Draft - Taxonomy of AI Risk

October 15, 2021

Note: This paper has been developed to address and build on responses to the Request for Information (RFI) released by NIST to help develop the AI Risk Management Framework (AI RMF). Among other things, in that RFI, NIST proposed eight characteristics of trustworthy AI. This paper aims to provide context to the eight characteristics of trustworthy AI mentioned in the RFI, clarify the distinction between characteristics and principles, and advance discussions about AI risks and forge agreements across organizations and internationally to the benefit AI design, development, use, and evaluation.

Background and Purpose

The National Institute of Standards and Technology (NIST) aims to cultivate trust in the design, development, use, and governance of Artificial Intelligence (AI) technologies and systems in ways that enhance economic security and improve quality of life. NIST focuses on improving measurement science, technology, standards, and related tools – including evaluation and data.

This white paper focuses on the preconditions of trust in AI and aims to further engage the AI community in a collaborative process to encourage consensus regarding terminology related to risk so that these types of risk may be identified and managed.

The paper starts by identifying several relevant policy directives that identify sources or types of risk across the AI lifecycle. For example, the Organisation for Economic Co-operation and Development (OECD) AI principles[1] specify that AI needs to have:

- Traceability to human values such as rule of law, human rights, democratic values, and diversity, and ensuring fairness and justice
- Transparency and responsible disclosure so people can understand and challenge AI-based outcomes
- Robustness, security, and safety, through the AI lifecycle to manage risks
- Accountability in line with these principles

Similarly, the European Union Digital Strategy's Ethics Guidelines for Trustworthy AI[2] identifies seven key principles of trustworthy AI:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination, and fairness
- Environmental and societal well-being
- Accountability

Finally, US Executive Order 13960, *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*[3] specifies that AI should be:

- Lawful and respectful of our Nation's values.
- Purposeful and performance-driven… using AI, where the benefits of doing so significantly outweigh the risks, and the risks can be assessed and managed
- Accurate, reliable, and effective

[1] https://www.oecd.org/going-digital/ai/principles/
[2] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[3] https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government

- Safe, secure, and resilient
- Understandable…by subject matter experts, users, and others, as appropriate
- Responsible and traceable
- Regularly monitored
- Transparent
- Accountable

Categories of Risk

Those three documents indicate that AI system stakeholders must account for several different sources of risk in the AI lifecycle. This proposed taxonomy seeks to simplify the categorization of these risks so that stakeholders may better recognize and manage them. The approach is hierarchical. First, it is recognized that there are three broad categories of risk sources related to AI systems:

1) **Technical design attributes.** This refers to the factors that are under the direct control of system designers and developers, and which may be measured using standard evaluation criteria that have traditionally been applied to machine learning systems, or that may be applied in an automated way in the future. Examples include accuracy and related measures (e.g., false positive and false negative rates, precision, recall, F-score) but also sources of statistical error that might be measured by applying AI tools to new data (e.g., discrepancies between performance on test and holdout sets). Finally, data generated from experiments that are designed to evaluate system performance also fall into this category, and might include tests of causal hypotheses, assessments of robustness to adversarial attack, etc.
2) **How AI systems are perceived.** This refers to mental representations of models, including whether the output provided is sufficient to evaluate compliance (transparency), whether model operations can be easily understood (explainability), and whether they provide output that can be used to make a meaningful decision (interpretability). In general, any judgment or assessment of an AI system, or its output, that is made by a human or needs human interpretation rather than by an automated process falls into this category.
3) **Guiding policies and principles.** This refers to broader societal determinations of value, such as privacy, accountability, fairness, justice, equity, etc., which cannot be measured consistently across domains because of their dependence on context.

Within each of these categories, several sources of potential risk have been identified.

Characteristics of Trustworthy Systems

1) Technical Attributes

The proposition put forward is that technical risks can be best managed by techniques to ensure the validity of machine learning methods. These risks can be explicitly measured using variations

of standard statistical or machine-learning metrics with specific thresholds specified in requirements. Specifically:

- **Accuracy.** This trustworthiness attribute captures the broad notion of whether the machine learning model is correctly capturing a relationship that exists within training data. It is analogous to statistical conclusion validity. Traditionally, accuracy may be measured and mitigated using standard metrics including, false positive and false negative rates, F1-score, precision, recall, etc. Beyond these traditional measures, the quality of a machine learning model must be assessed for whether it is underfit or overfit. One way to do so is to compare performance on training data to test and holdout data. More generally, it is widely acknowledged that current machine learning methods cannot provide a guarantee that the underlying model is capturing a causal relationship that generates the observed data (although see [3]-[5] for progress in this area). Establishment of internal (that is, causal) validity is an active area of research. (See also [6].)
- **Reliability.** A model is reliable if its output is insensitive to small changes in its input, and if it is free from measurement bias. Techniques designed to mitigate overfitting (e.g., regularization) and to adequately conduct model selection in the face of the bias/variance tradeoff can increase model reliability. The definition of reliability that is used here is analogous to construct validity in the social sciences, albeit without explicit reference to a theoretical construct. Specifically, this definition captures convergent-discriminant validity (whether the data reflects what the user intends to measure and not other things) and statistical reliability (whether the data may be subject to high levels of statistical noise and measurement bias). Measures of reliability might include Fleiss' Kappa scores, or goodness-of-fit tests from a factor analysis.
- **Robustness.** A model is robust if it applies to multiple settings beyond which it was trained. Threats to model robustness can be mitigated by explicitly recognizing limitations of the sampling strategy by which training, test, and holdout data were selected and ensuring that models are not applied "off label" (i.e., in domains that are not representative of these training data). Application of a "gradient of similarity" [7] may be helpful in mitigating this risk as well. Thus, robustness is analogous to "external validity" in the social sciences. Measures of robustness might include error measurements on novel datasets.
- **Resilience or Security.** A model that is insensitive to adversarial attacks, or more generally, to unexpected changes in its environment or use, may be said to be resilient and secure. This concept has some relationship to robustness except that it goes beyond the provenance of the data to encompass unexpected or explicitly hostile uses of the model or data. Mitigating these risks is an open area of research but may benefit from insights into flexible system design (e.g., [8]). Specific measures are still under development by the research community.

What these attributes have in common is that the extent of the corresponding risks may be directly measured using some kind of automatable process that does not require extensive human input. As a consequence, one may, at least in principle, develop domain-agnostic

measures of performance for each source of risk. Furthermore, these measures of performance may be expressed in requirements documents created by systems designers. Accordingly, attempts to manage these risks may follow standard systems engineering risk management practices.

2) Socio-Technical Attributes

How is one to know which technical measures are appropriate to a given task? Human judgment must be employed when deciding on the specific metrics, and the precise values of these metrics. Additionally, human users will also make judgments regarding what these metrics, and the associated models, mean when applied to daily life. Thus, a second broad category of risk pertains to how these human judgments are made. These include:

- **Explainability.** Attempts to increase explainability seek to provide a programmatic description of how model predictions are generated [9]. The underlying assumption is that perceptions of risk stem from a lack of technical background knowledge on the part of the user. Even given all the information required to make a model fully transparent, a human must apply what technical expertise they have to understand how the model works. Explainability refers to the user's perception of how the model works – such as what output may be expected for a given input. Risks due to explainability may arise if humans incorrectly infer a model's operation and it does not operate as expected. This risk may be managed by descriptions of how models work to users' skill levels.

  Explainability is related to transparency – a "white box" model is typically considered explainable, whereas a "black box" model is not. However, transparency does not guarantee explainability, especially if the user lacks an understanding of machine learning technical principles.

- **Interpretability.** Attempts to increase interpretability seek to fill a meaning deficit [10]. The underlying assumption is that perceptions of risk stem from a lack of ability to make sense of, or contextualize, model output appropriately. For example, models are developed for a particular functional use. Model interpretability refers to the extent to which a user is able to determine adherence to this function and the consequent implications of this output upon other consequential decisions for that user. Interpretations are typically contextualized in terms of values, and reflect simple, categorical distinctions. For example, a society may value privacy and safety, but individuals may have different determinations of how much safety is "safe enough" or how much privacy is sufficient.

  Risks to interpretability can often be addressed by communication of the interpretation intended by model designers, although this remains an open area of research. However, the prevalence of different interpretations can be readily measured with psychometric instruments. Interpretability is the glue that links transparency – information provided along with a model's output – to determinations that have to do with values (e.g., privacy, safety). Given a transparent stimulus shown to a decision-maker, they can then apply their values to *interpret* it and determine whether it is, for example, "safe" or "not safe."

However, transparency does not guarantee interpretability. Often, interpretability is associated with simple representations whereas transparency may create information overload. Interpretability is premised on the user's ability to "connect the dots" given information provided by a more transparent system. This means that there must be dots to connect (i.e., transparency is a needed precursor), but also that the information is presented in such a way that the user can craft a coherent understanding of the model's use in context. Similarly, apart from statistical measures of bias, interpretability allows decision-makers to make determinations about whether data, an algorithm, a process, etc., imposes an undesired bias (and is therefore unfair).

