12 June 2023

To: National Telecommunications and Information Administration

From: Ed Palacio, President, IEEE-USA

In re: *AI Accountability Policy Request for Comment, Docket No. 230407-0093*

IEEE-USA is pleased to submit the following comments in response to the NTIA's request for comments, published at 88 FR 22433 (13 April 2023) requesting input from stakeholders in the policy, legal, business, academic, technical, and advocacy arenas on how to develop a productive AI accountability ecosystem. We commend the NTIA for promoting the application of accountability measures to ensure that autonomous and intelligent systems (AI/S) are legal, effective, ethical, safe, and otherwise trustworthy.

IEEE-USA represents approximately 180,000 engineers, scientists, and allied professionals living and working in the US. Our members work in AI-related industries, developing and working with the emerging technologies used in artificial intelligence systems. This expertise provides us with a unique perspective on the benefits of these technologies. Our responses to select questions are below.

If you have any questions and wish to discuss our input, please contact Erica Wissolik at e.wissolik@ieee.org.

AI Accountability Objectives

1. What is the purpose of AI accountability mechanisms such as certifications, audits, and assessments?

The programming, output, and purpose of A/IS are often not discernible by the public. Based on the cultural context, application, and use of A/IS, people and institutions need clarity around the manufacture and deployment of these systems to establish responsibility and accountability, and to avoid potential harm. Additionally, manufacturers of these systems must be accountable in order to address legal issues of culpability. It should, if necessary, be possible to apportion culpability among responsible creators (designers and manufacturers) and operators to avoid confusion or fear within the public. **The purpose of accountability mechanisms is to ensure that A/IS meet these requirements (or, put more generally, to ensure that A/IS is created and operated to provide an unambiguous rationale for decisions made)**.

The principle of accountability is closely linked with each of the other principles intended to foster informed trust in A/IS: effectiveness, competence, and transparency. **With respect to effectiveness**, evidence of attaining key metrics and benchmarks to confirm that A/IS are functioning as intended

IEEE-USA | 2001 L Street, N.W., Suite 700, Washington, D.C. 20036-4928 USA
Office: +1 202 785 0017 | Fax: +1 202 785 0835 | E-mail: ieeeusa@ieee.org | Web: http://www.ieeeusa.org

may put questions of where, among creators, owners, and operators, responsibility for the outcome of a system lies on a sound empirical footing. **With respect to competence**, operator credentialing

and specified system handoffs enable a clear chain of responsibility in the deployment of A/IS. **With respect to transparency**, providing a view into the general design and methods of A/IS, or even a specific explanation for a given outcome, can help to advance accountability.

Providing further grounding of these points, we add that to be accountable means that those who rely on you – or in this case autonomous and intelligent systems (A/IS) – can trust the work that you do. In other words, others believe in and are comfortable with the outcomes. As A/IS are increasingly deployed for the purpose of making consequential decisions that impact the lives, liberty, and well-being of individuals in areas such as hiring, housing, credit, and criminal justice, A/IS trustworthiness is of crucial importance. A/IS are influencing serious large-scale societal governance decisions via the amplification of news, capital flows, and allocation of public resources.

If society cannot trust A/IS-enabled decisions, it cannot trust in the effective functioning of the core social institutions (medicine, law, finance, government) that rely on those decisions, putting the legitimacy of governance in doubt. A/IS offer the potential for more uniform, repeatable, and less-biased decisions, but that is far from automatic and far from the only criteria for trustworthiness. This technology can codify the biases, conscious and unconscious, of system builders, as well as ingest biases reflected in historical training data, thus cementing these biases into opaque software systems. A/IS can also suffer from a lack of transparency, accountability, or respect for human rights. To mitigate these risks, A/IS must be deployed in a socio-technical context where sound evidence of their fitness for purpose is demanded, there are incentives to identify and correct problems, and stakeholders are empowered with the information they need to advocate for their own interests.

While the question of trustworthiness is an important one, it is also a challenge because it requires answering multiple subsidiary questions. For instance, we can ask about the ends for which we deploy an A/IS: What are the objectives we set for the system? Or, more fundamentally, what are the values we are trusting a system to advance or protect? Or we can question the basis of trust in technology: What do we need to know about a system to trust it? Under what conditions will we trust that a system is capable of realizing the objectives we set for it? A third question is about evidence: What evidence is available to us for determining whether a system has met the conditions required for giving it our trust? All these questions can be answered by applying standards, audits, and assessments.

To have a practical impact, we must go beyond abstract discussions of evidence and consider norms that could be put in place for making the provision of the required evidence a regular practice of the developers of A/IS. We believe that, if our societies are to reach a state in which new technologies can be trusted by those who use and are affected by them, it is necessary to shift the burden from the end user (or decision subject) to creators, developers, and operators. Instead of asking the user to research and make determinations of trustworthiness (often without adequate access to the

information required to do so), we should be asking creators, developers, and operators to provide users and stakeholders with sufficient evidence to establish reliability.

We also realize, however, that shifting that burden will be accomplished only if practically feasible normative frameworks (standards) are developed that make providing the required evidence an expected, and accepted, part of the regular practice of creators, developers, and operators.

For example, adopted technical and governance standards can be used to assess the reliability of evidence and trustworthiness of A/IS and of the types of evidence that might meet the standard.

a. What kinds of topics should AI accountability mechanisms cover? How should they be scoped?

Accountability, as noted above, is closely connected with other principles for the responsible use of AI, both value-oriented principles (such as fairness, privacy, liberty, respect for human rights, respect for individual dignity, and so on) and trust-oriented principles (such as effectiveness, competence, and transparency). An organization's realization of those other principles is dependent on its design and implementation of an effective accountability regime that is responsive to the questions raised by those principles. Accordingly, the set of accountability mechanisms must, collectively, be broad in scope and designed to address the range of issues raised by both value- and trust-oriented principles.

b. What are assessments or internal audits most useful for? What are external assessments or audits most useful for?

Internal assessments provide a view to where, in an organization's procurement, adoption, implementation, operation, and retirement of AI-enabled systems, accountability mechanisms are weak or lacking altogether. Internal assessments should be a standard component of an organization's on-going efforts to maintain and improve AI- and data-governance policies and processes.

External assessments avoid conflicts of interest and can be used to inform external stakeholders whether or not a given system is trustworthy; as such, they are generally more rigorous in their design and in their collection of evidence.

Whether internal or external, an assessment must address the practical question of evidence, specifically the evidence necessary to assess whether a system in fact meets the trust conditions. What tests or evidence can give us adequate reason to believe that a technology is effective, is operated competently, is implemented under an adequate accountability protocol, and is open to meaningful inspection and audit (i.e., transparent)? In the case of more rigorous external assessments, an organization should look to external, or independent, verification and validation (V&V). The V&V process must be independent to avoid conflicts of interest that could lead to catastrophic failure. To this end, IEEE 1012 for System, Software and Hardware Verification and Validation (V&A) (*https://standards.ieee.org/ieee/1012/5609/*) provides specific advice about avoiding such conflicts by requiring technical, managerial, and financial independent V&V when

testing software and hardware. Allowing developers to certify their own software is a clear conflict of interest.

c. An audit or assessment may be used to verify a claim, verify compliance with legal standards, or assure compliance with non-binding trustworthy AI goals. Do these differences impact how audits or assessments are structured, credentialed, or communicated?

Yes. The purpose for which an assessment if being conducted shapes the nature of the evidence that is collected pursuant to the assessment, the manner in which that evidence is collected, and the way in which the evidence is analyzed and, ultimately, reported. Different goals will mean different levels of rigor and depth in the collection and analysis of evidence and different levels of intrusion in the operations of the systems being assessed. Assessments should be properly calibrated to the goals they are intended to achieve, lest they either gather insufficient evidence (and so fail to demonstrate compliance) or gather too much evidence (and so become an undue burden on operations).

d. Should AI audits or assessments be folded into other accountability mechanisms that focus on such goals as human rights, privacy protection, security, and diversity, equity, inclusion, and access? Are there benchmarks for these other accountability mechanisms that should inform AI accountability measures?

