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

December 1, 2022

1. Background

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.

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.

2. Risk-based approaches: Bringing EU and U.S. approaches closer

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.

Both sides apply risk-based approaches that consider the combination of societal and technical factors (socio-technical perspective) to advance trustworthy AI. EU examples are represented in the proposed EU AI Act and the work of the High-Level Expert Group (HLEG) on AI. United States examples can be seen in the National Institute of Standards and Technology (NIST) draft AI Risk Management Framework as well as the White House Office of Science and Technology Policy (OSTP) Blueprint for an AI Bill of Rights. (See Appendix for summaries of these examples.) While the EU and United States may have different views on regulatory approaches – including allocation of responsibility for risk assessment, possible legal responsibility for the

establishment of a risk management system, and the appropriate balance between regulatory and voluntary measures – the EU and United States risk-based approaches recognize that our shared values can guide the advancement of emerging technologies.

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. 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.

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.

3. Roadmap activities

3.1 Advance shared terminologies and taxonomies

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

3.2 EU-U.S. Leadership and cooperation on international technical standards and 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, non-discrimination, 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.

3.2.1. 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.

3.2.2. 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:

- 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.

- 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.

3.3 Monitoring and measuring 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:

- 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.

- 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.

4. Implementation plan

Advancing shared terminology and taxonomy provides an interoperable lexicon to communicate about risk and appropriate risk treatment, which in turn promotes interoperable measurements and evaluations of AI risks and impact. Jointly developing tools such as a shared repository of metrics likewise fosters transparency, interoperability, and uniformity of risk measurements. Collectively, such efforts improve effectiveness, transparency, and interoperability of risk assessment and risk management.

The objectives described in this joint roadmap can be achieved through several mechanisms including:

Short-term objectives:

- Establish inclusive cooperation channels:
  - Establish three (3) expert working groups on 1) AI terminology and taxonomy, 2) AI standards and tools for trustworthy AI and risk management, and 3) monitoring and measuring existing and emerging AI risks.
  - Develop work plans for each of the three expert groups.
  - Establish stakeholder and expert consultation plans, including expert workshops.

- Advancing shared terminologies and taxonomies:
  - Map terminology and taxonomy in key EU and United States documents and international standards that include, but are not limited to – risk, risk management, risk tolerances, risk perception, and the socio-technical characteristics of trustworthy AI.

- AI Standards:
  - Conduct a landscape analysis of international standards of interest to the EU or United States and evaluate the level of each parties' participation in and contribution to international standards development.
  - Identify international standards of interest for cooperation.
  - Promote participation of experts and relevant stakeholders in respective international standardization bodies.

- Development of tools:
  - Establish a process for tool selection, inclusion and revision.
  - Establish the criteria to evaluate tools for trustworthy AI.

- Monitoring and measuring existing and emerging AI risks:
  - Establish the objectives and methodology for tracking existing AI risks based on use cases and incidents reporting, which may be based on pilot attempts at categorization.
  - Identify the research methodology for tests and evaluations of emerging AI risks.

Long-term objectives:

- Establish inclusive cooperation channels to inform input to and leadership in international standards:
  - Conduct expert workshops.
  - Review and assess progress made and update the roadmap if needed.
  - Identify opportunities to cooperate and share roadmap outputs and learnings.

- Advancing shared terminologies and taxonomies:
  - Develop or revise shared understanding of terminology and taxonomy.

- AI Standards:
  - Organize possible cooperation in international standardization fora with respect to certain identified items.
  - Work with and support experts in the development or deployment of standards of mutual interest.

- Development of tools:
  - Identify metrics and methodologies to add to the shared hub/repository.
  - Update and maintain the shared hub/repository.

- Monitoring and measuring existing and emerging AI risks:
  - Create benchmarks and evaluations of AI risks that could be informed by empirical studies of AI incidents.
  - Conduct theoretically informed and analytical forecasting of emerging and future risks.