- **Privacy.** Like safety and security, specific technical features of a system may promote privacy and assessors can identify how the processing of data could create privacy-related problems. However, determinations of likelihood and severity of impact of these problems are contextual and vary between cultures and individuals. Furthermore, ensuring fairness may require violating privacy and vice versa (since fairness determinations often require obtaining data that some consider private).
- **Safety.** In the context of medical devices and drugs, safety is a categorical determination made by domain experts: a drug is either deemed "safe and efficacious" or it is not. These determinations are made relative to the state of the art in the field, and relative to society's expectations. Regulatory agencies, such as the U.S. Food and Drug Administration (FDA), typically maintain measures for safety in a given context; however, these measures are subject to revision, often with input from practitioners. For example, FDA convenes panels of experts to determine standards of safety for innovative medical devices – a process that is not without social influence [11], [12]. Determinations of security, as a value, are similar [13]. Leveson [14] proposes a systems theoretic definition of safety that may serve as the basis for preliminary metrics.
- **Managing bias.** Schwartz et al. [15] point out that bias is neither new nor unique to AI, nor can bias be eliminated entirely. Rather, biases which are harmful must be identified and, to the extent possible, understood, measured, managed, and reduced. Furthermore, perceptions of bias are also human judgments. Thus, perceptions of bias are intimately related to interpretations of model output.

3) Guiding Principles Contributing to AI Trustworthiness

Human judgments are premised on guiding policies and principles – broad social constructs that indicate societal priorities. AI has the potential to benefit nearly all aspects of our society, but the development and use of new AI-based technologies, products, and services bring technical and societal challenges and risks, including risks to ethical values. While there is no objective standard for ethical values, as they are grounded in the norms and legal expectations of specific societies or cultures, it is widely agreed that AI must be developed in a trustworthy manner. This trustworthiness can support the development and deployment of AI in ways that meet a given set of ethical values.

When specified as policy, human experts apply their judgments to "flow down" these principles into technical requirements. Several of the policy documents cited above outline

broad statements of values to which AI should adhere. Systems engineers frequently derive requirements from these values, which are later translated to measures of performance and effectiveness.

Principles relevant to AI include:

- **Fairness.** Like safety, standards of fairness are culturally determined, and perceptions of fairness differ between cultures, with societal determinations of fairness litigated in courts. Engineers often assume that machine learning algorithms are inherently fair because the same procedure applies regardless of user; however, this perception has eroded recently as awareness of biased algorithms and biased datasets has increased. Arguably, absence of harmful bias is a necessary condition for fairness.
- **Accountability.** Determinations of accountability are closely related to notions of risk and "blame" – that is, the responsible party in the event that a risky outcome is realized. Anthropologists, including Mary Douglas [16], have written extensively on how perceptions of risk and blame associated with technology differ systematically between cultures, and legal scholars [17] have developed psychometric measures of cultural cognition that are theorized to vary with these risk perceptions.
- **Transparency.** Attempts to increase transparency seek to fill a perceived information deficit. The underlying assumption is that perceptions of risk stem from an absence of information. Transparency reflects the extent to which information is available to a decision-maker when making a judgment about an AI system, and may span the scope from what data were included in model training, the structure of the model, its intended use case, to how decisions were made, by whom, when, etc. Absent transparency, users are left to guess about these factors and may make unwarranted assumptions regarding model provenance.

  Although it is impossible to remove a subject's background knowledge from their evaluations of a model, making adequate knowledge available is a precursor to building trust. This risk may be mitigated by a *transparent process* – one in which users can get answers regarding what decisions were made and what resources (e.g., data, energy, etc.) were used throughout the lifecycle, and why these decisions were made. This highlights the importance of documenting information in a standardized manner throughout the development lifecycle of an AI algorithm (i.e., the need for a "transparency toolkit.") Beyond such a toolkit, users' perceptions of systems as transparent are crucial. This emphasizes the need to develop approaches (e.g., a convenient user interface and cataloguing system, and possibly human contact) to surface this information when needed or requested, potentially in a context-sensitive manner. Finally, transparency is often framed as an instrumental value – a means to the end of achieving a broader value, such as accountability.

Table 1 provides a mapping of the proposed taxonomy to those provided by OECD, the EU, and the US Executive Order 13960.

Table 1: Mapping of taxonomy proposed in this paper to those provided in relevant policy documents

	Proposed Taxonomy	OECD	EU	US EO 13960
Technical Design Attributes	• Accuracy • Reliability • Robustness • Security & Resilience	• Robustness • Security	• Technical robustness	• Purposeful and performance-driven • Accurate, reliable, and effective • Secure and resilient
Socio-Technical Attributes	• Explainability • Interpretability • Privacy • Safety • Absence of Bias	• Safety	• Safety • Privacy • Non-discrimination	• Safe • Understandable by subject matter experts, users, and others, as appropriate
Guiding Principles Contributing to Trust-worthiness	• Fairness • Accountability • Transparency	• Traceability to human values • Transparency and responsible disclosure • Accountability •	• Human agency and oversight • Data governance • Transparency • Diversity and fairness • Environmental and societal well-being • Accountability	• Lawful and respectful of our Nation's values • Responsible and traceable • Regularly monitored • Transparent • Accountable