As noted above, accountability is closely connected with other principles for the responsible use of AI, both value-oriented principles and trust-oriented principles. Accordingly, accountability mechanisms must be designed to address the range of issues raised by both value- and trust-oriented principles.

e. Can AI accountability practices have meaningful impact in the absence of legal standards and enforceable risk thresholds? What is the role for courts, legislatures, and rulemaking bodies?

AI accountability practices can have impact outside of legal standards; however, such voluntary practices can be strengthened when legal standards and policies are also established. A combination of voluntary standards or practices with laws and regulation can help mitigate risks posed by AI/S. For example, technical standards provide practical guidelines and specifications for implementing regulatory requirements and can provide more specific direction for particular sectors or applications. We must recognize that science and technology—for all their power to create, preserve, and destroy—are not the only engines of innovation in the world. Other social institutions also innovate, and they play an invaluable part in realigning the aims of science and technology with those of culturally disparate human societies. Foremost among these is the law. Laws ensure that A/IS, in both design and operation, are aligned with principles of ethics and human well-being.

2. Is the value of certifications, audits, and assessments mostly to promote trust for external stakeholders or is it to change internal processes? How might the answer influence policy design?

The value is in both. On the one hand, accountability and mechanisms designed to maintain accountability (such as certifications and audits) serve the interest of promoting trust among external stakeholders. An essential condition of stakeholders having an informed trust in a technological system is confidence that it is possible, if the need arises, to apportion responsibility among the human agents engaged along the path of its creation and application: from design through to development, procurement, deployment, operation, and, finally, validation of effectiveness. Unless there are mechanisms to hold the agents engaged in these steps accountable, it will be difficult or impossible to assess responsibility for the outcome of the system under any framework, whether a formal legal framework or a less formal normative framework. A model of A/IS creation and use that does not have such mechanisms will also lack important forms of deterrence against poorly thought-out design, casual adoption, and inappropriate use of A/IS. Simply put, a system that produces outcomes for which no one is responsible is one which stakeholders cannot trust.

On the other hand, accountability (and mechanisms (such as certifications and audits) designed to maintain accountability will (and should) change internal processes. An environment of AI creation and use that does not have such mechanisms will lack important forms of deterrence against poorly thought-out design, casual adoption, and inappropriate use of A/IS. Those engaged in creating, procuring, deploying, and operating a system under such a model will lack the discipline engendered by the clear assignment of responsibility. Conversely, the presence of meaningful accountability mechanisms will foster an environment of responsibility and attention to core values (such as fairness, privacy, individual dignity, etc.) throughout the lifecycle of an AI application.

It is important to note, however, that there is an interdependency between the realization of trust and the achievement of change: a change in internal processes (and in the behavior of the agents implementing those processes) is not only a result of the maintenance of accountability, it is also a precondition of the maintenance of an effective trust-oriented accountability regime. The first step in implementing such a regime is ensuring that all those engaged in the creation, procurement, deployment, operation, and testing of A/IS recognize that, if accountability is not maintained, these systems will not be trusted. In the interest of maintaining accountability, these stakeholders should take steps to clarify lines of responsibility throughout this continuum, and make those lines of responsibility, when appropriate, accessible to meaningful inquiry and audit.

3. AI accountability measures have been proposed in connection with many different goals, including those listed below. To what extent are there tradeoffs among these goals? To what extent can these inquiries be conducted by a single team or instrument?

a. The AI system does not substantially contribute to harmful discrimination against people.

b. The AI system does not substantially contribute to harmful misinformation, disinformation, and other forms of distortion and content-related harms.

c. The AI system protects privacy.

d. The AI system is legal, safe, and effective.

e. There has been adequate transparency and explanation to affected people about the uses, capabilities, and limitations of the AI system.

f. There are adequate human alternatives, consideration, and fallbacks in place throughout the AI system lifecycle.

g. There has been adequate consultation with, and there are adequate means of contestation and redress for, individuals affected by AI system outputs.

h. There is adequate management within the entity deploying the AI system such that there are clear lines of responsibility and appropriate skillsets.

The consistent achievement of all these goals is dependent on having in place an effective accountability regimen. However, the nature of the evidence collected, the manner in which it is collected, the metrics generated, the reporting of results will all differ from one goal to the next. Specific accountability mechanisms should be specifically designed for each goal which then, collectively, should cover all goals.

5. **Given the likely integration of generative AI tools such as large language models (e.g., ChatGPT) or other general-purpose AI or foundational models into downstream products, how can AI accountability mechanisms inform people about how such tools are operating and/or whether the tools comply with standards for trustworthy AI?**

An essential component of trust in a technology is trust that it works and meets the purpose for which it is intended. This is true of all technologies, including applications based on large language models. What this means is that one goal of accountability mechanisms is to inform stakeholders (which, in many cases, includes the public) of the **effectiveness** of the technology.

What is meant by effectiveness?

In gathering evidence of effectiveness, we are seeking to gather empirical data that will tell us whether a given technology, or its application will serve as an effective solution to the problem it is intended to address. Serving as an effective solution means more than meeting narrow specifications or requirements; it means that the A/IS can address their target problems in the real world. It also means remaining practically feasible once collateral concerns and potential unintended consequences are considered.

Viewed in this light, assessing the effectiveness of an AI-enabled application in accomplishing the target task (narrowly defined) is not sufficient; it may also be necessary to assess the extent to which the application is aligned with applicable laws, regulations, and standards, and whether (and to what extent) it impinges on values such as privacy, fairness, or freedom from bias. It is only from such a complete view of the impact of A/IS that a balanced judgment can be made of the appropriateness of their adoption.

What we are measuring is therefore an application's general "**fitness for purpose**."

What form do the results of an effectiveness evaluation take?

The results of an evaluation typically take the form of a number—a quantitative gauge of effectiveness. This can be, for example, the decreased likelihood of developing a given medical condition; safety ratings for automobiles; recall measures for retrieving responsive documents; and so on. Certainly, qualitative considerations are not (and should not) be ignored; they often provide context crucial to interpreting the quantitative results. Nevertheless, at the heart of the results of an evaluation exercise is a number, a metric that serves as a telling indicator of effectiveness.

In some cases, the research community engaged in developing any new system will have reached consensus on salient effectiveness metrics. In other cases, the research community may not have reached a consensus, requiring further study. In the case of AI, given both their accelerating development and the fact that they are often applied to tasks for which the effectiveness of their human counterparts is seldom precisely gauged, we are often still at the stage of defining metrics. An example of an application of AI for which there is a general consensus around measures of effectiveness is legal electronic discovery, where there is a working consensus around the use of the evaluation metrics referred to as "recall" and "precision." Conversely, in the case of AI applied in support of sentencing decisions, a consensus on the operative effectiveness metrics does not yet exist.

Who is the audience for the results of the effectiveness evaluation?

In defining metrics, it is important to keep in mind the consumers of the results of an evaluation of effectiveness. Broadly speaking, it is helpful to distinguish between two categories of stakeholders who will be interested in measurements of effectiveness:

- Experts are the researchers, designers, operators, and advanced users with appropriate scientific or professional credentials who have a technical understanding of the way in which a system works and are well versed in evaluation methods and the results they generate.
- Nonexperts are the policymakers, professionals, decision subjects, communities, and passive users whose work or outcomes may, even if only indirectly, be affected by the results of a given system. These individuals, however, may not have a technical understanding of the way in which a system operates. Furthermore, they may have little experience in conducting scientific evaluations and interpreting their results.

Effectiveness metrics must meet the needs of both expert and nonexpert consumers.

What measurement practices generate sound and meaningful metrics (and so foster an informed trust)?

By equipping both experts and nonexperts with accurate information regarding the capabilities and limitations of a given system, measurements of effectiveness can provide society with information needed to adopt and apply AI in a thoughtful, carefully considered, beneficial manner.

