TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management

This Joint Roadmap aims to guide the development of tools, methodologies, and approaches to AI risk management and trustworthy AI by the EU and the United States and to advance our shared interest in supporting international standardization efforts and promoting trustworthy AI on the basis of a shared dedication to democratic values and human rights. The roadmap takes practical steps to advance trustworthy AI and uphold our shared commitment to the Organisation for Economic Co-operation and Development (OECD) Recommendation on AI.

Tangible global leadership by the United States and the European Union can provide scalable, science-based methods to advance trustworthy approaches to AI that serve all people in responsible, equitable, and beneficial ways. Effective risk management and assessment can help earn and increase trust in the development, deployment, and use of AI systems. Recognizing the power of AI to address the world's challenges, we also acknowledge AI systems entail risk. By minimizing the negative impacts of AI systems on individuals, culture, the economy, societies, and the planet, we can maximize the positive impacts and benefits of AI systems that support the shared values underpinning like-minded democracies. Towards that goal, the U.S.-EU Joint Statement of the Trade and Technology Council (May 2022) expressed an intention to develop a joint roadmap ("Joint Roadmap") on evaluation and measurement tools for trustworthy AI and risk management.

Contents

Trade and Technology Council (TTC)

Stakeholder(s):

United States

European Union

Industry

Academia

Civil Society

Vision

Trustworthy approaches to AI

Mission

To guide the development of tools, methodologies, and approaches to AI risk management and trustworthy AI by the EU and the United States

Values

Trust:

The United States and EU acknowledge that a risk-based approach and a focus on trustworthy AI systems can provide people with confidence in AI-based solutions, while inspiring enterprises to develop trustworthy AI technologies. This approach supports common values, protects the rights and dignity of people, sustains the planet, and encourages market innovation. Both parties are pursuing risk-based approaches that operationalize these values.

Dignity

Sustainability

Innovation

Science:

This Joint Roadmap underscores the importance of the EU and United States approaches being supported by science, international standards, shared terminology, and validated metrics and methodologies. It suggests activities which are intended to be compatible with the respective regulatory, policy, and legislative initiatives of the two sides.

Standards

Metrics

Engagement:

The active engagement and participation of stakeholders throughout the whole AI community (including industry, academia, and civil society) is key to fulfilling the objectives of this roadmap. In this respect, all activities are intended to be conducted with engagement and support by stakeholders and experts via consultation plans, including expert workshops.

Strategic Alignment:

Roadmap suggestions for concrete activities aimed at aligning EU and United States risk-based approaches are advancing: 1) shared terminologies and taxonomies; 2) leadership and cooperation in international technical standards development activities and analysis and collection of tools for trustworthy AI and risk management; and 3) monitoring and measuring existing and emerging AI risks.

Democracy

Human Rights

Safety

Security

Fairness

Nondiscrimination

Interoperability

Transparency

Diversity

Compatibility

Openness

Impartiality

Inclusiveness

1. Terminologies & Taxonomies

Advance shared terminologies and taxonomies

Stakeholder(s)

Organization for Standardization (ISO)

NIST

OECD

European Standardisation Organisations

Institute of Electrical and Electronics Engineers (IEEE)

Shared terminologies and taxonomies are essential for operationalizing trustworthy AI and risk management in an interoperable fashion. The activities in this section support the EU's and United States' work on interoperable definitions of key terms such as trustworthy, risk, harm, risk threshold, and socio-technical characteristics such as bias, robustness, safety, interpretability, and security. Developing a shared understanding of basic terms will offer an interoperable taxonomy when developing standards and identifying responsibilities, practices, and policies.

This work will leverage the global work already done and ongoing (such as within the International Organization for Standardization [ISO], OECD, and Institute of Electrical and Electronics Engineers [IEEE]). It will consider related work by the United States (such as the NIST AI Risk Management Framework and the Blueprint for an AI Bill of Rights) and the EU (such as the EU AI Act, HLEG, and European Standardisation Organisations). The EU and United States affirm the importance of a shared understanding and consistent application of concepts and terminology that include, but are not limited to - risk, risk management, risk tolerances, risk perception, and the socio-technical characteristics of trustworthy AI.

