TRUSTWORTHY EVIDENCE FOR TRUSTWORTHY TECHNOLOGY

An Overview of Evidence for Assessing the Trustworthiness of Autonomous and Intelligent Systems

AUTHORS

Jeanna Matthews (Co-Chair IEEE-USA AI Policy Committee)
Bruce Hedin (Member, IEEE Law Committee)
Marc Canellas (Former Chair, IEEE-USA AI Policy Committee)

Law Committee of the IEEE Global Initiative
and IEEE-USA AI Policy Committee

IEEE
USA

INTRODUCTION

Autonomous and intelligent systems (A/IS) are increasingly being 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. In addition, these systems are able to influence serious large-scale societal governance decisions via the amplification of news, capital flows, and allocation of public resources. The question of A/IS trustworthiness is of crucial importance. If society cannot trust the decisions enabled by A/IS, 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 key components of the social order 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 need to 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 in order 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?* In this paper, we explore these three components of trustworthiness, with a particular focus on the latter: the evidence required for establishing the trustworthiness of a system.

To have a practical impact, however, we need to 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.

In this paper, we explore these questions, with a particular eye to practical measures that can be taken to allow our societies to answer questions about the trustworthiness of A/IS on the basis of sound evidence. More specifically, the paper begins with an overview of the three-tier framework (values, trust, and evidence) for ensuring that the adoption and use of A/IS protects and advances our societies' core values. From that point on, the focus of the paper is on the third tier, evidence. In discussing evidence, we begin with a historical review of some of the standards that have been adopted for assessing the reliability of evidence and then turn to a discussion of the specific features evidence should have if it is to establish the trustworthiness of A/IS and of the types of evidence that might meet those requirements. We follow up our discussion of the types of evidence and their features with a discussion of a practically feasible, and already available, standard for gathering and producing the evidence required to establish the trustworthiness of A/IS: IEEE 1012.

A FRAMEWORK FOR ETHICS, INFORMED TRUST, AND EVIDENCE

Suppose we were asked to comment on the advisability of adopting an A/IS that provides a decision-maker with guidance as to whether an applicant for public benefits is entitled to those benefits. Our initial focus in procuring such a decision-making tool would be: Can we trust the system? Or, more specifically, can we trust that it will provide sound guidance that is well-adapted to the specific circumstances of any given applicant?

Within the realm of trust there is a tangle of component questions. Some of these are questions about ends, both the ends for which the system was designed and deployed. *Was the system designed for the purpose to which it is being put?* (Or to expand that question: What are the requirements we expect such a system to meet? If we want to require that it be accurate, what definition of accuracy is used and what level and distribution of errors is considered acceptable? If we want to require that it render decisions in accordance with the law, what evidence will be needed for meeting this requirement? What other norms or values do we expect the system to adhere to? That it be unbiased? That it be fair? If fair, on what definition of fairness?)

Another line of questioning will cause us to examine the grounds for trust. *What are the traits of an A/IS to which we would give our trust?* (Or to expand that question: Put another way, what do we need to know about the system in order to trust it? That it has been designed and developed in accordance with a certain protocol? That it has been trained on representative data? That it has been tested and found to be effective? That those operating the system have the skills and experience required to operate it as intended? That lines of accountability for the outcomes of the system are clearly drawn?)

Still another line of questions raises the issue of evidence. *If a precondition of trusting a system is knowing (or having solid grounds for believing) that it is effective, how do we know that it is (or is not) effective?* (Or to expand that question: By past performance? By testing? What kind of testing? By what metrics? By whom? If a precondition of trusting a system is knowing (or having solid grounds for believing) that its operators are competent, how do we determine whether the operators in question meet that requirement? By academic or professional credentials? By specifications set by the designers of the system? By prior experience? By testing?)

If we wish to ensure that our assessment of the trustworthiness of an A/IS is coherent, thorough, and open to inspection, it is necessary to disentangle these trust-related questions and give a well-focused and reasoned response. In order to provide a mechanism for doing so, we propose a three-tier analytical framework for addressing the considerations raised by the question of trust in the creation and operation of a decision-making tool or services embodying A/IS. In other words, this framework is meant to support finding the level of evidence required to assess a system's trustworthiness.

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

Providing stakeholders with the basis for informed trust in an A/IS requires that we adequately address the questions on all three tiers and do so at every stage of an A/IS's lifecycle. Accordingly, these questions should frame the work of all agents engaged in the design and deployment of a system: designers, developers, procurement officers, operators, testers, regulators, and even those charged with retiring a system.

The three tiers, their intent, and the IEEE resources available for addressing them are represented in the following table. In this paper, we focus on Tier 3, Evidence, but address this question in the context of Tier 1 and 2.

