NIST AI Risk Management Framework Playbook
– MEASURE

Abstract

The MEASURE function employs quantitative, qualitative, or mixed-method tools, techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and related impacts.

MEASURE uses knowledge relevant to AI risks identified in the MAP function and informs the MANAGE function risk monitoring and response efforts.

Contents

MEASURE-1: Appropriate methods and metrics are identified and applied. 2
MEASURE 1.1 2
MEASURE 1.2 3
MEASURE 1.3 4
MEASURE-2: AI systems are evaluated for trustworthy characteristics. 6
MEASURE 2.1 6
MEASURE 2.2 7
MEASURE 2.3 9
MEASURE 2.4 10
MEASURE 2.5 11
MEASURE 2.6 13
MEASURE 2.7 14
MEASURE 2.8 16
MEASURE 2.9 17
MEASURE 2.10 19
MEASURE 2.11 21
MEASURE 2.12 25
MEASURE 2.13 27
MEASURE-3: Mechanisms for tracking identified AI risks over time are in place. 28
MEASURE 3.1 28
MEASURE 3.2 29
MEASURE 3.3 30
MEASURE-4: Feedback about efficacy of measurement is gathered and assessed. 32
MEASURE 4.1 32
MEASURE 4.2 34
MEASURE 4.3 36

MEASURE-1: Appropriate methods and metrics are identified and applied.

MEASURE 1.1

Approaches and metrics for measurement of AI risks enumerated during the Map function are selected for implementation starting with the most significant AI risks. The risks or trustworthiness characteristics that will not – or cannot – be measured are properly documented.

About

The development and utility of trustworthy AI systems depends on reliable measurements and evaluations of underlying technologies and their use. Compared with traditional software systems, AI technologies bring new failure modes, inherent dependence on training data and methods which directly tie to data quality and representativeness. Additionally, AI systems are inherently socio-technical in nature, meaning they are influenced by societal dynamics and human behavior. AI risks – and benefits – can emerge from the interplay of technical aspects combined with societal factors related to how a system is used, its interactions with other AI systems, who operates it, and the social context in which it is deployed. In other words, What should be measured depends on the purpose, audience, and needs of the evaluations.

These two factors influence selection of approaches and metrics for measurement of AI risks enumerated during the Map function. The AI landscape is evolving and so are the methods and metrics for AI measurement. The evolution of metrics is key to maintaining efficacy of the measures.

Suggested Actions

- Establish approaches for detecting, tracking and measuring known risks, errors, incidents or negative impacts.

- Identify testing procedures and metrics to demonstrate whether or not the system is fit for purpose and functioning as claimed.
- Identify testing procedures and metrics to demonstrate AI system trustworthiness
- Define acceptable limits for system performance (e.g. distribution of errors), and include course correction suggestions if/when the system performs beyond acceptable limits.
- Define metrics for, and regularly assess, AI actor competency for effective system operation,
- Identify transparency metrics to assess whether stakeholders have access to necessary information about system design, development, deployment, use, and evaluation.
- Utilize accountability metrics to determine whether AI designers, developers, and deployers maintain clear and transparent lines of responsibility and are open to inquiries.
- Document metric selection criteria and include considered but unused metrics.
- Monitor AI system external inputs including training data, models developed for other contexts, system components reused from other contexts, and third-party tools and resources.
- Report metrics to inform assessments of system generalizability and reliability.
- Assess and document pre- vs post-deployment system performance. Include existing and emergent risks.
- Document risks or trustworthiness characteristics identified in the Map function that will not be measured, including justification for non- measurement.

Transparency and Documentation

Organizations can document the following:

- How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed?
- What corrective actions has the entity taken to enhance the quality, accuracy, reliability, and representativeness of the data?
- Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)
- Did your organization address usability problems and test whether user interfaces served their intended purposes?

- What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e.adversarial or stress testing)?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- Datasheets for Datasets. URL

References

Sara R. Jordan. "Designing Artificial Intelligence Review Boards: Creating Risk Metrics for Review of AI." 2019 IEEE International Symposium on Technology and Society (ISTAS), 2019. URL

IEEE. "IEEE-1012-2016: IEEE Standard for System, Software, and Hardware Verification and Validation." IEEE Standards Association. URL

ACM Technology Policy Council. "Statement on Principles for Responsible Algorithmic Systems." Association for Computing Machinery (ACM), October 26, 2022. URL

David Piorkowski, Michael Hind, and John Richards. "Quantitative AI Risk Assessments: Opportunities and Challenges." arXiv preprint, submitted January 11, 2023. URL

Daniel Schiff, Aladdin Ayesh, Laura Musikanski, and John C. Havens. "IEEE 7010: A New Standard for Assessing the Well-Being Implications of Artificial Intelligence." 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020. URL

MEASURE 1.2

Appropriateness of AI metrics and effectiveness of existing controls is regularly assessed and updated including reports of errors and impacts on affected communities.