This work could be informed by:

- Alignment with international standards development organizations

- Ongoing efforts within OECD Working Party on AI Governance (AIGO) and OECD Network of AI Experts (ONE. AI)

- NIST's efforts in developing an AI Risk Management Framework and its related guides and tools - The National AI Initiative Act and Blueprint for an AI Bill of Rights - The EU AI Act

- Work developed by the European standards organizations.^

- The deliverables of the EU High-Level Expert Group, such as the ALTAI Assessment List for Trustworthy AI

2. Tools & Standards

Foster leadership and cooperation in international technical standards development activities and analysis and collection of tools for trustworthy AI and risk management

The EU and United States affirm that AI technologies should be shaped by our shared democratic values and commitment to protecting and respecting human rights. Leadership in standards for AI and emerging technologies should promote safety, security, fairness, nondiscrimination, interoperability, innovation, transparency, diverse markets, compatibility, and inclusiveness. Both sides are committed to supporting multi-stakeholder approaches to standards development, and recognize the importance of procedures that advance transparency, openness, fair processes, impartiality, and inclusiveness.

2.1. Standards

Engage with stakeholders to identify standards that are of mutual interest

International technical standards Standards shape the design, development and use of technologies that underpin our economies, cultures, and societies. Technologies provide opportunities for positive impact. They can also cause cascading negative consequences without proper safeguards.

AI standards that articulate requirements, specifications, test methodologies, or guidelines relating to trustworthy characteristics can help ensure that AI technologies and systems meet critical objectives (e.g., functionality, interoperability) and performance characteristics (e.g., accuracy, reliability, and safety). In contrast, standards that are not fit for purpose, not yet available, not broadly accessible (notably to start-ups and small and medium-sized enterprises), or not designed around valid technological solutions may hamper innovation and the timely development and deployment of trustworthy AI technologies.

Global leadership, participation, and cooperation on international AI standards will be critical for consistent "rules of the road" that enable market competition, preclude barriers to trade, and allow innovation to flourish. This may enable governments to align with an international approach when developing internal policies for safeguarding and advancing respect for human rights and democratic values.

As like-minded partners, the EU and United States seek to support and provide leadership in international standardization efforts. This can be achieved by contributing and cooperating on technical AI standards development, currently underway in international standards organizations. These standards impact the design, operation, and evaluation and measurement of trustworthy AI and risk management.

Without prejudice to the specificities and needs of their respective legal systems, the EU and United States aim to act as a model for others by adhering to the WTO TBT principles. This includes support and use of international standards, as appropriate, as the basis for technical regulations, conformity assessments and regional standards. At the same time, the EU and United States, working with our respective stakeholders and mechanisms, aim to identify critical gaps in existing international AI standards development activities. The EU and the United States can cooperate on AI pre-standardization research and development (R&D) to advance the technical and scientific foundation for international standards development.

The EU and United States intend to actively promote the participation of a wide range of stakeholders – including their standards experts, impacted communities, domain experts, and other cross-disciplinary experts – in ongoing AI standards development work. Both sides plan to promote continual expert-level information sharing to improve understanding of the respective approaches and possible uptake of common technical solutions. The EU and United States governments can play a convening role with their respective stakeholders to promote appropriate representation at important standards-setting bodies and organizations. Furthermore, both sides intend to promote the development and voluntary use of international AI standards that are established in an open and transparent manner and that are technically sound, performance-based, and suitable for public and private sector use. Both sides also plan to support the consideration of small and medium-sized enterprises and start-up communities in standards development activities.

In the short term, this activity will involve engaging with stakeholders to identify standards that are of mutual interest, starting with AI trustworthiness, bias, and risk management.