	Exemplar Question	IEEE Principles and Protocols	Ensures the creation and operation of A/IS that
TIER 1: Ethics and Values	What is the purpose of this technology? Does it adhere to the necessary values and ethical considerations?	• IEEE EAD's principles of human rights, well-being, data agency, and awareness of misuse. • IEEE 7000 series Standards[1]	Uphold the principles of human rights, well-being, data agency, and awareness of misuse.
TIER 2: Trust Conditions	Under what conditions can a stakeholder in the output of a system trust that the technology adheres to the values identified in Tier 1?	• IEEE EAD's principles of effectiveness, competence, accountability and transparency.	Uphold the ethics and values established by creators and operators, as well as other stakeholders in the outcomes generated by a system.
TIER 3: Evidence	What evidence is available for demonstrating that a system meets the trust conditions identified in Tier 2?	• IEEE Standard 1012: System, Software, and Hardware Verification and Validation. • IEEE CertifAIEd Methodology	Meet the conditions for an informed trust in the responsible use of an A/IS.

TIER1 THE QUESTION OF ETHICS: ANCHORING THE DESIGN, ADOPTION, AND USE OF AI IN UNIVERSAL VALUES

The Tier 1 analysis focuses on questions of values and ethics. In *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems First Edition* (EAD), IEEE identified three key pillars: 1) universal human values (respecting human rights, aligning with human values, and holistically increasing well-being), 2) political self-determination

[1] IEEE P7000 Standards include a range of standards in the range P7000-P7015 that all focus on Ethics in Action in Autonomous and Intelligent Systems (https://ethicsinaction.ieee.org/p7000/). They include: IEEE P7000 "Model Process for Addressing Ethical Concerns During System Design", IEEE P7001 "Transparency of Autonomous Systems", IEEE P7002 "Data Privacy Process", IEEE P7003 "Algorithmic Bias Considerations", IEEE P7004 "Standard on Child and Student Data Governance", IEEE P7004.1 "Recommended Practices for Virtual Classroom Security, Privacy and Data Governance", IEEE P7005 "Standard on Employer Data Governance", IEEE P7007 "Ontological Standard for Ethically driven Robotics and Automation Systems", IEEE P7008 "Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems", IEEE P7009 "Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems", IEEE Std 7010 "IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being", IEEE P7010.1 "Recommended Practice for Environmental Social Governance (ESG) and Social Development Goal (SDG) Action Implementation and Advancing Corporate Social Responsibility", IEEE P7011 "Standard for the Process of Identifying & Rating the Trust-worthiness of News Sources", IEEE P7012 "Standard for Machine Readable Personal Privacy Terms", IEEE P7014 "Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems", and IEEE P7015 "Standard for Data and Artificial Intelligence (AI) Literacy, Skills, and Readiness".

and data agency (nurturing political freedom and democracy, in accordance with the cultural precepts of individual societies), and 3) technical dependability (reliably, safely, and actively accomplishing the objectives and values for which they were created)[2]. Going beyond technical features, these pillars speak to the processes of society, policy, and lawmaking. They are the ethical grounding upon which A/IS are created and operated.

From these pillars, they derive additional principles that should guide ethical and values-based design, development, and implementation of A/IS. Among them are four ethical principles that are especially relevant here: *human rights*, *well-being*, *data agency*, and *awareness of misuse*.

- **Human rights**: An A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
- **Well-being**: A/IS creators shall adopt increased human well-being as a primary success criterion for development.
- **Data Agency**: A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
- **Awareness of misuse**: A/IS creators shall guard against all potential misuses and risks of A/IS in operation.

Creating and operating A/IS in accordance with these ethical principles is needed to make society more equitable, inclusive, and just; making operations more transparent and accountable; and encouraging public participation and increasing the public's trust in the organizations and institutions that use A/IS.[3] Accordingly, organizations and institutions are encouraged to procure and operate only AI systems that adhere to the principles in IEEE's Ethically Aligned Design for creating and operating A/IS that further human values and ensure trustworthiness.[4]

TIER2 THE CONDITIONS FOR TRUST: THE TRAITS REQUIRED OF A TRUSTWORTHY SYSTEM

Having considered the question of what values and ethical considerations guide the purpose of the technology, we can now turn to what specific attributes would make an A/IS trustworthy. Said differently, what conditions must be met if stakeholders of a system are to have an informed trust that it will in fact meet its intended purpose (with "purpose" understood in both narrow and broad terms)? If those attributes or trust conditions are identifiable, we will be in a position to take inventory of the evidence that can serve as indicators for the presence of a sought-for attribute or that a trust condition is met.

