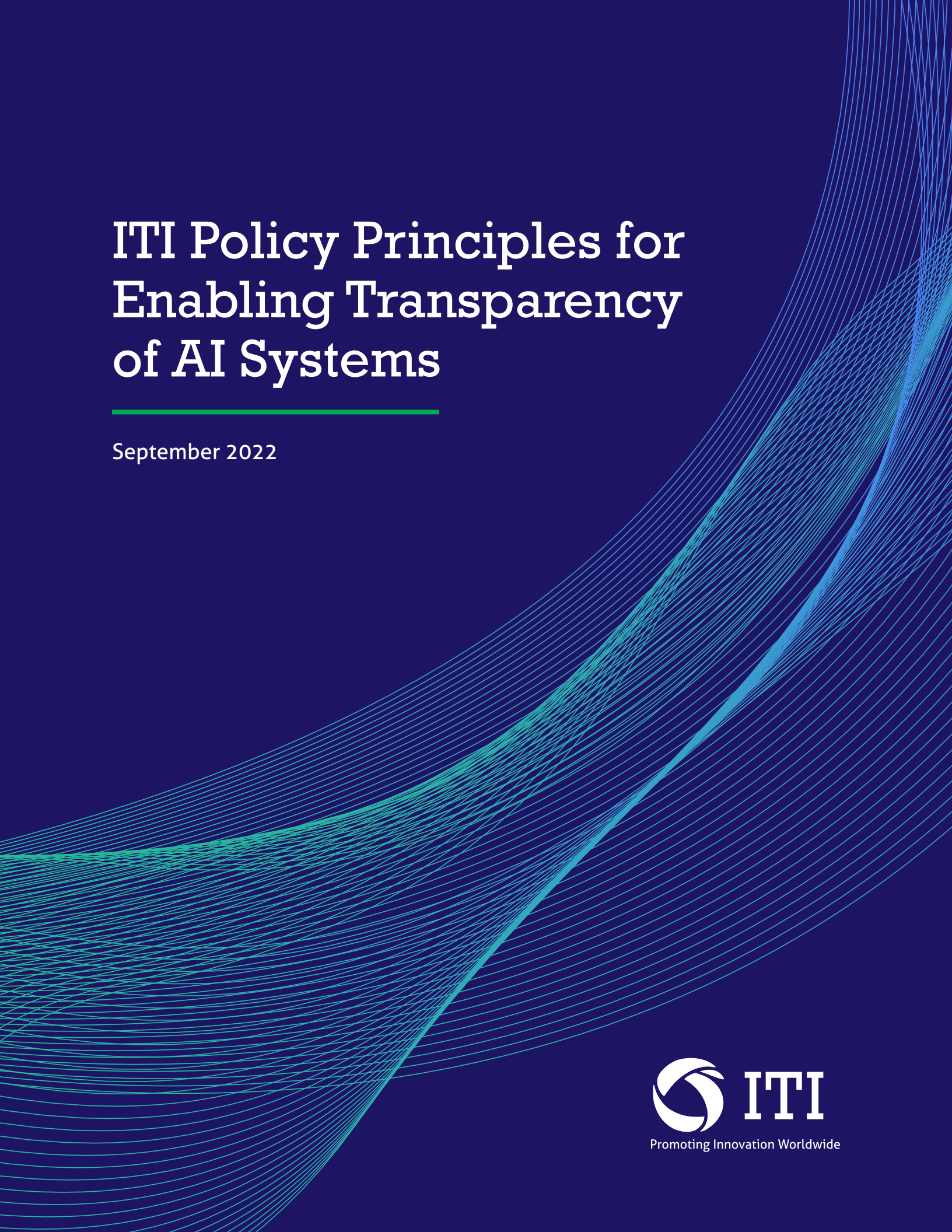
ITI Policy Principles for Enabling Transparency of AI Systems

September 2022



ITI
Promoting Innovation Worldwide

As Artificial Intelligence (AI) continues to evolve, policymakers are increasingly focused on how to best approach potential risks stemming from use of the technology.

