AI and the Pursuit of Justice: Questions To Ask and Evidence To Require

Jeanna Matthews

Clarkson University

October 5 2022

Clarkson UNIVERSITY defy convention™

A bit about me

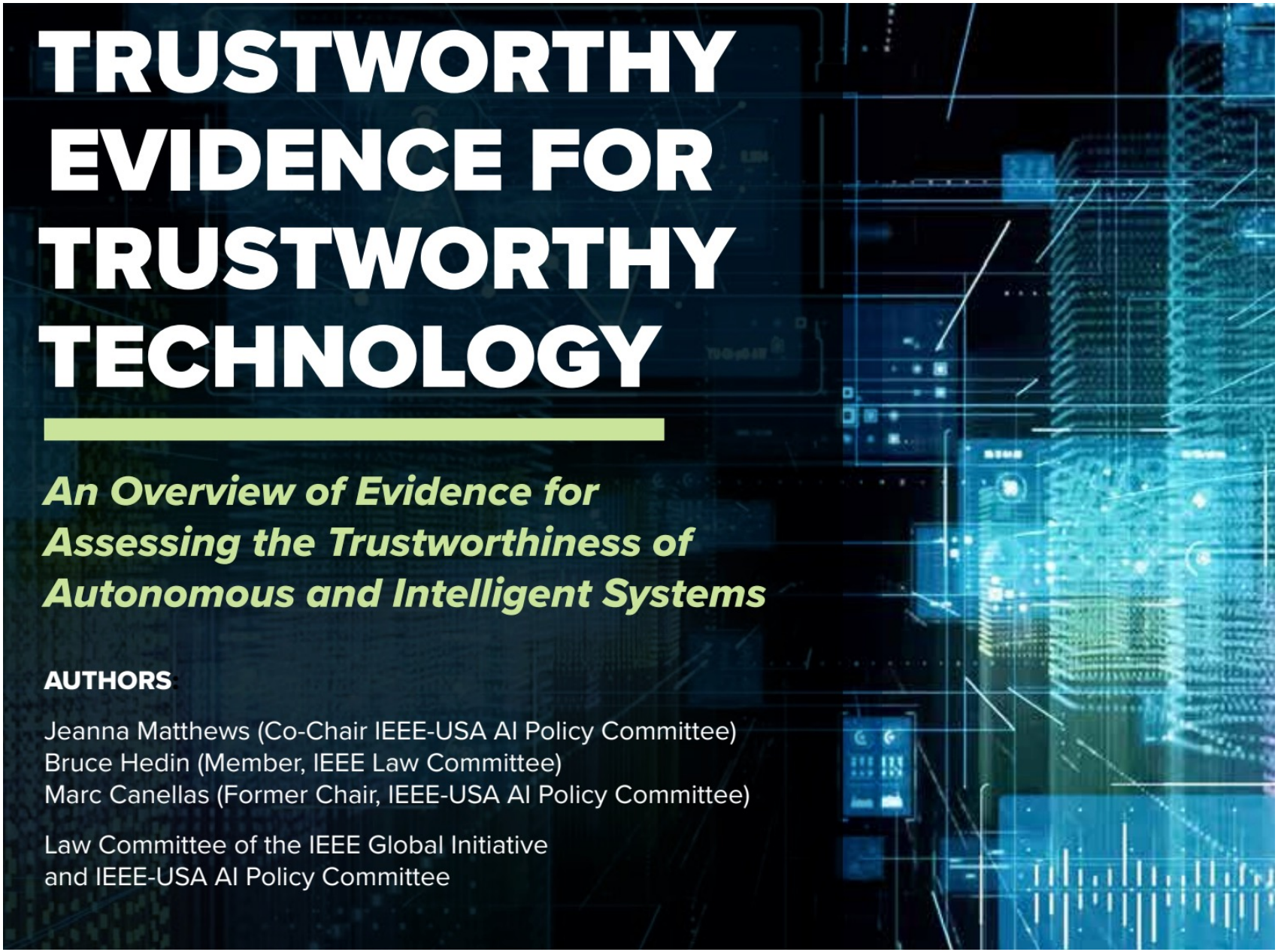
- PhD, UC Berkeley, 1994-1999
  - Network of Workstations project
- Professor
  - Clarkson University, since 2000
  - Cornell University, 2002- 2003
- Sabbaticals
  - VMware, Boston, 2008 – 2009
  - Data and Society New York, 2017-2019
- Industrial collaborations
  - Intel, EMC, Greenplum and others
- ACM and IEEE
  - ACM Council, Distinguished Speaker, SIGOPS Chair
  - ACM Technology Policy Council (TPC) and US-TPC
  - Vice-chair IEEE-USA AI Policy Committee

Data&Society

Association for Computing Machinery

Advancing Computing as a Science & Profession

Advancing Technology for Humanity


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

Articles

- *Jeanna Matthews, Bruce Hedin, Marc Canellas*
  Trustworthy Evidence for Trustworthy Technology: An Overview of Evidence for Assessing the Trustworthiness of Autonomous and Intelligent Systems
  IEEE-USA, September 29 2022.