Fortunately, when it comes to identifying the attributes that make A/IS worthy of trust, IEEE has made considerable progress. EAD highlights four key trust conditions for AI systems designed to be individually necessary and collectively sufficient; globally applicable but culturally flexible;

[2] EAD, p. 10-12.

[3] IEEE-USA, "Artificial Intelligence: Accelerating Inclusive Innovation By Building Trust", p. 1, 3. (Available at: https://ieeeusa.org/assets/public-policy/positions/ai/AITrust0720.pdf)

[4] IEEE-USA, "Artificial Intelligence: Accelerating Inclusive Innovation By Building Trust", p. 6. (Available at: https://ieeeusa.org/assets/public-policy/positions/ai/AITrust0720.pdf)

and capable of being operationalized.[5] 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.

TIER3 ON EVIDENCE: DEMONSTRATING THAT A TECHNOLOGY MEETS THE CONDITIONS REQUIRED FOR TRUST

Having identified the conditions an A/IS must meet if it is to be trusted, we turn to the practical question of evidence, specifically the evidence necessary to assess whether a system in fact meets the trust conditions. On this level of analysis we again meet the question of trust. However, this time it is not a matter of the traits that makes technology trustworthy, but rather of the traits that make evidence trustworthy including what types of data can be gathered that will be meaningful, accurate, and practically viable indicators of whether a technology has met a given trust condition. Here we ask 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).

[5] Bar, G., Wiktorzak, G., and Matthews, J., "Four Conditions for Building Trusted AI Systems," 13 July 2021. (Available at: https://news.bloomberglaw.com/us-law-week/insight-four-principles-for-the-trustworthy-adoption-of-ai-in-legal-systems)

LEGAL AND HISTORICAL BACKGROUND OF TRUSTWORTHY EVIDENCE

In this section, we begin with an historical consideration of what makes evidence trustworthy. Subsequently, we take some inspiration from a US legal context including the *Frye* and *Daubert* standards as well as a 2016 President's Council of Advisors on Science and Technology (PCAST) report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods."

A Historical Perspective

It is helpful to put the question of what makes evidence trustworthy in a broader historical perspective. In other words, what type of information is acceptable evidence of something we don't (and, in some cases, cannot) know ourselves? What type of evidence can be introduced in court? What is the role of technology and statistics?

The matter of determining what evidence is required before trusting someone or something that you cannot personally verify is not new. Especially in the context of legal proceedings, people have been grappling for centuries with what to characterize as sufficient and trustworthy evidence. Before tackling this issue in the context of AI and automated systems, we can look to the evolution of evidence in the legal system for some important guidance.

Some early, and now much discredited, tests of the soundness of evidence could be classified as enabling God to weigh in directly on the matter of truth. Practices used, for example, in the Salem witch trials to test whether someone was a witch relied on the idea that if the person was telling the truth about not being a witch, God would save them from drowning or other demise. At times, monarchs or other rules were viewed as God's representative in assessing justice and truth. More recently, trial by jury has replaced a single monarch or judge with the collective will of a group of one's peers. In both cases, human judges are often assessing the trustworthiness of the testimony of others and may address questions of reliability, character, competence and potential bias of a witness. The obvious danger of the intrusion of bias in such a system is clear. For example, in some legal frameworks, the social status of a witness might affect the degree of credibility attached

to their testimony. In Roman law, for example, the testimony of a slave was regarded as credible only if it was extracted via torture. In Islamic law, the testimony of one male and two female witnesses was considered equivalent to that of two male witnesses.

Daubert and Frye

Next, we turn to the American legal system for an example of how the testimony of experts is considered. Modern jurisprudence generally treats the trustworthiness of eyewitness testimony differently from the testimony of an expert. For an early case involving expert testimony on probability and statistics in the context of handwriting analysis, see the matter of the Howland will, decided in 1868 (the expert was Benjamin Peirce, father of C.S. Peirce, and at question was whether the signature on a will was a forgery).[6] While the distinction between eyewitness and expert testimony has long been recognized, it was not until the twentieth century, at least in US law, with first the *Frye* decision and then the *Daubert* decision, that standards for the trustworthiness of evidence produced by technological means and testimony given by experts came to be more precisely defined.