For the practice of measuring effectiveness to realize its full potential for fostering trust and mitigating the risks of uninformed adoption and uninformed avoidance of adoption, it must have certain features:

- **Meaningful metrics**: As noted above, an essential element of a measurement practice is a metric that provides an accurate and readily understood gauge of effectiveness. The metric should provide clear and actionable information as to the extent to which a given application has, or has not, met its objective so that potential users of the results of the application can respond accordingly. For example, in legal discovery, both recall and precision have done this well and have contributed to the acceptance of the use of A/IS for this purpose.
- **Sound methods**: Measures of effectiveness must be obtained by scientifically sound methods. If, for example, measures are obtained by sampling, those sample-based estimates must be the result of sound statistical procedures that hold up to objective scrutiny.
- **Valid data**: Data on which evaluations of effectiveness are conducted should accurately represent the actual data to which the given A/IS would be applied and should be vetted for potential bias. Any data sets used for benchmarking or testing should be collected, maintained, and used in accordance with principles for the protection of individual privacy and agency.
- **Awareness and consensus**: Measurement practices must not only be technically sound (in terms of metrics, methods, and data), but they must be widely understood and accepted as evidence of effectiveness.
- **Implementation**: Measurement practices must be both practically feasible and actually implemented (i.e., widely adopted by practitioners) .
- **Transparency**: Measurement methods and results must be open to scrutiny by experts and the public. Without such scrutiny, the measurements will not be trusted and will be incapable of fulfilling their intended purpose.

In seeking to advance informed trust in AI, and accountability for its results, policymakers should formulate policies and promote standards that encourage sound measurement of effectiveness.

6. **The application of accountability measures (whether voluntary or regulatory) is more straightforward for some trustworthy AI goals than for others. With respect to which trustworthy AI goals are there existing requirements or standards? Are there any trustworthy AI goals that are not amenable to requirements or standards? How should accountability policies, whether governmental or non-governmental, treat these differences?**

On this question, it is helpful to distinguish value-oriented goals (such as fairness, privacy, liberty, respect for human rights, respect for individual dignity, and so on) from trust-oriented goals (such as effectiveness, competence, accountability, and transparency).

The former (value-oriented goals) often involve hard-to-define, even "essentially contested," concepts. This makes identifying evidence and defining metrics that might be used in objective

assessments of compliance with these goals a much greater challenge. As an example, see the discussion of the challenges with measuring fairness here (https://digitaltechitp.nz/2021/04/19/operationalizing-values-in-ai-the-case-of-fairness/)

The latter (trust-oriented goals) often, but not always, involve concepts for which there is already a working definition. This makes identifying evidence and defining metrics that might be used in objective assessments of compliance with these goals a challenge that is more easily met.

To assist in the operationalization of principles for the trustworthy development and use of AI, the IEEE has advanced a three-tier framework the connects high-level, value-oriented principles through the conditions of trust that an application adheres to those principles to the concrete evidence that allows an evaluation of whether the trust conditions are met.

The framework is as follows.

Tier 1 - Ethics and Values: What is the purpose of a given technology? What values and ethical considerations should it adhere to?

> Available assessment tools include, *IEEE Ethically Aligned Design* (EAD), a document providing principles of human rights, wellbeing, data agency, and awareness of misuse; and the IEEE 7000 standards series addressing specific issues at the intersection of technological and ethical considerations.

Tier 2 - Trust Conditions: What conditions must be met to allow for an informed trust in a technology Under what conditions could an end user (or other stakeholders) trust that a given technology is, in fact, fit for its purpose and adheres to the values and ethical considerations identified in the Tier 1 analysis?

> Available assessment tools include, IEEE EAD's principles of effectiveness, competence, accountability and transparency, and the following IEEE standards:
>
> IEEE P2817 Guide for Verification of Autonomous Systems
>
> IEEE P2894 Guide for an Architectural Framework for Explainable Artificial Intelligence
>
> IEEE P2863 Recommended Practice for Organizational Governance of Artificial Intelligence

Tier 3 - Evidence: What evidence is available for assessing whether (or demonstrating that) a given technology meets the trust conditions identified in the Tier 2 analysis?

> Available assessment tools include, IEEE 1012 Standard for System, Software, and Hardware Verification, and Validation; and IEEE CertifAIEd, a Certification Program that enables, enhances, and reinforces trust through AI Ethics specifications, training, criteria, and certification. The rationale is that an entity benefits from an independent ethical

evaluation and certification of its A/IS. The IEEE CertifAIEd Mark communicates additional confidence for entities that have their AIS ethically aligned with expected and consistent behaviors and conveys trust to customers and consumers.

For any given high-level goal, the framework can be used to think through the means for assessing whether the goal is being met by a given application, thereby allowing discernment of those for which assessment mechanisms are readily available from those for which effective assessment mechanisms have still to be defined and designed.

Features of sound evidence. A key condition of putting the framework outlined above to effective use is the ability to gather meaningful evidence. Elaborating on that point, we note the following characteristics of trustworthy evidence.

- **Objective**: A statement of fact to which one can arrive with little application of expert judgment and about which competent individuals could not reasonably disagree.
- **Repeatable/Reproducible**: An outcome that can be consistently reproduced with additional trials.
- **Transparent/Auditable**: The evidence is obtained via a process that is transparent and open to audit by competent experts.
- **Empirically validated**: Validated by empirical testing. More specifically, both the accuracy and the consistency of the evidence have been empirically tested and quantified via meaningful and statistically sound metrics.
- **Competent agency**: Obtained by agents with the skills and experience required to maintain its accuracy and integrity.
- **Adherence to operative norms**: Obtained via a protocol that adheres to operative scientific, legal, and ethical norms.
- **Authoritative**: Supported by the testimony of credentialed experts and by appeal to a reasonably strong consensus within the relevant scientific community.
- **Probative value**: The evidence contributes unique and meaningful information to the question at hand.

Types of evidence. With these features in mind (and noting that it is not expected that all evidence will exhibit all these features), we can take inventory of the types of evidence that might be gathered in an assessment of the trustworthiness of an AI-enabled application. We organize the evidence under the rubrics of the four trust conditions (effectiveness, competence, accountability, and transparency).

For effectiveness. Evidence relevant to the assessment of a system's effectiveness will take the form, first and foremost, of scientifically sound empirical trials. Other types of evidence (such as that provided by a more subjective qualitative analysis of the results of a trial) can fill out the picture. It is also important to assess a system's effectiveness in the context of its overall fitness for its intended purpose, a principle that has become well established in the law. Specific types of evidence relevant to effectiveness include:

- Local validation exercises (both during development and after deployment and operation);
- Benchmarking studies — especially independent benchmarking studies;
- Algorithmic risk assessments: evaluations of the potential harms that might arise from the use of the system before it is launched into the world (e.g., environmental impact assessments);
- Algorithmic impact evaluations: evaluations of the system and its effects on its subjects after it has been launched into the world;
- Qualitative analysis of the results of validation exercise or benchmarking evaluations;
- Documentation of compliance with technical standards, certifications, and with any relevant regulations (including those relating to data security and privacy);
- Descriptions of the process followed in designing and developing the system; and
- Descriptions of the process followed in implementing the system.

For competence. Evidence relevant to an assessment of the competence of the human agents involved in the deployment and operation of a system will generally take the form of authoritative documentation of the agents' attainment of professional standards, but evidence from prior work in operating similar systems will also be highly relevant. Specific types of evidence relevant to competence include descriptions or documentation of:

- Purposes, capabilities, and limitations of the technology;
- Intended human roles in the development and operation of the system;
- Qualifications of the individuals actually filling the roles (including relevant certifications and documentation of past experience and education);
- Results of any prior testing of the individual's accuracy in using the system;
- Provisions for human oversight and evaluation of operators;
- Educational resources available to developers, operators, and end users; and
- Compliance with standards and regulations (those specifically relevant to operators).