APPENDIX: EU and United States approaches to AI risk management

Examples of the US's risk-based approach to AI

NIST's draft AI Risk Management Framework (AI RMF)

The AI RMF is intended to address challenges unique to AI systems and encourage and equip different AI stakeholders to manage AI risks proactively and purposefully. The Framework describes a process for managing AI risks across a wide spectrum of types, applications, and maturity – regardless of sector, size, or level of familiarity with a specific type of technology. Cultivating trust by understanding and managing the risks of AI systems helps preserve civil liberties and rights and enhances safety while creating opportunities for innovation and realizing the full potential of this technology.

The AI RMF is a voluntary framework seeking to provide a flexible, structured, and measurable process to address AI risks prospectively and continuously throughout the AI lifecycle. It is intended to help organizations manage both enterprise and societal risks related to the design, development, deployment, evaluation, and use of AI systems through improved understanding, detection, and preemption. Using the AI RMF can assist organizations, industries, and society to understand and determine their acceptable levels of risk.

The AI RMF is not a compliance mechanism, nor is it a checklist intended to be used in isolation. It is law- and regulation-agnostic, as AI policy discussions are live and evolving. While risk management practices should incorporate and align to applicable laws and regulations, the NIST AI RMF is not intended to supersede existing regulations, laws, or other mandates; it should support organizations' abilities to operate under applicable domestic and international legal or regulatory regimes. Engagement with the broad AI community during development of the AI RMF informs AI research, development, and evaluation by NIST and others. The AI RMF is currently in its second draft and is expected to be released in early 2023.

NIST AI RMF employs the following definitions:
Note: additional considerations are underway to further align with international AI standards (including ISO/IEC 22989, ISO/IEC 23894 etc.)

- **Risk**: In the context of the AI RMF, 'risk' refers to the composite measure of an event's probability of occurring and the magnitude (or degree) of the consequences of the corresponding events. 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**: Risk management refers to coordinated activities to direct and control an organization with regard to risk (Source: ISO 31000:2018).
- **Risk tolerance:** Refers to the organization's or stakeholder's readiness to bear risks in order to achieve its objectives. Risk tolerance can be influenced by legal or regulatory requirements (Adapted from: ISO Guide 73).
- **Socio-technical characteristics of AI trustworthiness**: A system is considered trustworthy if it is valid and reliable, safe, fair and with managed bias, secure and

resilient, accountable and transparent, explainable and interpretable, and privacy-enhanced.

The Blueprint for an AI Bill of Rights

The Blueprint for an AI Bill of Rights is a set of five principles and associated practices to help guide the design, use, and deployment of automated systems to protect the rights of the American public in the age of artificial intelligence. Developed through extensive consultation with the public, these principles are a blueprint for building and deploying automated systems that are aligned with human rights and democratic values. The Blueprint for an AI Bill of Rights gives concrete steps that can be taken by many kinds of organizations—from governments at all levels to companies of all sizes—to uphold these values.

The Blueprint for an AI Bill of Rights lays out five core protections to which the American public should be entitled:

- **Safe and Effective Systems**: You should be protected from unsafe or ineffective systems.
- **Algorithmic Discrimination Protections**: You should not face discrimination by algorithms and systems should be used and designed in an equitable way.
- **Data Privacy**: You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used.
- **Notice and Explanation**: You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you.
- **Human Alternatives, Consideration, and Fallback**: You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter.

To protect the civil rights of Americans, and ensure technology is working for the American people, and to move these principles into practice, the Blueprint for an AI Bill of Rights also includes concrete steps which governments, companies, communities, and others can take in order to build these key protections into policy, practice, or technological design to ensure automated systems work in ways that protect human rights and democratic values.

The Blueprint for an AI Bill of Rights is focused on protecting human rights and democratic values, so the systems defined as in scope are based on *impact* as opposed to the underlying technical choices made in any system, since such choices can and do change with the speed of technological innovation. Specifically, the Blueprint should be applied with respect to all automated systems that have the potential to meaningfully impact individuals' or communities' rights, opportunities, or access, defined as below:

- **Rights, opportunities, or access**: The set of: civil rights, civil liberties, and privacy, including freedom of speech, voting, and protections from discrimination, excessive punishment, unlawful surveillance, and violations of privacy and other freedoms in both

public and private sector contexts; equal opportunities, including equitable access to education, housing, credit, employment, and other programs; or, access to critical resources or services, such as healthcare, financial services, safety, social services, non-deceptive information about goods and services, and government benefits.