About

Different AI tasks, such as neural networks or natural language processing, benefit from different evaluation techniques. Use-case and particular settings in which the AI system is used also affects appropriateness of the evaluation techniques. Changes in the operational settings, data drift, model drift are among factors that suggest regularly assessing and updating appropriateness of AI metrics and their effectiveness can enhance reliability of AI system measurements.

Suggested Actions

- Assess external validity of all measurements (e.g., the degree to which measurements taken in one context can generalize to other contexts).
- Assess effectiveness of existing metrics and controls on a regular basis throughout the AI system lifecycle.
- Document reports of errors, incidents and negative impacts and assess sufficiency and efficacy of existing metrics for repairs, and upgrades
- Develop new metrics when existing metrics are insufficient or ineffective for implementing repairs and upgrades.
- Develop and utilize metrics to monitor, characterize and track external inputs, including any third-party tools.
- Determine frequency and scope for sharing metrics and related information with stakeholders and impacted communities.
- Utilize stakeholder feedback processes established in the Map function to capture, act upon and share feedback from end users and potentially impacted communities.
- Collect and report software quality metrics such as rates of bug occurrence and severity, time to response, and time to repair (See Manage 4.3).

Transparency and Documentation

Organizations can document the following:

- What metrics has the entity developed to measure performance of the AI system?
- To what extent do the metrics provide accurate and useful measure of performance?
- What corrective actions has the entity taken to enhance the quality, accuracy, reliability, and representativeness of the data?
- How will the accuracy or appropriate performance metrics be assessed?
- What is the justification for the metrics selected?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

ACM Technology Policy Council. “Statement on Principles for Responsible Algorithmic Systems.” Association for Computing Machinery (ACM), October 26, 2022. URL

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer-Verlag, 2009. URL

Harini Suresh and John Guttag. “A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle.” Equity and Access in Algorithms, Mechanisms, and Optimization, October 2021. URL

Christopher M. Bishop. Pattern Recognition and Machine Learning. New York: Springer, 2006. URL

Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W. Duncan Wadsworth, and Hanna Wallach. “Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs.” Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, July 2021, 368–78. URL

MEASURE 1.3

Internal experts who did not serve as front-line developers for the system and/or independent assessors are involved in regular assessments and updates. Domain experts, users, AI actors external to the team that developed or deployed the AI system, and affected communities are consulted in support of assessments as necessary per organizational risk tolerance.

About

The current AI systems are brittle, the failure modes are not well described, and the systems are dependent on the context in which they were developed and do not transfer well outside of the training environment. A reliance on local evaluations will be necessary along with a continuous monitoring of these systems. Measurements that extend beyond classical measures (which average across test cases) or expand to focus on pockets of failures where there are potentially significant costs can improve the reliability of risk management activities. Feedback from affected communities about how AI systems are being used can make AI evaluation purposeful. Involving internal experts who did not serve as front-line developers for the system and/or independent assessors regular assessments of AI systems helps a fulsome characterization of AI systems’ performance and trustworthiness .

Suggested Actions

- Evaluate TEVV processes regarding incentives to identify risks and impacts.
- Utilize separate testing teams established in the Govern function (2.1 and 4.1) to enable independent decisions and course-correction for AI systems. Track processes and measure and document change in performance.

- Plan and evaluate AI system prototypes with end user populations early and continuously in the AI lifecycle. Document test outcomes and course correct.
- Assess independence and stature of TEVV and oversight AI actors, to ensure they have the required levels of independence and resources to perform assurance, compliance, and feedback tasks effectively
- Evaluate interdisciplinary and demographically diverse internal team established in Map 1.2
- Evaluate effectiveness of external stakeholder feedback mechanisms, specifically related to processes for eliciting, evaluating and integrating input from diverse groups.

- Evaluate effectiveness of external stakeholder feedback mechanisms for enhancing AI actor visibility and decision making regarding AI system risks and trustworthy characteristics.