Frye. Similarly to A/IS today, the D.C. Circuit case *Frye v. U.S.* case in 1923,[7] concerned whether or not the results of a new type of system were trustworthy. In particular, the *Frye* case considered whether results of a blood pressure test (a precursor to the lie detector test) should be admissible in court. The *Frye* standard for admissibility essentially focuses on one condition: whether a scientific theory is generally accepted in the scientific community and has the technique been applied correctly. Currently, nine states use the Frye standard, or a modified version of it, when admitting evidence.[8] It has been criticized as too vague and unable to reliably manage complex scientific testimony.[9]

Daubert. The *Daubert* standard, established in the U.S. Supreme Court case *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, in 1993,[10] overruled the Frye standard in U.S. federal courts and is now the standard for federal courts, codified in Rule 702 of the U.S. Federal Rules of Evidence, and adopted, in whole or in part, in approximately 27 states.[11] Under Daubert, rather than the “general acceptance” test, the key criteria are relevance and reliability. The Supreme Court and the U.S. Federal Rules of Evidence list a non-exhaustive set of factors to consider including whether an expert’s technique or theory can be tested and assessed for reliability, the technique or theory has been subject to peer review and publication, the known or potential rate of error of the technique or theory, the existence and maintenance of standards and controls, and whether the technique or theory has been generally accepted in the scientific

[6] *Robinson* v. *Mandell*, 20 *Fed. Cas.* 1027, 1868. For discussion of the use of statistics and probabilistic reasoning in the case (and of the use of expert testimony on those topics), see Meier & Zabell 1980.

[7] (Frye v. United States, 293 F. 1013 (D.C. Cir. 1923))

[8] See Funk 2022.

[9] See Cappellino 2022.

[10] 509 U.S. 579 (1993). In practice, the Daubert standard refers to the Daubert case and its progeny including the Supreme Court decisions in *General Electric Co. v. Joiner*, 522 U.S. 136 (1997) (emphasizing the importance of expert methodology as opposed to focusing solely on the conclusory opinion of the expert) and *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999) (explaining that *Daubert* factors apply to expert testimony that is not scientific in nature but which still requires “technical, or other specialized knowledge”).

[11] Fed R. Evid. 702. For further discussion see Cappellino 2022.

community. Even with *Daubert*'s list of factors, there are still questions as to whether the judiciary is equipped to evaluate the merit of scientific testimony, especially because testing and reliability are oftentimes the crux of the analysis.[12]

Beyond Daubert and Frye

While it is not the case that all assessments of the trustworthiness of evidence take place in a legal context, we take inspiration from the *Frye* and *Daubert* standards as well as lessons from the assessment of various types of forensic evidence by the National Research Council (NRC), National Academy of Sciences (NAS) in 2009, and the President's Council of Advisors on Science and Technology (PCAST) in 2016. As shown by the NAS and PCAST studies, neither *Frye* nor *Daubert* have a track record of ensuring that the evidence produced for use in courtrooms is actually trustworthy.[13]

NRC/NAS. The 2009 NRC/NAS report, "Strengthening Forensic Science in the United States: A Path Forward", documented patterns of deficiencies common to forensic methods. They concluded that much forensic evidence has been introduced into criminal trials "without any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline."

PCAST. In September 2016, PCAST released a report "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods" which evaluated the scientific validity of seven feature-comparison methods that have been used to generate evidence used in courtroom proceedings: 1) DNA analysis of single-source and simple-mixture samples, 2) DNA analysis of complex-mixture samples, 3) Bite mark analysis, 4) Latent fingerprint analysis, 5) Firearms analysis, 6) Footwear analysis and 7) Hair analysis. For each, the report questioned whether the methods were foundationally valid and validly applied, equivalent to key prongs of *Daubert*.[14] In sum, the only foundationally valid method was DNA analysis of single-source and simple-mixture samples.[15] Specifically, the report divided the other methods into three categories: methods for which foundational validity had not been adequately established (bitemark analysis, firearms analysis, footwear analysis, and hair analysis),[16] methods for which foundational validity has only been established in narrow circumstances (DNA analysis of complex-mixture

[12] As *Daubert* pointed out, "a key question to be answered…will be whether [an expert's technique or theory] can be (and has been) tested," implying the particular importance of the testing for reliability factor.

[13] See e.g., Marc Canellas. 2021. Defending IEEE Software Standards in Federal Criminal Court. Computer 54, 6 (2021), 14–23; Rediet Abebe, Moritz Hardt, Angela Jin, John Miller, Ludwig Schmidt, and Rebecca Wexler. 2022. Adversarial Scrutiny of Evidentiary Statistical Software. In 2022, ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea. ACM, New York, NY, USA. https://doi.org/10.1145/3531146.3533228; Jim Hilbert. 2018. The disappointing history of science in the courtroom: Frye, Daubert, and the ongoing crisis of junk science in criminal trials. Oklahoma Law Review 71(2018), 759.

[14] *Daubert* is embodied in Federal Rule of Evidence 702 with 4 elements: (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case.

[15] PCAST at 7.

[16] PCAST at 9, 11, 13.

samples),[17] and methods for which foundational validity is largely subjective with substantial false positive rates (latent fingerprint at 9).[18] The fact that evidence generated through methods without established foundational validity has been accepted under the *Frye* and *Daubert* standards, suggested insufficiency of those standards for distinguishing trustworthy from untrustworthy evidence at least as applied in practice. We especially point to limitations of peer review as the standard for software-based evidence as peer review is not a replacement for independent verification and validation of the software.