For accountability. Evidence relevant to an assessment of accountability will generally take the form of descriptions of protocols for maintaining lines of communication and responsibility throughout a system, but documentation of prior responses to actual events can also be highly relevant. Specific types of evidence relevant to accountability include:

- A comprehensive map of roles and responsibilities so that each material decision can be traced back to a responsible individual;
- Established, understandable, and customary documentation (such as requirements, specifications, test plans, maintenance records, change logs, etc. as are common best practices in software engineering) that can provide the appropriate level of explanation to internal compliance departments, outside auditors, lawyers, judges, and the ordinary citizen, including:
    - Documentation of oversight and communication protocols;

    - Documentation of provisions for maintaining data security, integrity, and for disposing of data when no longer needed;
    - Descriptions of documentation maintained in the normal course of operation; and
    - Descriptions of documentation developer/operator is (is not) willing to disclose to an independent auditor;
- Descriptions of responses to prior adverse events.

For transparency. Evidence relevant to an assessment of transparency will generally take the form of descriptions of resources available to stakeholders interested in understanding the operation of a system and finding an explanation for its results in a given circumstance. Specific types of evidence relevant to transparency include documentation of:

- Access to reliable information about the A/IS including the training procedure, training data, machine learning algorithms, and methods of testing and validation;
- Access to a reliable explanation calibrated for different audiences – i.e., why an autonomous system behaves in a certain way under certain circumstances or would behave in a certain way under hypothetical circumstances;
- Engineering steps throughout the lifecycle of the system: design documentation (requirements, thread models), development (coding standards, unit tests, code review processes), procurement (who made the decisions and on what basis), deployment/ operation (workflows followed, qualifications of personnel), and validation (records of errors found, how repaired); and
- What degree of oversight, if any, is provided by human decision makers when considering the output of the A/IS.

7. Are there ways in which accountability mechanisms are unlikely to further, and might even frustrate, the development of trustworthy AI? Are there accountability mechanisms that unduly impact AI innovation and the competitiveness of U.S. developers?

Yes. Accountability assessments and transparency requirements, if not properly calibrated to the actual need, can be unduly burdensome and raise significant IP questions. Arriving at an effective and unduly burdensome governance model will therefore require the participation of those engaged in the creation and operation of AI, those affected by the results of its use, and those with the expertise to understand why and how the governance model is being implemented in each circumstance.

<u>Existing Resources and Models</u>

9. What AI accountability mechanisms are currently being used? Are the accountability frameworks of certain sectors, industries, or market participants especially mature as compared to others? Which industry, civil society, or governmental accountability instruments, guidelines, or policies are most appropriate for implementation and operationalization at scale in the United States? Who are the people currently doing AI accountability work?

When AI-enabled systems are used to automate critical decisions in regulated areas such as hiring, housing, credit, and allocation of public resources, it is important to use transparent and accountable processes. IEEE standard 1012 for System, Software, and Hardware Verification and Validation is a universally applicable and broadly accepted process for ensuring that the right product, whether it is software or hardware, is correctly built for its intended use. It consists of a process, which if followed, provides a reasonable degree of assurance that the system adheres to the intent of the design and requirements. It has been used to verify and validate Department of Defense nuclear weapons systems and NASA manned space systems.

Verification and validation (V&V) are interrelated and complementary processes that build quality into any system. Verification is focused on a product, providing objective evidence for whether it conforms to requirements, standards, and practices. Validation focuses on customers and stakeholders, providing evidence for whether a product is accurate and effective, solves the right problem, and satisfies the intended use and user needs in the operational environment. In short, verification ensures that a product is correctly built, while validation ensures that the right product is built.

To appropriately perform V&V, IEEE1012 requires that each software and hardware component be assigned an integrity level that increases depending on the likelihood and consequences of a failure: negligible, marginal, critical (causing "major and permanent injury, partial loss of mission, major system damage, or major financial or social loss"), and catastrophic (causing "loss of human life, complete mission failure, loss of system security and safety, or extensive financial or social loss"). When the integrity level increases, so too does the intensity and rigor of the required verification and validation tasks.

The V&V process must be independent to avoid conflicts of interest that could lead to catastrophic failure. To this end, IEEE1012 provides specific advice about avoiding such conflicts by requiring technical, managerial, and financial independent verification and validation (IV&V) when testing software and hardware where catastrophic consequences could occasionally occur and where critical consequences will probably occur. Moreover, letting developers certify their own software is a clear conflict of interest, and the IEEE Code of Ethics and the Association for Computing Machinery Code of Ethics are both clear about the obligation of developers to manage competing aims.

IEEE-USA has repeatedly stated that these "high-risk" systems, where catastrophic consequences are occasional or critical consequences are probable, must be independently verified and validated in accordance with IEEE 1012 prior to being deployed and be subject to post-deployment auditing. In a November 2021 letter to NIST, they stated that if high-risk systems have not been independently verified and validated, that "is an indictment of the lack of trustworthiness for these systems and software" and therefore, "neither these systems, nor their results can be considered reliable or trustworthy."

Technical independence ensures there is an objective perspective of the problem. Developers can find it difficult to imagine flaws in their own system. Technical independence encourages people to

search for and investigate when an extreme or outlier possibility occurs. They must be honest about whether the system is proper or not. In sum, the goal is to look for reasons why the system is performing improperly (falsification) not to prove that the system is performing properly. IEEE 1012 "[r]equires the IV&V effort to use personnel who are not involved in the development of the system or its elements. The IV&V effort should formulate its own understanding of the problem and how the proposed system is solving the problem." "Technical independence means that the IV&V effort uses or develops its own set of test and analysis tools separate from the developer's tools." And if sharing tools is necessary, "IV&V conducts qualification tests on tools to assure that the common tools do not contain errors that may mask errors in the system being analyzed and tested." This independence requires the exclusion of parties with a stake in the outcome.

The goal of managerial independence is to ensure that the people performing the V&V are not pressured in any way to reach a certain conclusion about the system's performance – preventing conflicts of interest. IEEE 1012 "[r]equires that the responsibility for the IV&V effort be vested in an organization separate from the development and program management organizations. Managerial independence also means that the IV&V effort independently selects the segments of the software, hardware, and system to analyze and test, chooses the IV&V techniques, defines the schedule of IV&V activities, and selects the specific technical issues and problems to act on." The IV&V effort must be "allowed to submit to program management the IV&V results, anomalies, and findings without any restrictions (e.g., without requiring prior approval from the development group) or adverse pressures, direct or indirect, from the development group."

For its part, financial independence ensures that funding is protected and provided for without being hijacked. IEEE 1012 specifically "[r]equires that control of the IV&V budget be vested in an organization independent of the development organization. This independence prevents situations where the IV&V effort cannot complete its analysis or test or deliver timely results because funds have been diverted or adverse financial pressures or influences have been exerted."

The application of V&V, IV&V, and IEEE 1012 to ensure informed trust has raised some questions that can be addressed here.

- Who can perform IV&V: Those creating, procuring, operating, or using the outputs of A/IS cannot perform IV&V of those systems. Managerial independence works to prevent any conflicts of interest or any pressure on those performing the V&V to produce a particular result. In this situation, these stakeholders do have options. They can pay for independent parties to perform the IV&V or require their developers to do IV&V and then review those results and make a procurement decision based upon those results. Nevertheless, all stakeholders can and should do V&V of their systems in addition to requiring appropriate IV&V. IEEE 1012 provides extensive guidance for organizations to do V&V and is commonly used by developers in a non-independent way to ensure for themselves that their systems are properly designed and implemented.