Transparency and Documentation

Organizations can document the following:

- What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?
- How easily accessible and current is the information available to external stakeholders?
- To what extent does the entity communicate its AI strategic goals and objectives to the community of stakeholders?
- To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?
- To what extent is this information sufficient and appropriate to promote transparency? Do external stakeholders have access to information on the design, operation, and limitations of the AI system?
- What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

Board of Governors of the Federal Reserve System. “SR 11-7: Guidance on Model Risk Management.” April 4, 2011. URL

“Definition of independent verification and validation (IV&V)”, in IEEE 1012, IEEE Standard for System, Software, and Hardware Verification and Validation. Annex C, URL

Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. “Participation Is Not a Design Fix for Machine Learning.” Equity and Access in Algorithms, Mechanisms, and Optimization, October 2022. URL

Rediet Abebe and Kira Goldner. “Mechanism Design for Social Good.” AI Matters 4, no. 3 (October 2018): 27–34. URL

MEASURE-2: AI systems are evaluated for trustworthy characteristics.

MEASURE 2.1

Test sets, metrics, and details about the tools used during test, evaluation, validation, and verification (TEVV) are documented.

About

Documenting measurement approaches, test sets, metrics, processes and materials used, and associated details builds foundation upon which to build a valid, reliable measurement process. Documentation enables repeatability and consistency, and can enhance AI risk management decisions.

Suggested Actions

- Leverage existing industry best practices for transparency and documentation of all possible aspects of measurements. Examples include: data sheet for data sets, model cards, [commenters provided examples]
- Regularly assess the effectiveness of tools used to document measurement approaches, test sets, metrics, processes and materials used
- Update the tools as needed

Transparency and Documentation

Organizations can document the following:

- Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
- To what extent has the entity documented the AI system's development, testing methodology, metrics, and performance outcomes?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework- WEF - Companion to the Model AI Governance Framework, 2020. URL

References

Emily M. Bender and Batya Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." Transactions of the Association for Computational Linguistics 6 (2018): 587–604. URL

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model Cards for Model Reporting." FAT *19: Proceedings of the Conference on Fairness, Accountability, and Transparency, January 2019, 220–29. URL

IEEE Computer Society. "Software Engineering Body of Knowledge Version 3: IEEE Computer Society." IEEE Computer Society. URL

IEEE. "IEEE-1012-2016: IEEE Standard for System, Software, and Hardware Verification and Validation." IEEE Standards Association. URL

Board of Governors of the Federal Reserve System. "SR 11-7: Guidance on Model Risk Management." April 4, 2011. URL

Abigail Z. Jacobs and Hanna Wallach. "Measurement and Fairness." FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, 375–85. URL

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. URL

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. "Hard Choices in Artificial Intelligence." Artificial Intelligence 300 (November 2021). URL

MEASURE 2.2

Evaluations involving human subjects meet applicable requirements (including human subject protection) and are representative of the relevant population.

About

Measurement and evaluation of AI systems often involves testing with human subjects or using data captured from human subjects. Protection of human subjects is required by law when carrying out federally funded research, and is a domain specific requirement for some disciplines. Standard human subjects protection procedures include protecting the welfare and interests of human subjects, designing evaluations to minimize risks to subjects, and completion of mandatory training regarding legal requirements and expectations.

Evaluations of AI system performance that utilize human subjects or human subject data should reflect the population within the context of use. AI system activities utilizing non-representative data may lead to inaccurate assessments or negative and harmful outcomes. It is often difficult – and sometimes impossible, to collect data or perform evaluation tasks that reflect the full operational purview of an AI system. Methods for collecting, annotating, or using these data can also contribute to the challenge. To counteract these challenges, organizations can connect human subjects data collection, and dataset practices, to AI system contexts and purposes and do so in close collaboration with AI Actors from the relevant domains.

Suggested Actions

- Follow human subjects research requirements as established by organizational and disciplinary requirements, including informed consent and compensation, during dataset collection activities.
- Analyze differences between intended and actual population of users or data subjects, including likelihood for errors, incidents or negative impacts.
- Utilize disaggregated evaluation methods (e.g. by race, age, gender, ethnicity, ability, region) to improve AI system performance when deployed in real world settings.
- Establish thresholds and alert procedures for dataset representativeness within the context of use.
- Construct datasets in close collaboration with experts with knowledge of the context of use.
- Follow intellectual property and privacy rights related to datasets and their use, including for the subjects represented in the data.
- Evaluate data representativeness through
  - investigating known failure modes,
  - assessing data quality and diverse sourcing,
  - applying public benchmarks,
  - traditional bias testing,
  - chaos engineering,
  - stakeholder feedback
- Use informed consent for individuals providing data used in system testing and evaluation.