If we wish to establish that a given A/IS meets the trust conditions, it is necessary to have evidence and that evidence must itself be trustworthy. *Frye* and *Daubert* offer some guidance as to how to vet evidence that might be used for this purpose, but, as the PCAST and NAS reports document, those standards are insufficient, especially when it comes to vetting evidence generated using advanced scientific or statistical techniques. Thus, a more effective framework for vetting the evidence used for assessing the trustworthiness of A/IS is necessary.

Errors that led to the use of some forensic techniques without first establishing foundational validity provide good examples of pitfalls to consider when evaluating the trustworthiness of AI and automated decision-making systems.

Features of Trustworthy Evidence from the PCAST Report

Empirical measurements of accuracy such as false positive rate, specificity, and sensitivity are important to understand the validity of a method. It is not sufficient that a method has been tested on a large collection of known data; representative samples must also be drawn from relevant populations. If data in a particular use case falls outside the range for which these empirical measurements have been established, the foundational validity of the method is called into question.

Drawing upon the criteria used in the *Frye* and *Daubert* standards, as well as those applied in the PCAST report to the analysis of evidence, we can identify some of the key features that contribute to the trustworthiness of evidence (especially that of evidence generated by technological means).

The PCAST report makes a number of recommendations including:

1. Demonstrating the capability of an analyst through routine, blinded proficiency testing.
2. Demonstrating that the techniques were reliably applied in the case by providing a complete description of the procedures, results and laboratory notes.

[17] PCAST at 8.

[18] PCAST at 9.

3. Utilizing comprehensive and accurate reporting and testimony, including information from the empirical studies of false positive rates and sensitivity, information on the comparability between the types of samples used in these empirical studies and the sample(s) available in the particular case, and an accurate portrayal of the probative value of the observed features (i.e., how common or rare the features are, based on empirical studies).

Taking inspiration from the PCAST report, we present the following list of features that could be used to characterize 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.

These are features that contribute to the trustworthiness of evidence; whether, and to what extent, we actually trust a specific item of evidence is a matter of circumstance-dependent reasoned judgment. In addition, it bears noting that it is not expected that *all* evidence will, if it is to be considered trustworthy, exhibit *all* of these features. Some types of evidence, for example, may be the product of the application of a significant amount of expert judgment and may not meet the criterion of objectivity. Other types of evidence may not be readily empirically tested for accuracy.

For our consideration of A/IS, it is especially interesting to note that the PCAST report considers DNA analysis of single-source mixture samples and DNA analysis of complex-mixture samples separately. In general, the introduction of probabilistic versus deterministic analysis requires additional considerations. For example, a technique that tests some number of samples fundamentally introduces a level of uncertainty versus a technique that tests every single item. Of course, testing every single item is not always possible. For example, when the tests themselves are fundamentally destructive such as with testing of ammunition versus tests that can be performed many times (e.g. document retrieval). Similarly, it is easier to manage uncertainty when the correct answer can be computed manually versus a technique where the correct answer is fundamentally unknowable (e.g. prediction of future behavior).

EVIDENCE SUPPORTING INFORMED TRUST

When IEEE outlined its four conditions of informed trust in *Ethically Aligned Design*, it also specified the types of evidence relevant to each condition.

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[19] (both during development and after deployment and operation);
- Benchmarking studies[20] — 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;
- Descriptions of the process followed in implementing the system.

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;

[19] By a "local" validation, we mean an evaluation of the effectiveness of the system in the specific circumstances in which it is being applied. For example, in the case of legal discovery, a test of the effectiveness of a document retrieval system on the population of documents in scope for discovery in the proceeding at hand.

[20] By a "benchmark" evaluation, we mean an evaluation of the effectiveness of the system under generic conditions designed to model, generally but not specifically, the actual circumstances in which a system might be applied in a real-world setting. For example, in the case of legal discovery, a test of the effectiveness of a document retrieval system under conditions that, at an abstract level, might be encountered in a wide range of real-world proceedings.

- Educational resources available to developers, operators, and end users;
- Compliance with standards and regulations (those specifically relevant to operators).

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;
    - » Descriptions of documentation developer/operator is (is not) willing to disclose to an independent auditor;
- Descriptions of responses to prior adverse events.

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).
- What degree of oversight, if any, is provided by human decision makers when considering the output of the A/IS.