In 2021, ITI released our *Global AI Policy Recommendations*, which offered a comprehensive set of policy recommendations for global policymakers seeking to foster innovation in AI while also addressing specific harms. A key tenet of those recommendations was that fostering acceptance and trust in AI systems1 is a shared responsibility and requires developers, industry, and policymakers to work together to collectively achieve that trust. Therefore, in those recommendations, we focused on facilitating public trust in and understanding of AI systems.

We encouraged governments to consider how to best promote the development of meaningfully explainable AI systems as *one way* to foster accountability, which therefore builds trust. Indeed, understanding how and/or why a system made the decision it did is critical to facilitating accountability. However, the broader concept of transparency is an important aspect of and necessary to developing accountable and trustworthy AI systems and avoiding unintended outcomes or other harmful impacts.

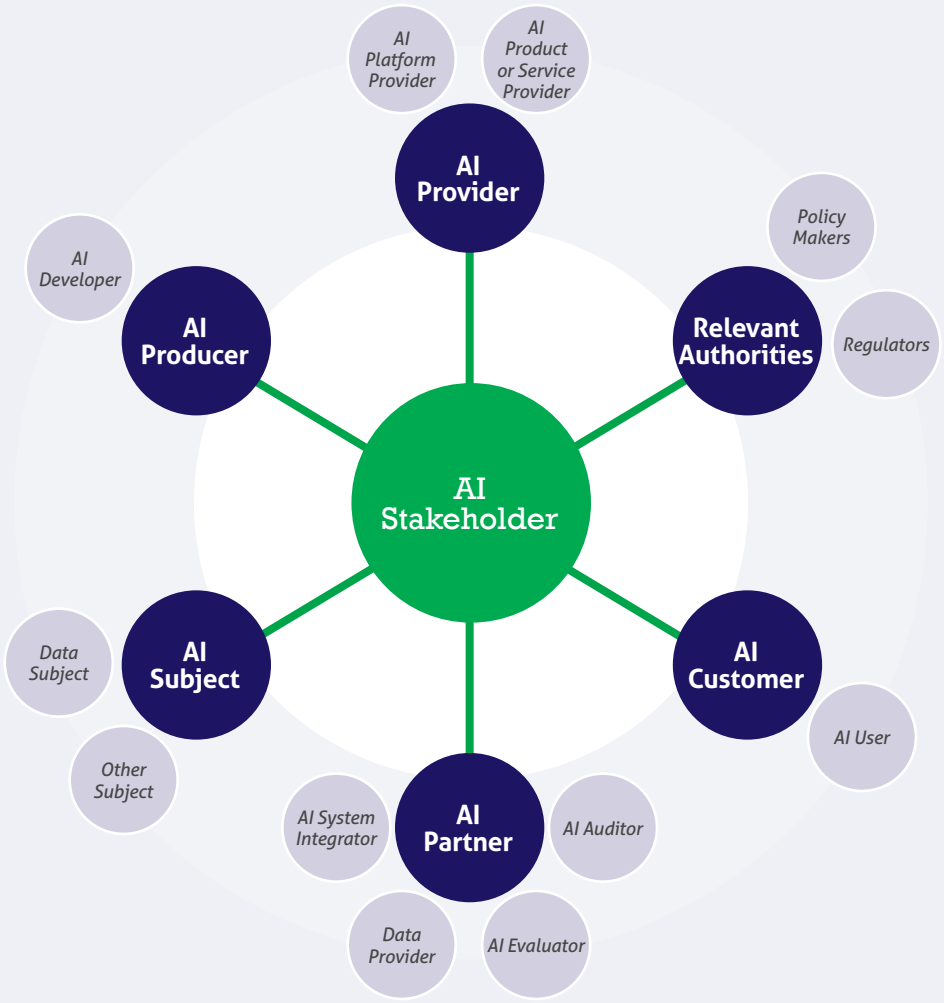
Accountability generally refers to the commitment that organizations (or individuals) will work to ensure that the AI systems they design, develop, or deploy function properly throughout their lifecycle and that they will implement mechanisms to demonstrate responsible AI systems development through their actions, including governance at the organizational level.[2] In the context of transparency, accountability refers to the need for organizations to make sure users are aware of the fact that they are interacting with an AI system, including, where appropriate, to provide them with an explanation and justification for a particular decision or outcome.[3]

As policymakers increasingly consider applying broad transparency requirements, they should be aware of the complexity inherent within that concept, as well as understand the various different types of transparency that can be achieved. They should also consider the ultimate objective of and audience for transparency requirements; take a risk-based approach to transparency requirements; include clear definitions of what is meant by transparency; consider that there are different ways to approach transparency; consider including provisions within legislation that are intended to provide users with sufficient information to understand decisions of an AI system that may negatively affect their fundamental rights and provide users with the ability to review and/or challenge such decisions; ensure transparency requirements do not implicate sensitive IP or source code; leverage voluntary international standards; and consider the role of disclosure requirements.