- Errors: It is impossible for V&V to introduce an error. Only the design and implementation can create errors, and V&V is not design, nor implementation. The worst outcome of IEEE 1012 could be that it did not identify an error that was created by the designer or developer.
- Size of organization and system development process: IV&V should be performed on any high-risk system (as described above) regardless of the size of the organization or the development process used to create hardware or software. IEEE1012 is only focused on the hardware and software and ensuring they are fit for their purpose.
- Intellectual property and access to source code: Access to the software source code is typically necessary for IV&V because black-box testing is not always enough. Entering specific inputs will not always trigger the errors or flaws within the system. Therefore, it is important to also examine the software directly via white box testing where the specific implementation of requirements is examined. For these reasons, IEEE-USA, IEEE Standards Association, and the IEEE Computer Society have stated that (1) "[i]ntellectual property protections should not be used as a shield to prevent duly limited disclosure of information needed to ascertain whether [A/IS] meet acceptable standards of effectiveness, fairness, and safety" and (2) governments "should not procure AI[systems] that... are shielded from independent validation and verification, and public review." IEEE-USA, IEEE Standards Association, and the IEEE Computer Society have also outlined the minimum information to be disclosed when source code is court ordered to be provided to a counterparty. In "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System", Rebecca Wexler argues that criminal trade secret privilege is ahistorical, harmful to defendants, and unnecessary to protect the interests of the secret holder. She concluded that, compared to substantive trade secret law, the privilege overprotects intellectual property and that privileging trade secrets in criminal proceedings fails to serve the theoretical purpose of either trade secret law or privilege law.

10. What are the best definitions of terms frequently used in accountability policies, such as fair, safe, effective, transparent, and trustworthy? Where can terms have the same meanings across sectors and jurisdictions? Where do terms necessarily have different meanings depending on the jurisdiction, sector, or use case?

As noted above, it is often helpful to distinguish value-oriented goals from trust-oriented goals.

In the case of value-oriented goals, which often involve "essentially contested" concepts (such as fairness), general definitions will often be elusive. In such cases, it will be important to use context-informed definitions, allow for multiple metrics for the same value, and be transparent about the criteria and metrics being applied.

In the case of trust-oriented goals, we can draw upon the considerable body of work IEEE has compiled on what makes a system trustworthy. IEEE's Ethically Aligned Design highlights four key trust conditions for AI systems designed to be individually necessary and collectively sufficient; globally applicable but culturally flexible; and capable of being operationalized. The four conditions for which an A/IS can be trusted for adoption and use are:

- **Effectiveness**: Sound empirical evidence can be provided that a system is indeed fit for its intended purpose.
- **Competence**: Creators and operators of the system have specified the skills and knowledge required for its effective operation and have adhered to the creators' competency specifications.
- **Accountability**: Those engaged in the system's design, development, procurement, deployment, operation, and validation maintain clear and transparent lines of responsibility for the outcomes generated and are open to inquiries as may be appropriate.
- **Transparency**: Stakeholders in the results of the system have access to pertinent and appropriate information about its design, development, procurement, deployment, operation, and validation of effectiveness.

The trust conditions offer an actionable framework for designing assessments of trustworthiness, including those focused on value-oriented goals (assuming those goals have been properly defined as described above).

Accountability Subjects

15. The AI value or supply chain is complex, often involving open source and proprietary products and downstream applications that are quite different from what AI system developers may initially have contemplated. Moreover, training data for AI systems may be acquired from multiple sources, including from the customer using the technology. Problems in AI systems may arise downstream at the deployment or customization stage or upstream during model development and data training.

The challenge. Maintaining accountability in AI-enabled systems can be a particularly steep challenge. This is because of both the perceived "black box" nature of AI and the diffusion of responsibility it brings.

The perception of AI as a black box stems from the opacity that is an inevitable characteristic of a system that is a complex nexus of algorithms, computer code, and input data. As observed by Joshua New and Daniel Castro of the Information Technology and Innovation Foundation:

> "The most common criticism of algorithmic decision-making is that it is a "black box" of extraordinarily complex underlying decision models involving millions of data points and thousands of lines of code. Moreover, the model can change over time, particularly when using machine learning algorithms that adjust the model as the algorithm encounters new data."

This opacity of the systems makes it challenging to trace cause to effect, which, in turn, makes it difficult, or even impossible, to draw lines of responsibility.

The diffuseness challenge stems from the fact that even the most seemingly straightforward A/IS can be complex, with a wide range of agents—systems designers, engineers, data analysts, quality control specialists, operators, and others—involved in design, development, and deployment.

Moreover, some of these agents may not even have been engaged in the development of the A/IS in question; they may have, for example, developed open-source components that were intended for an entirely different purpose but that were subsequently incorporated into the A/IS. This diffuseness of responsibility poses a challenge to the maintenance of accountability. As Matthew Scherer, a frequent writer and speaker on topics at the intersection of law and A/IS, observes:

> "The sheer number of individuals and firms that may participate in the design, modification, and incorporation of an AI system's components will make it difficult to identify the most responsible party or parties. Some components may have been designed years before the AI project had even been conceived, and the components' designers may never have envisioned, much less intended, that their designs would be incorporated into any AI system, still less the specific AI system that caused harm. In such circumstances, it may seem unfair to assign blame to the designer of a component whose work was far-removed in both time and geographic location from the completion and operation of the AI system."

As a result of the challenges presented by the opacity and diffuseness of responsibility in A/IS, the present-day answer to the question "Who is accountable?" is, in far too many instances, "It's hard to say." This is a response that, in practice, means "no one" or, equally unhelpful, "everyone." Such failure to maintain accountability will undermine efforts to bring A/IS (and all their potential benefits) into legal systems based on informed trust.

Meeting the challenge. Although maintaining accountability in complex systems can be a challenge, it is one that must be met in order to engender informed trust in the use of A/IS in the legal domain. "Blaming the algorithm" is not a substitute for taking on the challenge of maintaining transparent lines of responsibility and establishing norms of accountability. This is true even if we allow that, given the complexity of the systems in question, some number of "systems accidents" are inevitable. Informed trust in a system does not require a belief that zero errors will occur; however, it does require a belief that there are mechanisms in place for addressing errors when they do occur. Accountability is an essential component of those mechanisms.

In meeting the challenge, it should be recognized that there are existing norms and controls that have a role to play in ensuring that accountability is maintained. For example, contractual arrangements between an A/IS provider and a party acquiring and applying a system may help to specify who is (and is not) to be held liable in the event the system produces undesirable results. Professional codes of ethics may also go some way toward specifying the extent to which lawyers, for example, are responsible for the results generated by the technologies they use, whether they operate them directly or retain someone else to do so. Judicial systems may have procedures for assessing responsibility when a citizen's rights are improperly infringed. As illustrated by the cases described above, however, existing norms and controls, while helpful, are insufficient in themselves to meet the specific challenge represented by the opacity and diffuseness of A/IS. To meet the challenge further steps must be taken.

The first step is ensuring that all those engaged in the creation, procurement, deployment,

operation, and testing of A/IS recognize that, if accountability is not maintained, these systems will not be trusted. In the interest of maintaining accountability, these stakeholders should take steps to clarify lines of responsibility throughout this continuum, and make those lines of responsibility, when appropriate, accessible to meaningful inquiry and audit.

The goal of clarifying lines of responsibility in the operation of A/IS is to implement a governing model that specifies who is responsible for what, and who has recourse to which corrective actions (i.e., a trustworthy model that ensures that it will admit actionable answers should questions of accountability arise). Arriving at an effective model will require the participation of those engaged in the creation and operation of A/IS, those affected by the results of their use, and those with the expertise to understand how such a model would be used in a given legal system. For example:

- Individuals responsible for the design of A/IS will have to maintain a transparent record of the sources of the various components of their systems, including identification of which components were developed in-house and which acquired from outside sources (whether open source or acquired from another firm).
- Individuals responsible for the design of A/IS will have to specify the roles, responsibilities, and potential subsequent liabilities of those who will be engaged in the operation of the systems they create.
- Individuals responsible for the operation of a system will have to understand their roles, responsibilities, and potential liabilities, and will have to maintain documentation of their adherence to requirements.
- Individuals affected by the results of the operation of A/IS (e.g., a defendant in a criminal proceeding) will have to be given access to information about the roles and responsibilities of those involved in relevant aspects of the creation, operation, and validation of the effectiveness of the A/IS affecting them.
- Individuals with legal and political training (e.g., jurists, regulators, as well as legal and political scholars) will have to ensure that any model that is created will provide information that is in fact actionable within the operative legal system.