IEEE STANDARD 1012: VERIFICATION AND VALIDATION OF SOFTWARE

It is helpful to have an inventory, such as that compiled in the previous section, of the types of evidence that may be relevant in assessing the trustworthiness of an AI or automated system. However, to be operationally useful we need to couple that inventory with a protocol for obtaining that evidence. After all, the evidence will be meaningful only if it is obtained via procedures that maintain its accuracy and integrity. We close this article with a description of one such protocol, *IEEE Standard 1012*.[21]

A Protocol for Obtaining Evidence of Trustworthiness

When AI/S 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, the IEEE Standard 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, IEEE Standard 1012 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”).[22] When the integrity level increases, so too does the intensity and rigor of the required verification and validation tasks.

[21] For review of IEEE 1012 and its application to issues of trustworthiness within the legal field, see M. Canellas, “Defending IEEE Software Standards in Federal Criminal Court,” in Computer, vol. 54, no. 6, pp. 14-23, June 2021, doi: 10.1109/MC.2020.3038630.

[22] The term “extensive social loss” requires an expansive definition because there are sectors and domains where, because of existing and persistent social inequities, normative judgments may be biased. For example, school assignment algorithms can lead to “extensive social loss.” The algorithm itself may be fairly simple and straightforward but when used in an American school district with high segregation levels or uneven resources distribution there is, in fact, a social loss to being placed in an under-resourced school that is not a problem of the algorithm in isolation. The definition of an A/IS does not look at the algorithm in isolation but, instead, as a sociotechnical system of software and hardware which requires an understanding of who operates it, how it is operated, and the environment in which it operates. Even so-called “neutral”

The V&V process must be independent to avoid conflicts of interest that could lead to catastrophic failure. To this end, IEEE Standard 1012 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.[23] 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 Standard 1012 prior to being deployed and be subject to post-deployment auditing.[24] 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.”[25]

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 Standard 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.”[26] “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.”[27] And if sharing tools is necessary, “IV&V conducts qualification tests on tools to assure that the common tools do not contain errors that

technologies being applied in inequitable ways or in inequitable environments can lead to unfair and unjust outcomes that cause “extensive social loss,” and thus require standards, certification, and independent verification and validation.

[23] Any software or hardware used to investigate, incarcerate, or convict in the criminal legal system is

[24] IEEE-USA, “Artificial Intelligence: Accelerating Inclusive Innovation By Building Trust”, p. 5. “Before being deployed, high-risk AI[systems] ought to be independently verified and validated (IV&V) in accordance with IEEE Standard 1012, IEEE Standard for System, Software, and Hardware Verification and Validation, and be subject to recurring post-deployment audit, including with respect to their operators. Furthermore, governmental entities should make the reports documenting the required IV&V and audits of their high-risk AI[systems] public.”

[25] IEEE-USA, Letter to the National Institute of Standards and Technology (NIST), responding to request for comments on NIST Internal Report 8351-DRAFT “DNA Mixture Interpretation: A NIST Scientific Foundation Review”, November 18 2021, p. 4. (hereinafter referred to as “IEEE-USA Letter to NIST”). IEEE-USA, IEEE Standards Association, and IEEE Computer Society, Letter to the National Institute of Standards and Technology (NIST) providing comments from IEEE-USA, the IEEE Standards Association (IEEE SA), and the IEEE Computer Society on NIST’s Digital Investigative Techniques: A NIST Scientific Foundation Review (8354-DRAFT, “the Review”). July 11, 2022, p. 5. (“[I]t is unacceptable for any system influencing decisions with potentially catastrophic consequences - especially forensic techniques used in the criminal legal system - to not be managerially, technically, and financially independently verified and validated.”) (hereinafter referred to as “IEEE Letter to NIST”)

[26] IEEE 1012, p. 198.

[27] IEEE 1012, p. 198.

may mask errors in the system being analyzed and tested."[28] 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 Standard 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."[29] 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."[30]

For its part, financial independence ensures that funding is protected and provided for without being hijacked. IEEE Standard 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."[31]

The application of V&V, IV&V, and IEEE Standard 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.[32] 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 Standard 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 Standard 1012 could be that it did not identify an error that was created by the designer or developer.

[28] IEEE 1012, p. 198.

[29] IEEE 1012, p. 198.

[30] IEEE 1012, p. 198.

[31] IEEE 1012, p. 198.

[32] For example, a software used by law enforcement and prosecutors in criminal proceedings cannot be properly independently verified and validated directly by the software developers, law enforcement, forensic labs, or prosecutors.

- **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.[33] IEEE Standard 1012 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"[34] and (2) governments "should not procure AI[systems] that… are shielded from independent validation and verification, and public review."[35] 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.[36] 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.

IEEE 1012 Providing Evidence for Informed Trust, Daubert, and Frye

The particular value of IEEE 1012 as a protocol for providing evidence of trustworthiness, is that its comprehensive, independent testing of software and hardware will address most, if not all, of the conditions established by *Frye*, *Daubert*, or IEEE.