The purpose of this document is twofold – first, to educate policymakers about transparency in the context of AI systems, and second, to offer suggested policy approaches to facilitating greater transparency of AI systems. We stress that the better policymaking approach is to apply transparency requirements to specific, high-risk uses of *AI systems*, as opposed to imposing requirements on the transparency of algorithms which, although an important part of AI systems, are also an integral component of a much broader set of technologies used in a wide variety of settings. Indeed, it is the context in and purpose for which the algorithm is used that matters—not the fact that an algorithm is used in and of itself—and policymakers should recognize that there are many contexts in which providing transparency is likely not necessary.

AI Ecosystem Stakeholders

It is important to understand the various stakeholders that participate in the AI ecosystem. This graphic depicts key stakeholder groups, including types of stakeholders that fall within those groups modeled after ISO/IEC 22989. Understanding stakeholders will also help policymakers further contemplate questions around audience.



What is transparency?

At the highest level, transparency is about being clear about how an AI system is built, operated, and functions. When executed well, AI transparency can help analyze outputs and hold appropriate AI stakeholders accountable.

Transparency in the context of an AI system can take many forms, including explainability, interpretability, and disclosure. Transparency can mean making a user aware that they are interacting with an AI system, or that appropriate information about an AI system is made available to relevant stakeholders. Transparency can also enable a user or a regulator to understand the way that a system has made a particular prediction, decision or series of decisions. While the term transparency is sometimes used interchangeably with its component parts, which are further elaborated on below, the terms are not synonymous. In considering how to facilitate transparency of AI systems, it is thus important that policymakers understand the difference between these key terms, including to what and to whom they apply.

Understanding Explainability

Explainability

Explainability is the property of an AI system to express important factors influencing the AI system results in a way that humans can understand. It is intended to answer the question "why" an AI system made the decision it did without arguing that the course of action that was taken was necessarily optimal. While the term interpretability is oftentimes used interchangeably with the term explainability, for the purposes of driving interoperability and alignment, we use the term "explainability" as defined and published in the applicable ISO standard.[4]

Disclosure

Disclosure generally refers to making a user aware of the fact that they are interacting with or using an AI system – usually in "real time" or during the use of the system.

A helpful conceptual distinction may be viewing explainability as explaining the outputs of an AI model in a way that humans understand, focused more on the how, and interpretability as allowing humans to understand the inputs and outputs of the AI model, focusing more on the cause of the decision.

Policy Principles for Enabling Transparency of AI Systems

- **Consider what the ultimate objective of transparency requirements are.** To the extent that policymakers are considering including transparency requirements in a policy proposal, we encourage them to think about what the ultimate objective of such requirements are. Is it to ensure that the user knows that they are interacting with an AI system? Is it to provide a post-hoc explanation to users about a decision that was made and provide them with an appropriate redress mechanism should the decision negatively impact them? Is it to help researchers and developers[5] test and validate the AI model or system? Is it to enable the AI system deployer to investigate an incident?

  Is it to enable and authorize regulators or third parties such as auditors to evaluate a system's safety features? Or is it something else? Understanding the answers to these questions is critical to determining the ultimate direction of the policy proposal and what approach is most appropriate to help achieve that objective.

- **Consider the intended audience of any transparency requirements and at what point of the AI system lifecycle they would apply.** Policymakers should also consider the target audience at which transparency requirements are directed, including their level of expertise. They should also consider when such requirements would apply (e.g., pre-deployment or post-deployment). For example, transparency could be useful to several different audiences (e.g., regulators, consumers, developers, etc.), which will in turn influence requirements.