- *Julia Brickell, Jeanna Matthews, Denia Psarrou, Shelley Podolny*,
  AI, Pursuit of Justice & Questions Lawyers Should Ask,
  Bloomberg Law, April 2022.

- *Gabriela Bar, Gabriela Wiktorzak, Jeanna Matthews*,
  Four Conditions for Building Trusted AI Systems: Effectiveness, Competence, Accountability, and Transparency
  IEEE Beyond Standards, July 13 2021.

What is Artificial Intelligence?

- Definition can be contentious
- In the context of law, recommend broad definition that includes automated decision-making systems broadly
- Autonomous and intelligent systems (A/IS) that often involve a mix of human decision-making and automated decision -making

Profound impact of AI on legal systems

Pervasive use of AI tools impacts lawyers, their clients, judges, and society as a whole

- E-Discovery
- Gunshot tracking
- Probabilistic genotyping
- Sentencing
- Facial recognition
- Hiring systems
- Surveillance systems
- Analysis of patterns in judicial decisions

Principles for the Ethical Use of AI in Legal Systems

- Council of Europe, through the European Commission for the Efficiency of Justice, has propounded an ethical charter on the use of AI in legal systems
- American Bar Association issued Resolution 112, cautioning lawyers to recognize that competence is required to understand when the risk of AI outweighs its benefits

Machine Learning

- AI systems are often trained based on patterns of past behavior including patterns of past human decision-making
  - Examples: image recognition, corpora of text, hiring decisions
- Past data often encodes patterns of discrimination and bias
- Futuristic force? Conservative force!

- Bugs are not surprising
  - Software and complex systems need an iterative process of debugging and improvement!
- Point is where is the incentive for testing and debugging?

Interests of deciders vs. those decided about

- Accuracy
  - Good enough according to who?
  - Invest some of savings in robust investigation of errors
  - Who can test/verify?
  - Rare cases that matter to individuals

- Conflicts between efficiency or reduced risk for the decision maker versus protection for the individual

Cautionary Tales

- Tradesecrets in recidivism scores
  - Loomis vs. Wisconsin
- Bias in facial recognition
  - Gender Shades project
- Failure to disclose or repair bugs in probabilistic genotyping
  - "The Right To Confront Your Accusers: Opening the Black Box of Forensic DNA Software", Forensic Statistical Tool (FST)
- Lack of rigorous testing for gunshot detection
  - ShotSpotter

What evidence can and should be collected about automated systems throughout their lifecyle?

What questions should lawyers and the broader society be asking about the results of automated systems?

Framework from IEEE

- 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?

	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.

Tier 1: Four Ethical Principles

- 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.

	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.

Tier 2: Trust Conditions

- Effectiveness
  - Solid information about the capabilities and limitations of an AI system to ensure fitness for the intended purpose.
- Competence
  - Certainty that operators have the skills and knowledge required for the effective operation of the AI system and adhere to those competency requirements.
- Accountability
  - Clear lines of responsibility to provide the rationale for decisions made in the design, development, procurement, deployment, operation, and validation of effectiveness for system outcomes.
- Transparency
  - Those affected by the output of an AI system have access to appropriate information about its design, development, procurement, deployment, operation, and validation of effectiveness

	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.

Tier 3: Characteristics of Sound 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.

Evidence of Each of the Trust Conditions

- Effectiveness
  - Evidence relevant to the assessment of a system's effectiveness will take the form, first and foremost, of scientifically sound empirical trials
- Competence
  - Evidence relevant to an assessment of the competence of the human agents involved in the deployment and operation of a system
  - Authoritative documentation of the agents' attainment of professional standards
  - Evidence from prior work in operating similar systems

Evidence of Each of the Trust Conditions

- 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
  - Documentation of prior responses to actual events can also be highly relevant
- 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

Evidence of Effectiveness

- 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;
- Descriptions of the process followed in implementing the system.

Evidence of Competence

- 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;
- Compliance with standards and regulations (those specifically relevant to operators).

Evidence of Accountability

- 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;
- Compliance with standards and regulations (those specifically relevant to operators).

Evidence of Transparency

- 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

- Standard for System, Software, and Hardware Verification and Validation
- Verification ensures that a product is correctly built
- Validation ensures that the right product is built

Integrity Levels

- 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.

Independent Verification and Validation (IV&V)

- "High-risk" systems, where catastrophic consequences are occasional or critical consequences are probable, should be independently verified and validated
- Technical Independence
- Managerial Independence
- Financial Independence

A Few Key Questions for Lawyers to Start With

- Have the systems you or your clients are considering (or using) been verified and validated according to IEEE 1012?
- For critical systems, what evidence is there of *independent* verification and validation?
- Under what conditions did the developers test the systems? How might the environment of the intended use differ?
- Is there research on bias and the potential for disparate impact in systems of this type? Are there key demographic groups for which the system was not tested? Might this system in this instance with this data have a disparate impact?
- What evidence is there that the system has been tested on use cases similar to the proposed use?

Thank you!

jnm@clarkson.edu

http://www.clarkson.edu/~jnm

@jeanna_matthews