Frye. The results of an IEEE 1012 IV&V will show whether the technique or theory has been generally accepted in the scientific community, satisfying the *Frye* standard. This is because there is a consensus among the scientific community is that (1) IV&V answers whether a technique/theory actually generally works, (2) IV&V is the generally accepted technique for ensuring

[33] For example, it makes no difference whether the software or hardware was developed using lightweight methods like agile or heavyweight methods like waterfall. See, IEEE Letter to NIST, p. 5. ("[I]t is unacceptable for *any* system influencing decisions with potentially catastrophic consequences - especially forensic techniques used in the criminal legal system - to not be managerially, technically, and financially independently verified and validated.") (emphasis added).

[34] IEEE-USA, "Artificial Intelligence: Accelerating Inclusive Innovation By Building Trust", p. 6. In addition, that position statement explains, "Specifically, in legal disputes, tribunals should permit disclosure under appropriate protective orders of intellectual property related to AI/AS necessary to obtain evidence in compliance with other judicial requirements, including constitutional requirements, discovery laws, or subpoenas." See also, IEEE-USA Letter to NIST, p. 10; IEEE Letter to NIST, p. 10

[35] IEEE-USA, "Artificial Intelligence: Accelerating Inclusive Innovation By Building Trust", p. 6. See also, IEEE-USA Letter to NIST, p. 10-11. IEEE Letter to NIST, p. 10-11.

[36] IEEE-USA Letter to NIST, p. 10, n. 26. IEEE Letter to NIST, p. 10, n. 27.

reliability/trustworthiness, and (3) High-risk deployed systems should not be generally accepted as reliable/trustworthy without IV&V.

Daubert. The requirements of *Daubert* and Federal Rules of Evidence 703 ("Bases of an Expert's Opinion Testimony") include whether the expert's technique or theory can be or has been tested (can it be challenged in some objective sense and assessed for reliability), the known or potential rate of error of the technique or theory when applied, the existence and maintenance of standards and controls, and whether the technique or theory has been generally accepted in the scientific community (aka *Frye*).

Another prong of *Daubert* and Federal Rules of Evidence 703, whether the technique or theory has been subject to peer review and publication, is not met through IV&V but in practice, peer review and publication is a proxy. In their November 2021 letter to NIST, IEEE-USA notes that there are substantial limitations to peer review used in this way. "[P]eer-reviewed publications… [w]hile a priceless tool for scientific inquiry, are not a substitute, nor a valid approximation of IV&V when determining reliability or trustworthiness of a deployed system. Peer-reviewed publications form the foundation of scientific advancement, but peer reviewers of scientific publications are not tasked with answering questions like "Should the [A/IS] or results be admissible in court? Is the [A/IS] fit for the evidence in this legal case?" Peer reviewers do not have access to the system itself and are not tasked with assessing its reliability. Peer reviewers are assessing whether a publication deserves the attention of the scientific community, whether the results described deserve the attention of other scientists. With respect to specific legal cases, any individual case could go well beyond the bounds of the published studies".[37]

IV&V also satisfies other factors that federal courts and the Federal Rules of Evidence Advisory Committee[38] have found "relevant in determining whether expert testimony is sufficiently reliable to be considered by the trier of fact:"[39]

- Whether evidence results "naturally and directly out of research they have conducted independent of the litigation, or whether they have developed their opinions expressly for purposes of testifying,"[40]
- Whether the expert has unjustifiably extrapolated from an accepted premise to an unfounded conclusion,[41]
- Whether the expert has adequately accounted for obvious alternative explanations,[42] and,
- Whether the field of expertise claimed by the expert is known to reach reliable results for the type of opinion the expert would give.[43]

[37] IEEE-USA Letter to NIST, p. 4-5. IEEE Letter to NIST, p. 4-5.

[38] The Federal Rules of Evidence Advisory Committee on Evidence Rules advises Congress on what evidence rules to adopt. The Advisory Committee's notes are "usually a good source for determining the meaning of an evidence rule." (Capra at 1) (Advisory Committee Notes to the Federal Rules of Evidence That May Require Clarification, D. Capra, Reed Professor of Law Fordham University School of Law)

[39] Fed. R. Evid. 702, Advisory Committee Notes – 2000 Amendment.

[40] Daubert v. Merrell Dow Pharmaceuticals, Inc., 43 F.3d 1311, 1317 (9th Cir. 1995).