2.2. Tools

Develop tools for trustworthy AI and risk management

Tools for trustworthy AI and risk management ~ Regardless of respective policy landscapes, technical tools are needed to map, measure, manage, and govern AI risks. Tools – defined by OECD as instruments and structured methods (of either a technical, procedural, or educational nature) that can be leveraged by relevant stakeholders to make their AI applications trustworthy – should be built upon strong scientific foundations and aligned with standards development activities. Objectives of the EU-U.S. joint work on tools for trustworthy AI and risk management are as follows:

2.2.1. Metrics & Methodologies

Build a knowledge base of metrics and methodologies for measuring AI trustworthiness, risk management methods, and related tools

Shared hub/repository of metrics and methodologies ~ The EU and United States intend to work together to build a common knowledge base of metrics and methodologies for measuring AI trustworthiness, risk management methods, and related tools. The latter could include, for example, the measurement of AI's positive and negative environmental implications. Building on the common work related to terminology, this effort involves developing selection criteria for inclusion of metrics in the shared hub/repository. The knowledge base would be openly and publicly accessible online and could augment the ongoing OECD efforts in the area. The selection and inclusion of metrics and tools supports a useful repository for the two parties but does not constrain or prejudge the regulatory activities of the two parties.

2.2.2. Analyses

Study and characterize the landscape of standards and tools for trustworthy AI

Analysis of tools for trustworthy AI ~ The EU and United States expect to support studies to characterize the landscape of existing sector- or application-agnostic and sector- or application-specific standards and tools for trustworthy AI developed by standards development organizations, industry (including start-ups and small and medium-sized enterprises), open-source developers, academia, civil society organizations, governments, and other stakeholders. The results of these studies could inform and support AI standards development efforts. These studies could identify commonalities in approaches that operationalize shared values and frameworks as well as gaps in existing methodologies as they relate to our shared values. Collectively, these studies can support interoperable risk management strategies, evaluation, and measurement tools. As trustworthy AI tools begin to be deployed more widely and aligned with AI standards, the learnings from this activity would both inform standards development and shape AI standards.

3. Risks

Monitor and measure existing and emerging AI risks

The EU and United States intend to develop knowledge-sharing mechanisms on cutting-edge scientific research in AI and its related risks, which have the potential to significantly impact trade and technology.

Both parties intend to take actionable steps towards:

3.1. Tracking

Develop a tracker of risks and risk categories based on context, use cases, and empirical data on AI incidents, impacts, and harms

A tracker of existing and emergent risks and risk categories based on context, use cases, and empirical data on AI incidents, impacts, and harms. A values-based understanding of existing risks serves as a baseline for detecting and analyzing both existing and emergent risks. This activity seeks to provide a common ground for both parties to better define the origin of risks and their impact, and to better organize risk metrics and methodologies for risk avoidance or mitigation. The tracker would be continually extended or updated to include new risks emerging from the dynamics of development and use, improvements in understanding of the potential harms to shared values, compound risks due to the interaction of several systems, or unknown but predictable risks that could arise from new AI methods and/or contexts of use.

3.2. Tests & Evaluations

Conduct interoperable tests and evaluations of AI risks

Interoperable tests and evaluations of AI risks: Evaluations strengthen research communities, establish research methodology, support the development of standards, and facilitate technology transfer. Evaluations inform consumer choice and facilitate innovation through transparency of system functionality and trustworthiness and can be used for compliance tests. A significant challenge in the evaluation of trustworthy AI systems is that context of deployment matters. For example, accuracy measures alone do not provide enough information to determine if a system is acceptable to deploy. The accuracy measures must be evaluated based on the context within which the AI system operates and the associated harms and benefits that could occur. Other challenges include the quickly moving state of the art, the diversity of architectures of AI systems, and the complex behavior and emergent capabilities of large deep learning systems. New joint efforts in AI tests and evaluations are expected to focus on trustworthiness characteristics of system performance in addition to metrics such as accuracy.

Administrative Information

Start Date: 2022-12-01
End Date:

Publication Date: 2022-12-12
Source: https://www.nist.gov/system/files/documents/2022/12/04/Joint_TTC_Roadmap_Dec2022_Final.pdf

Submitter:

Given Name: Owen
Surname: Ambur
Email: Owen.Ambur@verizon.net
Phone:

PDF formatted using TopLeaf XML publisher
www.turnkey.com.au