A governing model of accountability that reflects the interests of all these stakeholders will be more effective both at deterring irresponsible design or use of A/IS before it happens and at apportioning responsibility for an undesirable outcome when it does happen.

Pulling together the input from the various stakeholders will likely not take place without some amount of institutional initiative. Organizations that employ A/IS for accomplishing legal tasks—private firms, regulatory agencies, law enforcement agencies, judicial institutions—should therefore develop and implement policies that will advance the goal of clarifying lines of responsibility. Such policies could take the form of, for example, designating an official specifically charged with oversight of the organization's procurement, deployment, and evaluation of A/IS as well as the organization's efforts to educate people both inside and outside the organization on its use of A/IS. Such policies might also include the establishment of a review board to assess the organization's use of A/IS and to ensure that lines of responsibility for the outcomes of its use are maintained. In

the case of agencies, such as police departments, whose use of A/IS could impact the general public, such review boards would, in the interest of legitimacy, have to include participation from various citizens' groups, such as those representing defendants in the criminal system as well as those representing victims of crime.

The goal of opening lines of responsibility to meaningful inquiry is to ensure that an investigation into the use of A/IS will be able to isolate responsibility for errors (or potential errors) generated by the systems and their operation. This means that all those engaged in the design, development, procurement, deployment, operation, and validation of the effectiveness of A/IS, as well as the organizations that employ them, must in good faith be willing to participate in an audit (whether the audit is a formal legal investigation or a less formal inquiry) and to create and preserve documentation of key procedures, decisions, certifications, and tests made in the course of developing and deploying the A/IS.

The combination of a governing model of accountability and an openness to meaningful audit will allow the maintenance of accountability, even in complex deployments of A/IS in the service of a legal system.

a. Where in the value chain should accountability efforts focus?

[See above.]

b. How can accountability efforts at different points in the value chain best be coordinated and communicated?

[See above.]

c. How should vendors work with customers to perform AI audits and/or assessments? What is the role of audits or assessments in the commercial and/or public procurement process? Are there specific practices that would facilitate credible audits (e.g., liability waivers)?

Governments should set procurement and contracting requirements that encourage parties seeking to use A/IS in the conduct of business with or for the government, particularly with or for the court system and law enforcement agencies, to adhere to the principles of effectiveness, competence, accountability, and transparency as described in this document. This can be achieved through legislation or administrative regulation. All government efforts in this regard should be transparent and open to public scrutiny.

Professionals engaged in the practice, interpretation, and enforcement of the law (such as lawyers, judges, and law enforcement officers), when engaging with or relying on providers of A/IS technology or services, should require, at a minimum, that those providers adhere to, and be able to demonstrate adherence to, the principles of effectiveness, competence, accountability, and transparency as described in this document. Likewise, those professionals, when operating A/IS themselves, should adhere to, and be able to demonstrate adherence to, the principles of

effectiveness, competence, accountability, and transparency. Demonstrations of adherence to the requirements should be publicly accessible.

When negotiating contracts for the provision of A/IS products and services for use in the legal system, providers and buyers of A/IS should include contractual terms specifying clear lines of responsibility for the outcomes of the systems being acquired.

d. Since the effects and performance of an AI system will depend on the context in which it is deployed, how can accountability measures accommodate unknowns about ultimate downstream implementation?

Unknowns, which we assume to be defined as failures or risks, may be addressed via an appropriate application of IV&V; more specifically using IEEE 1012 to assess the risk levels of AI/S and possible unknown outcomes. Appropriate application of IV&V requires that each software and hardware component be assigned an integrity level that increases depending on the likelihood and consequences of a failure: negligible, marginal, critical (causing "major and permanent injury, partial loss of mission, major system damage, or major financial or social loss"), and catastrophic (causing "loss of human life, complete mission failure, loss of system security and safety, or extensive financial or social loss"). When the integrity level increases, so too does the intensity and rigor of the required verification and validation tasks.

16. The lifecycle of any given AI system or component also presents distinct junctures for assessment, audit, and other measures. For example, in the case of bias, it has been shown that "[b]ias is prevalent in the assumptions about which data should be used, what AI models should be developed, where the AI system should be placed—or if AI is required at all." [82] How should AI accountability mechanisms consider the AI lifecycle?

An effective governance model should cover the entire AI lifecycle, from design through retirement and data disposal.

Responses could address the following:

a. Should AI accountability mechanisms focus narrowly on the technical characteristics of a defined model and relevant data? Or should they feature other aspects of the socio- technical system, including the system in which the AI is embedded? [83] When is the narrower scope better and when is the broader better? How can the scope and limitations of the accountability mechanism be effectively communicated to outside stakeholders?

As noted above (under Question 5), while there are times when a narrow focus is appropriate (e.g., when researching the impact of specific changes in process), ultimately an assessment of a system's trustworthiness requires a broad scope, taking into account the whole sociotechnical system (including operators) of which the technology is a part.

This broader scope introduces an essential, but often overlooked, component of the trustworthiness

of an AI-enabled system: **operator competence.**

An essential component of informed trust in a technological system, especially one that may affect us in profound ways, is confidence in the competence of the operator(s) of the technology. We trust surgeons or pilots with our lives because we have confidence that they have the knowledge, skills, and experience to apply the tools and methods needed to carry out their tasks effectively. We have that confidence because we know that these operators have met rigorous professional and scientific accreditation standards before being allowed to step into the operating room or cockpit. This informed trust in operator competence is what gives us confidence that surgery or air travel will result in the desired outcome. No such standards of operator competence currently exist with respect to A/IS, where the life, liberty, and rights of citizens can be at stake. That absence of standards hinders the trustworthy adoption of A/IS.

The human operator is an integral component of A/IS. Almost all current applications of A/IS in legal systems, like those in most other fields, require human mediation and likely will continue to do so for the near future. This human mediation, post design and post development, will take a number of forms, including decisions about (a) whether or not to use A/IS for a given purpose, (b) the data used to train the systems, (c) settings for system parameters to be used in generating results, (d) methods of validating results, (e) interpretation and application of the results, and so on. Because these systems' outcomes are a function of all their components, including the human operator(s), their effectiveness, and by extension trustworthiness, will depend on their human operator(s).

Despite this, there are few standards that specify how humans should mediate applications of A/IS in legal systems, or what knowledge qualifies a person to apply A/IS and interpret their results. This reality is especially troubling for the instances in which the life, rights, or liberty of humans are at stake. Today, while professional codes of ethics for lawyers are beginning to include among their requirements an awareness and understanding of technologies with legal application, the operators of A/IS in legal systems are essentially deemed to be capable of determining their own competence: lawyers or IT professionals operating in civil discovery, correctional officers using risk assessment algorithms, and law enforcement agencies engaging in predictive policing or using automated surveillance technologies. All are mostly able to use A/IS without demonstrating that they understand the operation of the system they are using or that they have any particular set of consensus competencies.

The lack of competency requirements or standards undermines the establishment of informed trust in the use of A/IS in legal systems. If courts, legal practitioners, law enforcement agencies, and the general public are to rely on the results of A/IS when applied to tasks traditionally carried out by legal professionals, they must have grounds for believing that those operating A/IS will possess the requisite knowledge and skill to understand the conditions and methods for operating the systems effectively, including evaluating the data on which the A/IS trained, the data to which they are applied, the results they produce, and the methods and results of measuring the effectiveness the systems. Applied incompetently, A/IS could produce the opposite intended effect. Instead of

improving a legal system (and bringing about the gains in well-being that follow from such improvements), they may undermine both the fairness and effectiveness of a legal system and trust in its fairness and effectiveness, creating conditions for social disorder and the deterioration of human well-being that would follow from that disorder. More generally, without the confidence that A/IS operators will apply the technology as intended and supervise it appropriately, the public will harbor fear, uncertainty, and doubt about the use of A/IS in legal systems and potentially about the legal systems themselves.