The EU risk-based approach to AI

The EU approach to AI is **human-centric**, aiming to foster the trust of and uptake by citizens while offering the conditions for companies and researchers to develop and deploy trustworthy AI in Europe. A **balanced** approach to AI is needed in order to reap the benefits of this technology while addressing potential risks its use can pose to safety and fundamental rights.

Promoting the development of **trustworthy AI** is a key aspect of the European strategy on AI, and it plays a crucial role in the promotion of a values-based European digital economy and society. The EU supports basic and applied research, testing and experimentation (including regulatory sandboxes), and deployment.

Trust is also needed for uptake and adoption, and thus a **precondition for the benefits of AI** to materialize in the EU digital market. The EU's human-centric approach to AI involves balancing and assessing on an ongoing basis the progress and benefits of AI against their potential risks to individuals and society. Values guiding socio-technical governance efforts are derived from the Treaties of the European Union and its Charter of Fundamental Rights that prescribes a series of fundamental rights that EU member states and EU institutions are legally obliged to respect when implementing EU law.

Coordinated Plan on AI

The Coordinated Plan on AI (2001) puts forward EU measures on supporting innovation and enabling conditions, such as access to data and computing infrastructure, promoting the development and deployment of trustworthy AI solutions, training and skills development, as well as promoting the EU's value-based approach to AI on the global stage.

To develop a European **ecosystem of excellence**, the EU is setting up AI Networks of Excellence to foster cooperation among Europe's AI research teams to tackle major scientific and technological challenges in AI hampering deployment of AI-based solutions, including the development of ethical and trustworthy AI. Furthermore, it set up a European public-private partnership on AI, data and robotics.

To bridge the gap between AI research and deployment. AI Testing and Experimentation Facilities are being set up. They will allow companies to test their AI-based technologies in real-world environments. This will be complemented by the development of a European marketplace for trustworthy AI solutions, connecting resources and services to support innovators in developing and deploying trustworthy AI solutions.

The EU AI Act

Certain specific features of AI technologies (e.g. opacity) can make the application and enforcement of existing legislation more challenging and generate high risks for which a tailored regulatory response is needed. Therefore, the EU AI Act introduces a set of rules applicable to the design, development and use of certain high-risk AI systems, as well as restrictions on certain uses of remote biometric identification systems.

By earning people's trust, the envisaged risk-based legislation should also foster the uptake of AI across Europe. To be future-proof and innovation-friendly, the proposed legal framework is designed to intervene only where this is strictly needed and in a way that minimises the burden for economic operators, with a light governance structure.
The proposed AI regulation puts forward rules to enhance transparency and to minimise risks to safety and fundamental rights before AI systems can be used in the European Union. It is a proportionate and risk-based approach.

The proposal focuses on high-risk AI use cases. Whether an AI system is classified as high-risk depends on its intended purpose and on the severity of the possible harm and the probability of its occurrence.

AI systems identified as high-risk would include AI technology used in: safety components of regulated products; critical infrastructures; educational and vocational training; employment, workers management and access to self-employment; essential private and public services; law enforcement that may interfere with people's fundamental rights; migration, asylum and border control management; and administration of justice and democratic processes.

High-risk AI systems are to comply with specific requirements, which include the setting up of a sound risk management system, the use of high-quality datasets, appropriate documentation to enhance traceability, the sharing of adequate information with the user, the design and implementation of appropriate human oversight measures, and the achievement of the highest standards of robustness, safety, cybersecurity and accuracy.

Such requirements will be supported by harmonised technical standards to be developed by the European Standardisation Organisations (ESOs) on the basis of a mandate from the European Commission. Appropriate agreements in place between the ESOs and international standardisation organisations ensure that fit-for-purpose international standards can be taken over by ESOs and proposed as European harmonised standards in response to a standardisation request.

High-risk AI systems must be assessed for conformity with these requirements before being placed on the market or put into service. Depending on the type of high-risk AI system, the conformity assessment procedure may be based on internal control or rely on the involvement of a third-party certification body.
The proposed regulation will also encourage the use of regulatory sandboxes establishing a controlled environment to test innovative technologies for a limited time.