Transparency and Documentation

Organizations can document the following:

- Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
- How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?

- To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?
- To what extent has the entity identified and mitigated potential bias—statistical, contextual, and historical—in the data?
- If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework- WEF - Companion to the Model AI Governance Framework, 2020. URL
- Datasheets for Datasets. URL

References

United States Department of Health, Education, and Welfare's National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Volume II. United States Department of Health and Human Services Office for Human Research Protections. April 18, 1979. URL

Office for Human Research Protections (OHRP). “45 CFR 46.” United States Department of Health and Human Services Office for Human Research Protections, March 10, 2021. URL Note: Federal Policy for Protection of Human Subjects (Common Rule). 45 CFR 46 (2018)

Office for Human Research Protections (OHRP). “Human Subject Regulations Decision Chart.” United States Department of Health and Human Services Office for Human Research Protections, June 30, 2020. URL

Jacob Metcalf and Kate Crawford. “Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide.” Big Data and Society 3, no. 1 (2016). URL

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. “Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing.” arXiv preprint, submitted April 20, 2021. URL

Divyansh Kaushik, Zachary C. Lipton, and Alex John London. “Resolving the Human Subjects Status of Machine Learning's Crowdworkers.” arXiv preprint, submitted June 8, 2022. URL

Office for Human Research Protections (OHRP). “International Compilation of Human Research Standards.” United States Department of Health and Human Services Office for Human Research Protections, February 7, 2022. URL

National Institutes of Health. “Definition of Human Subjects Research.” NIH Central Resource for Grants and Funding Information, January 13, 2020. URL

Joy Buolamwini and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” Proceedings of the 1st Conference on Fairness, Accountability and Transparency in PMLR 81 (2018): 77–91. URL

Eun Seo Jo and Timnit Gebru. “Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning.” FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 2020, 306–16. URL

Marco Gerardi, Katarzyna Barud, Marie-Catherine Wagner, Nikolaus Forgo, Francesca Fallucchi, Noemi Scarpato, Fiorella Guadagni, and Fabio Massimo Zanzotto. “Active Informed Consent to Boost the Application of Machine Learning in Medicine.” arXiv preprint, submitted September 27, 2022. URL

Shari Trewin. “AI Fairness for People with Disabilities: Point of View.” arXiv preprint, submitted November 26, 2018. URL

Andrea Brennen, Ryan Ashley, Ricardo Calix, JJ Ben-Joseph, George Sieniawski, Mona Gogia, and BNH.AI. AI Assurance Audit of RoBERTa, an Open source, Pretrained Large Language Model. IQT Labs, December 2022. URL

MEASURE 2.3

AI system performance or assurance criteria are measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s). Measures are documented.

About

Once deployed, AI system purposes can drift as the application is repurposed or used in unforeseen ways, and in unplanned settings or contexts. Different deployment contexts means a new set of risks to be considered.

Suggested Actions

- Conduct regular and sustained engagement with potentially impacted communities
- Maintain a demographically diverse and multidisciplinary and collaborative internal team
- Regularly test and evaluate systems in non-optimized conditions, and in collaboration with AI actors in user interaction and user experience (UI/UX) roles.
- Evaluate feedback from stakeholder engagement activities, in collaboration with human factors and socio-technical experts.
- Collaborate with socio-technical, human factors, and UI/UX experts to identify notable characteristics in context of use that can be translated into system testing scenarios.
- Measure AI systems prior to deployment in conditions similar to expected scenarios.
- Measure and document performance criteria such as accuracy (false positive rate, false negative rate, etc.) and efficiency (training times, prediction latency, etc.).
- Measure assurance criteria such as AI actor competency and experience.
- Document differences between measurement setting and the deployment environment(s).

Transparency and Documentation

Organizations can document the following:

- What experiments were initially run on this dataset? To what extent have experiments on the AI system been documented?
- To what extent does the system/entity consistently measure progress towards stated goals and objectives?
- How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed? How much distributional shift or model drift from baseline performance is acceptable?
- As time passes and conditions change, is the training data still representative of the operational environment?
- What testing, if any, has the entity conducted on theAI system to identify errors and limitations (i.e.adversarial or stress testing)?