Fostering informed trust in the competence of human operators. If negative outcomes such as those just described are to be avoided, it will be necessary to include among norms for the adoption of A/IS in a legal system a provision for building informed trust in the operators of A/IS. Building trust will require articulating standards and best practices for two groups of agents involved in the deployment of A/IS: creators and operators.

On the one hand, those engaged in the design, development, and marketing of A/IS must commit to specifying the knowledge, skills, and conditions required for the safe, ethical, and effective deployment and operation of the systems. On the other hand, those engaged in operating the systems, including both legal professionals and experts acting in the service of legal professionals, must commit to adhering to these requirements in a manner consistent with other operative legal, ethical, and professional requirements. The precise nature of the competency requirements will vary with the nature and purpose of the A/IS and what is at stake in their effective operation. The requirements for the operation of A/IS designed to assist in the creation of contracts, for example, might be less stringent than those for the operation of A/IS designed to assess flight risk, which could affect the liberty of individual citizens.

A corollary of these provisions is that education and training in the requisite skills should be available and accessible to those who would operate A/IS, whether that training is provided through professional schools, such as law school; through institutions providing ongoing professional training, such as, for federal judges in the United States, the Federal Judicial Center; through professional and industry associations, such as the American Bar Association; or through resources accessible by the general public. Making sure such training is available and accessible will be essential to ensuring that the resources needed for the competent operation of A/IS are widely and equitably distributed.

It will take a combined effort of both creators and operators to ensure both that A/IS designed for use in legal systems are properly applied and that those with a stake in the effective functioning of legal systems—including legal professionals, of course, but also decision subjects, victims of crime, communities, and the general public—will have informed trust, or, for that matter, informed distrust (if that is what a competence assessment finds) in the competence of the operators of A/IS as applied to legal problems and questions.

b. How should AI audits or assessments be timed? At what stage of design, development, and deployment should they take place to provide meaningful accountability?

A dynamic model, in which closer scrutiny and more frequent assessments are triggered either by changes to a system or by observed errors in output and more relaxed scrutiny and less frequent assessments are permitted after a defined succession of successful assessments, may be helpful in many cases.

c. How often should audits or assessments be conducted, and what are the factors that should inform this decision? How can entities operationalize the notion of continuous auditing and communicate the results?

See above.

18. Should AI systems be released with quality assurance certifications, especially if they are higher risk?

With the pervasive use and the dependence on AI-based systems, the quality of these systems becomes essential for their practical usage and should be considered in the context of the various risk levels. However, quality assurance for AI-based systems is an emerging area that has not been well explored. Quality assurance can be applied to various stages in AI systems' development and deployment, ranging from data being used to assessing ethics of the systems to help protect, differentiate, and grow product adoption. On the latter, IEEE CertifAIEd(™), a certification program that provides guidance, assessment and independent verification of AI systems and offers the ability to scale responsible innovation implementations, thereby helping to increase the quality of AI Systems, the associated trust with key stakeholders, and realizing associated benefits stands as an example. It is based on four criteria:

- **Transparency criteria** relate to values embedded in a system design, and the openness and disclosure of choices made for development and operation.
- **Accountability criteria** recognize that the system/service autonomy and learning capacities are the results of algorithms and computational processes designed by humans and organizations that remain responsible for their outcomes.
- **Algorithmic bias criteria** relate to the prevention of systematic errors and repeatable undesirable behaviors that create unfair outcomes.
- **Privacy criteria** are aimed at respecting the private sphere of life and public identity of an individual, group, or community, upholding dignity.

19. As governments at all levels increase their use of AI systems, what should the public expect in terms of audits and assessments of AI systems deployed as part of public programs? Should the accountability practices for AI systems deployed in the public sector differ from those used for private sector AI? How can government procurement practices help create a productive AI accountability ecosystem?

See above, under Question 15(c).

Accountability Inputs and Transparency

20. What sorts of records (e.g., logs, versions, model selection, data selection) and other documentation should developers and deployers of AI systems keep in order to support AI accountability? How long should this documentation be retained? Are there design principles (including technical design) for AI systems that would foster accountability-by-design?

Access to meaningful information. An essential component of informed trust in a technological system is confidence that the information required for a human to understand why the system behaves a certain way in a specific circumstance (or would behave in a hypothetical circumstance) will be accessible. Without transparency, there is no basis for trusting that a given decision or outcome of the system can be explained, replicated, or, if necessary, corrected. Without transparency, there is no basis for informed trust that the system can be operated in a way that achieves its ends reliably and consistently or that the system will not be used in a way that impinges on human rights. In the case of A/IS applied in a legal system, such a lack of trust could undermine the credibility of the legal system itself.

Transparency and trust. Transparency, by prioritizing access to information about the operation and effectiveness of A/IS, serves the purpose of fostering informed trust in the systems. More specifically, transparency fosters trust that:

- the operation of A/IS and the results they produce are explainable;
- the operation and results of A/IS are fair;
- the operation and results of A/IS are unbiased;
- the A/IS meet normative standards for operation and results;
- the A/IS are effective;
- the results of A/IS are replicable; and
- those engaged in the design, development, procurement, deployment, operation, and validation of effectiveness of A/IS can be held accountable, where appropriate, for negative outcomes, and that corrective or punitive action can be taken when warranted.

This trust can be fostered by transparency.

The elements of transparency. Transparency of A/IS requires disclosing information about the design and operation of the A/IS to various stakeholders. In implementing the principle, however, we must, in the interest of both feasibility and effectiveness, be more precise both about the categories of stakeholders to whom the information will be disclosed, and about the categories of information that will be disclosed to those stakeholders.

To take the example of AI deployed in the legal system, relevant stakeholders would include those who:

- operate A/IS for the purpose of carrying out tasks in civil justice, criminal justice, and law enforcement (such as a law enforcement officer who uses facial recognition tools to identify potential suspects);

- rely on the results of A/IS to make important decisions (such as a judge who draws on the results of an algorithmic assessment of recidivism risk in deciding on a sentence);
- are directly affected by the use of A/IS—a "decision subject" (such as a defendant in a criminal proceeding whose bail terms are influenced by an algorithmic assessment of flight risk);
- are indirectly affected by the results of A/IS (such as the members of a community that receives more or less police attention because of the results of predictive policing technology); and
- have an interest in the effective functioning of the legal system (such as judges, lawyers, and the public).

Different types of relevant information can be grouped into high-level categories. As illustrated below, a taxonomy of such high-level categories may, for example, distinguish between:

- nontechnical procedural information regarding the employment and development of a given application of A/IS;
- information regarding data involved in the development, training, and operation of the system;
- information concerning a system's effectiveness/performance;
- information about the formal models that the system relies on; and
- information that serves to explain a system's general logic or specific outputs.

These more granular distinctions matter because different sorts of inquiries will require different sorts of information and it is important to match the information provided to the actual needs of the inquiry. For example, an inquiry into a predictive policing system that misdirected police resources may not be much advanced by information about the formal models on which the system relied, but it may well be advanced by an explanation for the specific outcome. On the other hand, an inquiry, undertaken by a designer or operator, into ways to improve system performance may benefit from access to information about the formal models on which the system relies.

These distinctions also matter because there may be circumstances in which it would be desirable to limit access to a given type of information to certain stakeholders. For example, there may be circumstances in which one would want to identify an agent to serve as a public interest steward. For auditing purposes, this individual would have access to certain types of sensitive information unavailable to others. Such restrictions on information access are necessary if the transparency principle is not to impinge on other societal values and goals (such as security, privacy, and appropriate protection of intellectual property).