  Understanding the intended audience will also inform the type of information presented, the manner in which it is presented, and the amount of information presented. Indeed, if the purpose of a transparency requirement is to allow a user to understand how or why a decision was made and allow for redress, that will result in a very different set of information being provided than if such information is being provided to allow a regulator to evaluate a system for safety.

AI Transparency Requirements: *Mapping Audience to Objectives*

Policymakers should understand their target audience for and the objectives of transparency requirements. This graphic shows different potential audiences and associated objectives.

Additionally, policymakers should ensure that their objectives align with post-deployment requirements, rather than pre-deployment requirements, which will likely be difficult, if not impossible, to efficiently implement.

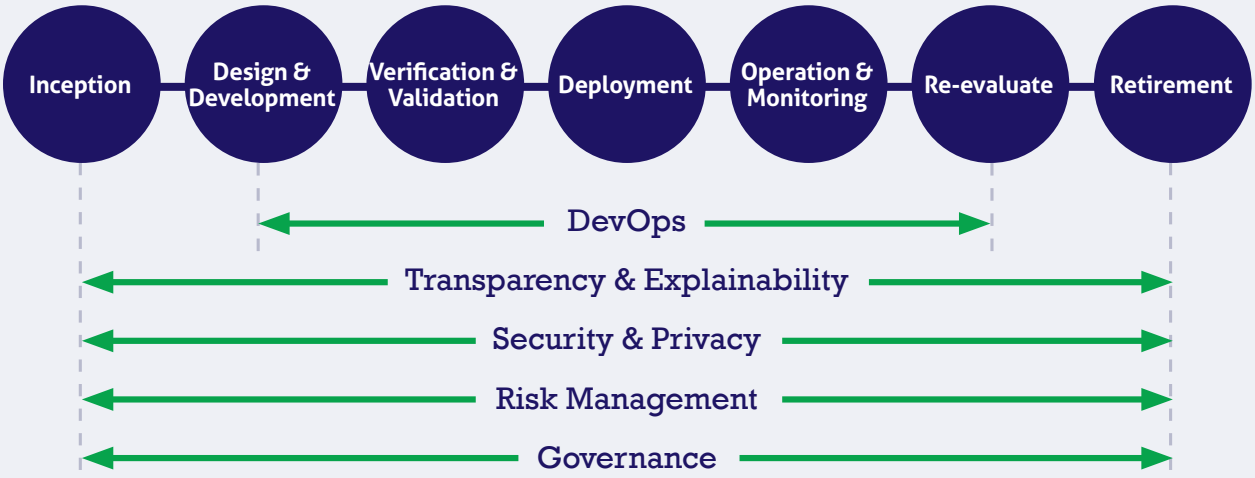
Finally, in thinking about the target audience, policymakers should also consider who will ultimately be required to comply with transparency requirements, and what information they may be required to provide. For instance, in the case of general purpose AI (GPAI) (i.e., general tools without an intended purpose and that can be used in a variety of use-cases and end products), the deployer of the end products with a high-risk use case would likely be best placed to implement potential transparency requirements, rather than the original GPAI developer, who would not have access to key information such as the context in which the system is deployed, its final use, the input data, the group(s) on which the system's outputs will have an impact, etc.

- ✔ **Take a risk-based approach to transparency when considering requirements.** In devising any requirements around transparency, policymakers should consider the diversity of AI applications and what their ultimate use-case will be, given the level of need and desire for transparency requirements from various users may significantly vary based on the AI application or intended use. Many use cases present little to no risk to the user, and so imposing transparency requirements in such situations will likely add little value to the user and hinder innovation by adding onerous, disproportionate requirements.

With this in mind, and in the context of post-deployment explainability requirements, we urge policymakers to remember that not all AI systems need to be explainable. It is our view that high-risk AI applications[6] are a more appropriate target for post-hoc explainability requirements.

AI System Lifecycle Model

This graphic depicts the AI system lifecycle. Transparency and explainability can exist at multiple points throughout the lifecycle, so it is important for policymakers to consider at what point it is most appropriate to apply requirements. This graphic is modeled after the AI lifecycle depiction in ISO/IEC 22989



- **Include clear definitions of what is meant by transparency in the context of a regulation or policy proposal.**[7] As a foundational matter, policymakers must take care to make clear what is meant by transparency in a regulation or policy proposal. Specifically, policymakers should include a definition of the term, what it applies to, who it applies to, and in what context it applies, including to clearly articulate any associated requirements. As outlined above, transparency can mean different things, so the lack of a clear definition of transparency in a policy or regulatory proposal may engender confusion and ultimately make it difficult for organizations to comply with or otherwise understand what it is they should be addressing. Policymakers should also be clear that when discussing transparency in the context of AI they are referring to the transparency of AI systems, as opposed to algorithmic transparency, which could apply more broadly.