Conclusion

Although the proposed taxonomy cannot be claimed to be collectively exhaustive, the three high-level categories that have been identified appear to take into account of existing frameworks and may be seen as providing an overarching approach. Within each category, new sources of risk will be identified as the AI landscape continues to evolve. Robust discussions of this taxonomy, which to date has been happening informally via workshops and small group discussions, would position industry, government, and academia to better anticipate these broad categories of future risks and to develop management strategies that could be flexibly implemented. NIST's development of an AI Risk Management Framework is an ideal opportunity to advance those discussions and forge agreements across organizations and internationally to the benefit AI design, development, use, and evaluation.

References

[1] B. Stanton and T. Jensen, *Trust and Artificial Intelligence,* preprint, Mar. 2021. doi: 10.6028/NIST.IR.8332-draft.

[2] E. Tabassi, K. J. Burns, M. Hadjimichael, A. D. Molina-Markham, and J. T. Sexton, *A taxonomy and terminology of adversarial machine learning*, preprint, Oct. 2019. doi: 10.6028/NIST.IR.8269-draft.

[3] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge: Cambridge University Press, 2009. doi: 10.1017/CBO9780511803161.

[4] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. 2018.

[5] E. Bareinboim and J. Pearl, *Causal inference and the data-fusion problem PNAS*, vol. 113, no. 27, pp. 7345–7352, Jul. 2016, doi: 10.1073/pnas.1510507113.

[6] D. A. Broniatowski and C. S. Tucker, *Assessing Causal Claims About Complex Engineered Systems with Quantitative Data: Internal, External, and Construct Validity*, *Systems Engineering*, no. 20, pp. 483–496, 2017, doi: 10.1002/sys.21414.

[7] W. M. Trochim, *The Research Methods Knowledge Base, 2nd Edition*, Oct. 20, 2006. http://www.socialresearchmethods.net/kb/ (accessed Aug. 01, 2017).

[8] D. A. Broniatowski and J. Moses, *Measuring Flexibility, Descriptive Complexity, and Rework Potential in Generic System Architectures*, Systems Engineering, vol. 19, no. 3, pp. 207–221, Sep. 2016, doi: 10.1002/sys.21351.

[9] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, and M. A. Przybocki, *Four Principles of Explainable Artificial Intelligence,* preprint, Aug. 2020. doi: 10.6028/NIST.IR.8312-draft.

[10] D. A. Broniatowski, *Psychological Foundations of Explainability and Interpretability in Artificial Intelligence,* National Institute of Standards and Technology, Apr. 2021. doi: 10.6028/NIST.IR.8367.

[11] D. A. Broniatowski and C. L. Magee, *Studying Group Behaviors: A tutorial on text and network analysis methods,* Signal Processing Magazine, IEEE, vol. 29, no. 2, pp. 22–32, 2012.

[12] D. A. Broniatowski and C. L. Magee, *Does Seating Location Impact Voting Behavior on Food and Drug Administration Advisory Committees*?" American Journal of Therapeutics, vol. 20, no. 5, pp. 502–506, Oct. 2013, doi: 10.1097/MJT.0b013e31821109d5.

[13] N. Leveson, *Safety and Security Are Two Sides of the Same Coin*, in *The* Coupling of Safety and Security, C. Bieder and K. Pettersen Gould, Eds. Cham: Springer International Publishing, 2020, pp. 17–27. doi: 10.1007/978-3-030-47229-0_3.

[14] N. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, 2011. Accessed: Jan. 07, 2015. [Online]. Available: http://books.google.com/books?hl=en&lr=&id=0gZ_7n5p8MQC&oi=fnd&pg=PR9&dq=safety+leveson&ots=xOIAU2dAqp&sig=_SLJQxtcJlZluhI-I4KUN7wCaIE

[15] R. Schwartz, L. Down, A. Jonas, and E. Tabassi, *A Proposal for Identifying and Managing Bias in Artificial Intelligence,* National Institute of Standards and Technology, Jun. 2021. doi: 10.6028/NIST.SP.1270-draft.

[16] M. Douglas, *Risk and Blame: Essays in Cultural Theory*, 1 edition. London; New York: Routledge, 1994.

[17] P. D. M. Kahan, *Cultural Cognition as a Conception of the Cultural Theory of Risk*, in *Handbook of Risk Theory*, S. Roeser, R. Hillerbrand, P. Sandin, and M. Peterson, Eds. Springer Netherlands, 2012, pp. 725–759. Accessed: Feb. 03, 2015. [Online]. Available: http://link.springer.com/referenceworkentry/10.1007/978-94-007-1433-5_28