Transparency in practice. As just noted, although transparency can foster informed trust in A/IS applied in a legal system, its practical implementation requires careful thought. Requiring public access to all information pertaining to the operation and results of A/IS is neither necessary nor feasible. What is required is a careful consideration of who needs access to what information for the specific purpose of building informed trust.

When it comes to deciding whether a specific type of information should be made available and, if so, which types of stakeholders should have access to it, there are various considerations, for example:

- The release of certain types of information may conflict with data privacy concerns, commercial or public policy interests (such as the promotion of innovation through appropriate intellectual property protections), and security interests (e.g., concerns about gaming and adversarial attacks). At the same time, such competing interests should not be permitted to be used, without specific justification, as a blanket cover for not adhering to due process, transparency, or accountability standards. The tension between these interests is particularly acute in the case of A/IS applied in a legal system, where the dignity, security, and liberty of individuals are at stake.
- There is tension between the specific goal of explainability (which may argue for limits on system complexity) and system performance (which may be served by greater complexity, to the detriment of explainability).
- One must carefully consider the question that is being asked in an inquiry into A/IS and what information transparency can produce to answer that question. Disclosure of A/IS algorithms or training data is, itself, insufficient to enable an auditor to determine whether the system was effective in a specific circumstance. By analogy, transparency into drug manufacturing processes does not, itself, provide information about the actual effectiveness of a drug. Clinical trials provide that insight. In a legal system, an excessive focus on transparency-related information-gathering and assessment may overwhelm courts, legal practitioners, and law enforcement agencies. Meanwhile, other factors, such as measurement of effectiveness or operator competence, coupled with information on training data, may often suffice to ensure that there is a well-informed basis for trusting A/IS in a given circumstance.

Given these competing considerations, arriving at a balance that is optimal for the functioning of a legal system and that has legitimacy in the eyes of the public will require an inclusive dialogue, bringing together the perspectives of those with an immediate stake in the proper functioning of a given technology, including those engaged in the design, development, procurement, deployment, operation, and validation of effectiveness of the technology, as well as those directly affected by the results of the technology; the perspectives of communities that may be indirectly impacted by the technology; and the perspectives of those with specialized expertise in ethics, government, and the law, such as jurists, regulators, and scholars. How the competing considerations should be balanced will also vary from one circumstance to another. Rather than aiming for universal transparency standards that would be applicable to all uses of A/IS, transparency standards should allow for circumstance-dependent flexibility, in the context of the four constitutive components of trust (effectiveness, competence, accountability, transparency).

21. What are the obstacles to the flow of information necessary for AI accountability either within an organization or to outside examiners? What policies might ease researcher and other third-party access to inputs necessary to conduct AI audits or assessments?

As noted above, pulling together the input from the various stakeholders will likely not take place without some amount of institutional initiative. Organizations that employ A/IS should therefore develop and implement policies that will advance the goal of clarifying lines of responsibility. Such policies could take the form of, for example, designating an official specifically charged with oversight of the organization's procurement, deployment, and evaluation of A/IS as well as the organization's efforts to educate people both inside and outside the organization on its use of A/IS. Such policies might also include the establishment of a review board to assess the organization's use of A/IS and to ensure that lines of responsibility for the outcomes of its use are maintained. In the case of agencies, such as police departments, whose use of A/IS could impact the general public, such review boards would, in the interest of legitimacy, have to include participation from various citizens' groups, such as those representing defendants in the criminal system as well as those representing victims of crime.

23. How should AI accountability "products" (e.g., audit results) be communicated to different stakeholders? Should there be standardized reporting within a sector and/or across sectors? How should the translational work of communicating AI accountability results to affected people and communities be done and supported?

See above (on Question 20).

Barriers to Effective Accountability

24. What are the most significant barriers to effective AI accountability in the private sector, including barriers to independent AI audits, whether cooperative or adversarial? What are the best strategies and interventions to overcome these barriers?

As noted above (Questions 15 and 21), a key challenge is convincing stakeholders (including designers, developers, and operators) that, in the interest of building trust in the systems they develop and operate, it is important that they design and participate in an effective governing model of accountability, and that participation in such a model will include an openness to independent assessments and audits. As also noted above, convincing stakeholders of this point may take some amount of institutional initiative.

27. What is the role of intellectual property rights, terms of service, contractual obligations, or other legal entitlements in fostering or impeding a robust AI accountability ecosystem? For example, do nondisclosure agreements or trade secret protections impede the assessment or audit of AI systems and processes? If so, what legal or policy developments are needed to ensure an effective accountability framework?

As noted above (Question 9), access to the software source code is typically necessary for IV&V because black-box testing is not always enough. Entering specific inputs will not always trigger the errors or flaws within the system. Therefore, it is important to also examine the software directly via white box testing where the specific implementation of requirements is examined. For these reasons, IEEE-USA, IEEE Standards Association, and the IEEE Computer Society have stated that

(1) “[i]ntellectual property protections should not be used as a shield to prevent duly limited disclosure of information needed to ascertain whether [A/IS] meet acceptable standards of effectiveness, fairness, and safety” and (2) governments ``should not procure AI[systems] that... are shielded from independent validation and verification, and public review.” IEEE-USA, IEEE Standards Association, and the IEEE Computer Society have also outlined the minimum information to be disclosed when source code is court ordered to be provided to a counterparty. Moreover, in “Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System ", Rebecca Wexler argues that criminal trade secret privilege is ahistorical, harmful to defendants, and unnecessary to protect the interests of the secret holder. She concluded that, compared to substantive trade secret law, the privilege overprotects intellectual property and that privilegeing trade secrets in criminal proceedings fails to serve the theoretical purpose of either trade secret law or privilege law.

29. How does the dearth of measurable standards or benchmarks impact the uptake of audits and assessments?

As noted above, a key condition of trust in an AI-enabled system is evidence of its effectiveness and a key mechanism for gathering that evidence of effectiveness is benchmarking exercises. In the absence of sound benchmarks (and the quantitative evidence (metrics) they generate, it is difficult, if not impossible, to implement an effective governance model for AI-enabled systems.

AI Accountability Policies

30. What role should government policy have, if any, in the AI accountability ecosystem? For example:

d. What accountability practices should government (at any level) itself mandate for the AI systems the government uses?

Government agencies should require that the AI-enabled systems they employ are open to assessment for adherence to the key trust conditions of effectiveness, competence, accountability, and transparency.

31. What specific activities should government fund to advance a strong AI accountability ecosystem?

Governments can advance the cause of accountability by funding benchmarking initiatives that provide stakeholders, and the public, with scientifically sound measures of the effectiveness of AI-enabled systems when applied to meet real-world objectives in real-world conditions.

32. What kinds of incentives should government explore to promote the use of AI accountability measures?

Governments should set procurement and contracting requirements that encourage parties seeking to use A/IS in the conduct of business with or for the government, particularly with or for the court

system and law enforcement agencies, to adhere to the principles of effectiveness, competence, accountability, and transparency as described in this document. This can be achieved through legislation or administrative regulation. All government efforts in this regard should be transparent and open to public scrutiny.

Further, Regulators should permit insurers to issue professional liability and other insurance policies that consider whether the insured (either a provider or operator of A/IS in a legal system) adheres to the principles of effectiveness, competence, accountability, and transparency (as they are articulated in this document).

Further reading:

Trustworthy Evidence For Trustworthy Technology: An Overview of Evidence for Assessing the Trustworthiness of Autonomous and Intelligent Systems (The Law Committee of the IEEE Global Initiative on Ethics of Autonomous and Intelligence Systems, and the IEEE-USA AI Policy Committee, September 2022. https://ieeeusa.org/committees/aipc/#:~:text=Trustworthy%20Evidence%20For%20Trustworthy%20Technology%3A%20An%20Overview%20of%20Evidence%20for%20Assessing%20the%20Trustworthiness%20of%20Autonomous%20and%20Intelligent%20Systems)

Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems https://standards.ieee.org/industry-connections/ec/ead-v1/)