicated in the published version that this Subcategory was referring to AI risks.

  Subcategory MEASURE 4.2 was modified in the published version to read:
  MEASURE 4.2: Measurement results regarding AI system trustworthiness in deployment context(s) and across AI lifecycle are informed by input from domain experts and relevant AI actors to validate whether the system is performing consistently as intended. Results are documented.

  Subcategory MEASURE 4.3 was modified in the published version to read:
  MEASURE 4.3: Measurable performance improvements or declines based on consultations with relevant AI actors, including affected communities, and field data about context-relevant risks and trustworthiness characteristics are identified and documented.

d. **Manage** Function

The **Manage** Function:

- Entails allocating risk resources to mapped and measured risks on a regular basis and as defined by the **GOVERN** function
- Risk treatment comprises plans to respond to, recover from, and communicate about incidents or events
- Contextual information gleaned from expert consultation and input from relevant AI actors – established in **GOVERN** and carried out in **MAP** – is utilized in this function to decrease the likelihood of system failures and negative impacts
- Systematic documentation practices established in **GOVERN** and utilized in **MAP** and **MEASURE** bolster AI risk management efforts and increase transparency and accountability
- Processes for assessing emergent risks are in place, along with mechanisms for continual improvement

Categories/Subcategories

- **MANAGE 1: AI risks based on impact assessments and other analytical output from the Map and Measure functions are prioritized, responded to, and managed.**

  Subcategory MANAGE 1.1 was modified in the published version to read:
  MANAGE 1.1: A determination is made about as to whether the AI system achieves its intended purposes and stated objectives and whether its development or deployment should proceed.

  Subcategory MANAGE 1.2 was modified in the published version to clarify that it was referring to AI risks.

  Subcategory MANAGE 1.3 was modified in the published version to read:
  MANAGE 1.3: Responses to the most significant AI risks deemed high priority, as identified by the Map function, are developed, planned, and documented. Risk response options can include mitigating, transferring, sharing, avoiding, or accepting.

  The following new Subcategory was added in the published version:
  MANAGE 1.4: Negative residual risks (defined as the sum of all unmitigated risks) to both downstream acquirers of AI systems and end users are documented. .

- **MANAGE 2: Strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, and documented, and informed by input from relevant AI actors.**

  Subcategory MANAGE 2.1 was modified in the published version to read:
  MANAGE 2.1: Resources required to manage AI risks are taken into account, - along with viable non-AI alternative systems, approaches, or methods, - to reduce the magnitude or likelihood of each potential actions.

  Subcategory 2.3 in the draft version was changed to Subcategory 2.4 in the published version and modified to read:
  MANAGE 2.4: Mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.

  The following new Subcategory was added in the published version:
  MANAGE 2.3: Procedures are followed to respond to and recover from a previously unknown risk when it is identified.

- **MANAGE 3: AI Risks from third-party entities are managed.**

  Subcategory MANAGE 3.1 was modified in the published version to read:
  MANAGE 3.1: AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.

  The following new Subcategory was added in the published version:
  MANAGE 3.2: Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance**.**

- **MANAGE 4: Risk treatments, including response and recovery, and communication plans for the identified and measured AI risks are documented and monitored regularly.**

  Subcategory MANAGE 4.1 was modified in the published version to read:
  MANAGE 4.1: Post-deployment system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management.

  Subcategory MANAGE 4.2 was modified in the published version to read:
  MANAGE 4.2: Measurable activities for continual improvement are integrated into AI system updates and include regular engagement with interested parties, including relevant AI actors.

- Finally, the published version included a new section on AI-specific risks. Al noted that thee risks, which are listed below, are all technical-related risks and don't address the socially-related risks that are of the greatest concern today.
    - The AI risks listed in the Framework are:
    - The data used for building an AI system may not be a true or appropriate representation of the context or intended use of the AI system, and the ground truth may either not exist or not be available
    - Harmful bias and other data quality issues can affect AI system trustworthiness, which could lead to negative impacts
    - AI system dependency and reliance on data for training tasks, combined with increased volume and complexity typically associated with such data
    - Intentional or unintentional changes during training may fundamentally alter AI system performance
    - Datasets used to train AI systems may become detached from their original and intended context or may become stale or outdated relative to deployment context
    - AI system scale and complexity (many systems contain billions or even trillions of decision points) housed within more traditional software applications
    - Use of pre-trained models that can advance research and improve performance can also increase levels of statistical uncertainty and cause issues with bias management, scientific validity, and reproducibility.
    - Higher degree of difficulty in predicting failure modes for emergent properties of large-scale pre-trained models
    - Privacy risk due to enhanced data aggregation capability for AI systems
    - AI systems may require more frequent maintenance and triggers for conducting corrective maintenance due to data, model, or concept drift
    - Increased opacity and concerns about reproducibility
    - Underdeveloped software testing standards and inability to document AI-based practices to the standard expected of traditionally engineered software for all but the simplest of cases.
    - Difficulty in performing regular AI-based software testing, or determining what to test, since AI systems are not subject to the same controls as traditional code development
    - Computational costs for developing AI systems and their impact on the environment and planet
    - Inability to predict or detect the side effects of AI-based systems beyond statistical measures

6. **Actions:** None

Next Steps

The next IDS WG Meeting will be July 13, 2023 at 3:00P ET / 12:00N PT, now that we are back on our normal cycle with IPP. Main topics will be the latest status of the HCD iTC and HIT and likely a special topic on a TBD topic (possibly the new Specification of Functional Requirements for Cryptography.that will be required on all cPPs)