[41] General Elec. Co. v. Joiner, 522 U.S. 136, 146 (1997)

[42] Claar v. Burlington N.R.R., 29 F.3d 499 (9th Cir. 1994)

[43] See Kumho Tire Co. v. Carmichael, 119 S.Ct. 1167, 1175 (1999) (Daubert's general acceptance factor does not "help show that an expert's testimony is reliable where the discipline itself lacks reliability, as, for example, do theories grounded in any so-called generally accepted principles of astrology or necromancy.")

CONCLUSION

In this article we have considered the question of the basis for trusting in the operation of the AI systems that are playing an increasingly central role in many of our core social institutions. We set the question of evidence in the context of three tiers: Ethics and Values (Tier 1), Trust Conditions (Tier 2) and Evidence (Tier 3), and surveyed IEEE resources available for addressing each of these 3 tiers. We have taken instruction on the creation of trustworthy evidence from the historical legal context and the context of legal evidence in the United States. We have reviewed (a) the attributes that make evidence trustworthy, (b) the types of evidence that may be relevant in assessing the trustworthiness of an AI or automated system, and (c) a protocol for obtaining such evidence, and connected this to the ways in which we consider evidence of the trustworthiness of AI systems. We hope that this review will serve as a helpful resource for practitioners and regulators seeking to develop operationally viable normative instruments for the responsible use of AI.

REFERENCES

A. Cappellino. Daubert vs. Frye: Navigating the Standards of Admissibility for Expert Testimony. *Expert Institute*, 2022. Available: https://www.expertinstitute.com/resources/insights/daubert-vs-frye-navigating-the-standards-of-admissibility-for-expert-testimony/.

M. Canellas, "Defending IEEE Software Standards in Federal Criminal Court," in Computer, vol. 54, no. 6, pp. 14-23, June 2021, doi: 10.1109/MC.2020.3038630.

C. Funk. Daubert Versus Frye: A National Look at Expert Evidentiary Standards. *Expert Institute*, 2022. Available: https://www.expertinstitute.com/resources/insights/daubert-versus-frye-a-national-look-at-expert-evidentiary-standards/.

J. Lapore, "Detection of Deception," "The Last Archive,", Season 1, Episode 2, https://www.thelastarchive.com/season-1/episode-2-detection-of-deception.

J. Franklin, The Science of Conjecture: Evidence and Probability before Pascal. Johns Hopkins University Press, 2002.

Executive Office of the President President's Council of Advisors on Science and Technology, "Ensuring Scientific Validity of Feature-Comparison Methods", September 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf

P. Meier and S. Zabell, Benjamin Peirce and the Howland will, Journal of the American Statistical Association, 75(371), 497-506, 1980.

IEEE Standard for System, Software, and Hardware Verification and Validation, IEEE Standard 1012, 2016.

IEEE, Ethically Aligned Design First Edition (EAD1e), https://standards.ieee.org/industry-connections/ec/ead1e-infographic/

IEEE-USA, "Artificial Intelligence: Accelerating Inclusive Innovation By Building Trust", July 21 2020, https://ieeeusa.org/assets/public-policy/positions/ai/AITrust0720.pdf

IEEE-USA, Letter to the National Institute of Standards and Technology (NIST), responding to request for comments on NIST Internal Report 8351-DRAFT "DNA Mixture Interpretation: A NIST Scientific Foundation Review", November 18 2021. https://ieeeusa.org/assets/public-policy/policy-log/2021/111821.pdf

IEEE-USA, IEEE Standards Association, and IEEE Computer Society, Letter to the National Institute of Standards and Technology (NIST) providing comments from IEEE-USA, the IEEE Standards Association (IEEE SA), and the IEEE Computer Society on NIST's Digital Investigative Techniques: A NIST Scientific Foundation Review (8354-DRAFT, "the Review"). July 11, 2022

G. Bar, G. Wiktorzak, J.Matthews, “Four Conditions for Building Trusted AI Systems: Effectiveness, Competence, Accountability, and Transparency”, IEEE Beyond Standards, July 2021.https://beyondstandards.ieee.org/four-conditions-for-building-trusted-ai-systems/

IEEE, Ethics In Action In Autonomous and Intelligent Systems, https://ethicsinaction.ieee.org/p7000/.

IEEE, Code of Ethics, https://www.ieee.org/about/corporate/governance/p7-8.html.

ACM, ACM Code of Ethics and Professional Conduct, https://www.acm.org/code-of-ethics.

R. Wexler. “Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System” (February 21, 2017). 70 Stanford Law Review 1343 (2018), Available at SSRN: https://ssrn.com/abstract=2920883 or http://dx.doi.org/10.2139/ssrn.2920883

ACKNOWLEDGEMENTS

This document was much improved after many detailed and helpful comments given by Nicolas Economou, Eileen Lach, Carlos Ignacio Gutierrez and Erica Wissolik. Of course, responsibility for any errors belongs to the authors.

IEEE
USA