- **Consider that there are different ways to approach transparency and improve trust, and that explainability is only one component.** While explainability is one way to approach transparency and therefore, engender trust, policymakers should recognize that it is not the only tool, nor is it a silver bullet solution.[8] Indeed, there are other ways that can help to create trust and deploy AI responsibly, including, for example, using technical, procedural, or educational tools to ensure that AI systems are fair and robust.[9] Another way one may approach transparency is to encourage certain stakeholders in the AI lifecycle to examine raw input data to understand the limitations of the dataset and account for and help manage potential bias, while respecting privacy. Policymakers should also understand there are limits to explainability in a variety of different contexts. For instance, in many cases explainability deployments are intended for machine-learning engineers to fix bugs in the model, as opposed to explaining the outcome to the users of those AI systems.[10] Another limitation to consider is that explainability does not necessarily equate to a higher confidence level and could in some cases lead to a misplaced sense of confidence in or understanding of the AI system. Keeping the above in mind, we expand upon several of the points we set forth in our *Global AI Policy Recommendations.*[11]

- **Consider including provisions within legislation that are intended to provide users with sufficient information to understand decisions of an AI system that may negatively affect their fundamental rights and provide users with the ability to review and/or challenge such decisions.** Similar to provisions in privacy legislation that allow data subjects to request review of decisions taken solely on the basis of automated processing of data affecting their interests, we are supportive of provisions that would similarly allow users to request clear information regarding a decision that negatively impacted their fundamental rights and to challenge such decisions as appropriate, keeping in mind the below considerations on protecting sensitive IP and source code. As mentioned above, we are supportive of proactive disclosures that enable a consumer to understand if they are interacting with an AI system, and to access additional information about the AI system itself in situations where their fundamental rights may have been negatively impacted, keeping in mind the below considerations on protecting sensitive IP and source code, expectations on how the system will be used, and any known limitations associated with the system. That being said, policymakers should avoid being too prescriptive and allow flexibility so that organizations can tailor such information depending on context, use of, and level or risk associated with the system.

✔ **Ensure that transparency requirements do not require companies to divulge sensitive IP or source code or otherwise reveal sensitive individual data.** Any requirements around transparency should avoid requiring companies to divulge sensitive IP or source code. Disclosure of source code could seriously put at risk trade secrets, undermine IP rights, and contravene widely accepted best practices for digital trade. It could also pose risks to safety and security and allow malicious actors to manipulate an AI system. Moreover, providing access to source code would not yield the information necessary to understand the way in which an AI system made a decision. One way to ensure that sensitive IP or source code is protected is to insist that any AI transparency requirements are post-deployment only, and not imposed as pre-deployment requirements, and that any requests to access source code clearly outline why such access is necessary and by who.

✔ **Leverage voluntary international standards in order to maintain interoperability of various AI transparency requirements to the extent possible.** Seeking to participate in and leveraging international, consensus-based standards will be useful in helping to increase alignment, interoperability, and trust in AI systems. In particular, ISO/IEC SC 42 is in the process of developing several standards, including on transparency taxonomy and objectives for explainability of ML models and AI systems.[12]

✔ **Consider that when an AI system is directly interacting with a user, that fact should be easily discoverable and that disclosure requirements can help facilitate this.** AI systems should be disclosed when they are playing a significant role in decision-making or interacting directly with users. Below, we offer three recommendations specific to transparency requirements as they relate to disclosure.

- **Ensure that disclosures use plain, clear language that is understandable to the user.** Because the primary purpose of disclosure is to make users aware of the fact that they are interacting with an AI system, language should be plain and clear so that it is understandable to a wide audience. Disclosure should generally include high-level information, including a topline explanation of how an AI system works, known limitations on performance, and expectations around how the system will be used.[13]

  In some cases, it may be beneficial to provide more technical information within a disclosure – for example, if the user is a regulator reviewing the AI system. This information might include things like how well the AI system performs for industry standard evaluation datasets measured against key metrics; providing an indication of the frequency and cost weighting assigned to different errors (e.g., false negatives/false positives); and, if relevant, how the AI system's performance compares to existing human performance benchmarks.

- **Regulations pertaining to disclosure should be flexible and avoid prescribing specific information or technical details to be included.** Although we have offered some thoughts around what may be useful to include in a disclosure, it is important that regulation allows for flexibility because it may not be appropriate or useful to provide the same type of details in every context or for every target audience.[14]

- **Only the actual deployer of the AI system should be responsible for disclosure.** The developer of an AI system cannot anticipate every single possible use case for its system, and as such, it should be the responsibility of the ultimate deployer – that is, the user that is deciding the means by and purpose for which the AI system is being used—to ensure that disclosure and other consumer-facing obligations are met. That said, the developer of the AI system should ensure that terms of sale do not prohibit disclosure.

References

[1]We define an AI system as a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy. This is based on the OECD definition of AI.

[2]OECD AI Principles, available here : https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[3] A user means the person or entity who is ultimately interacting with or utilizing the AI system.

[4]See ISO/IEC 22989: 2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology, which defines key AI terminology, including explainability. Available here: https://www.iso.org/obp/ui/#iso:std:iso-iec:22989:ed-1:v1:en. We also highlight additional ongoing work in ISO/IEC that will be useful for policymakers to stay abreast of in order to foster additional alignment moving forward, including ISO/IEC IS 12792 Artificial Intelligence – Transparency taxonomy of AI systems, ISO/IEC TS 5471 Artificial Intelligence – Quality evaluation for AI systems and ISO/IEC TS 6254 Artificial intelligence – Objectives and approaches for explainability of ML models and AI systems.

[5]A developer (sometimes used interchangeably with producer) is the entity that is producing the AI system. In some cases, the AI system can be built into other products that are then deployed by a different entity. A deployer is the entity (sometimes used interchangeably with provider) that is deciding the means by and purpose for which the AI system is ultimately being used and puts the AI system into operation.

[6]We consider an AI application to be high-risk when a negative outcome could have a significant impact on people — especially as it pertains to health, safety, freedom, discrimination, or human rights..

[7]For example, transparency is construed in different ways in the European Commission's proposal for Regulation on AI (i.e. EU AI Act). Both Article 13 (1) and Article 52 (1) implicate transparency. Article 13(1) applies to deployers of a high-risk AI system and Article 52 (1) applies to deployers of a system that interact with natural persons. It is not clear how the provisions are related to each other, especially because 13(1) appears to implicate interpretability, whereas 52 (1) is more focused on disclosure. As such, we do not believe that the text sufficiently differentiates between the component parts of transparency and may therefore cause confusion. It is important to note that at the time of publication of this document, amendments to the text are being proposed, and so this issue may be resolved in the future..

[8]We encourage policymakers to review the OECD's Tools for Trustworthy AI: A framework to compare implementation tools for trustworthy AI systems, available here: https://doi.org/10.1787/008232ec-en. This may offer insight into other tools that are available to facilitate trustworthiness.

[9]See the OECD Tools for Trustworthy AI: A framework to compare implementation tools for trustworthy AI systems, available here: https://doi.org/10.1787/008232ec-en. The framework outlines various tools that can be used to facilitate trustworthiness as laid out in the OECD AI Principles, as well as ways to compare said tools.

[10]See Bhatt, et. al, Explainable Machine Learning in Deployment, available here: https://dl.acm.org/doi/pdf/10.1145/3351095.3375624

[11]See ITI's Global AI Policy Recommendations here: https://www.itic.org/documents/artificial-intelligence/ITI_GlobalAIPrinciples_032321_v3.pdf

[12]https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0

[13]It is not possible to anticipate every single potential use-case for an AI system. However, AI systems are designed with particular use-cases in mind and disclosing that information can be helpful to a user in understanding what it was tested/marketed for.

[14]Using AI to limit fraud, spam, illegal, or malicious information are some examples of where including technical details or too prescriptive of a disclosure may be inappropriate.