AI Transparency Resources:

- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework- WEF - Companion to the Model AI Governance Framework, 2020. URL
- Datasheets for Datasets. URL

References

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer-Verlag, 2009. URL

Jessica Zosa Forde, A. Feder Cooper, Kweku Kwegyir-Aggrey, Chris De Sa, and Michael Littman. “Model Selection’s Disparate Impact in Real-World Deep Learning Applications.” arXiv preprint, submitted September

7, 2021. URL

Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. "The Fallacy of AI Functionality." FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 2022, 959–72. URL

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. "Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research." Patterns 2, no. 11 (2021): 100336. URL

Christopher M. Bishop. Pattern Recognition and Machine Learning. New York: Springer, 2006. URL

Md Johirul Islam, Giang Nguyen, Rangeet Pan, and Hridesh Rajan. "A Comprehensive Study on Deep Learning Bug Characteristics." arXiv preprint, submitted June 3, 2019. URL

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. "DQI: Measuring Data Quality in NLP." arXiv preprint, submitted May 2, 2020. URL

Doug Wielenga. "Paper 073-2007: Identifying and Overcoming Common Data Mining Mistakes." SAS Global Forum 2007: Data Mining and Predictive Modeling, SAS Institute, 2007. URL

Software Resources

- Drifter library (performance assessment)
- Manifold library (performance assessment)
- MLextend library (performance assessment)
- PiML library (explainable models, performance assessment)
- SALib library (performance assessment)
- What-If Tool (performance assessment)

MEASURE 2.4

The functionality and behavior of the AI system and its components – as identified in the MAP function – are monitored when in production.

About

AI systems may encounter new issues and risks while in production as the environment evolves over time. This effect, often referred to as "drift", means AI systems no longer meet the assumptions and limitations of the original design. Regular monitoring allows AI Actors to monitor the functionality and behavior of the AI system and its components – as identified in the MAP function - and enhance the speed and efficacy of necessary system interventions.

Suggested Actions

- Monitor and document how metrics and performance indicators observed in production differ from the same metrics collected during pre-deployment testing. When differences are observed, consider error propagation and feedback loop risks.
- Utilize hypothesis testing or human domain expertise to measure monitored distribution differences in new input or output data relative to test environments
- Monitor for anomalies using approaches such as control limits, confidence intervals, integrity constraints and ML algorithms. When anomalies are observed, consider error propagation and feedback loop risks.
- Verify alerts are in place for when distributions in new input data or generated predictions observed in production differ from pre-deployment test outcomes, or when anomalies are detected.
- Assess the accuracy and quality of generated outputs against new collected ground-truth information as it becomes available.
- Utilize human review to track processing of unexpected data and reliability of generated outputs; warn system users when outputs may be unreliable. Verify that human overseers responsible for these processes have clearly defined responsibilities and training for specified tasks.

- Collect uses cases from the operational environment for system testing and monitoring activities in accordance with organizational policies and regulatory or disciplinary requirements (e.g. informed consent, institutional review board approval, human research protections),

Transparency and Documentation

Organizations can document the following:

- To what extent is the output of each component appropriate for the operational context?
- What justifications, if any, has the entity provided for the assumptions, boundaries, and limitations of the AI system?
- How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed?
- As time passes and conditions change, is the training data still representative of the operational environment?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

Luca Piano, Fabio Garcea, Valentina Gatteschi, Fabrizio Lamberti, and Lia Morra. “Detecting Drift in Deep Learning: A Methodology Primer.” IT Professional 24, no. 5 (2022): 53–60. URL

Larysa Visengeriyeva, et al. “Awesome MLOps.“ GitHub. URL

MEASURE 2.5

The AI system to be deployed is demonstrated to be valid and reliable. Limitations of the generalizability beyond the conditions under which the technology was developed are documented.

About

An AI system that is not validated or that fails validation may be inaccurate or unreliable or may generalize poorly to data and settings beyond its training, creating and increasing AI risks and reducing trustworthiness. AI Actors can improve system validity by creating processes for exploring and documenting system limitations. This includes broad consideration of purposes and uses for which the system was not designed.

Validation risks include the use of proxies or other indicators that are often constructed by AI development teams to operationalize phenomena that are either not directly observable or measurable (e.g, fairness, hireability, honesty, propensity to commit a crime). Teams can mitigate these risks by demonstrating that the indicator is measuring the concept it claims to measure (also known as construct validity). Without this and other types of validation, various negative properties or impacts may go undetected, including the presence of confounding variables, potential spurious correlations, or error propagation and its potential impact on other interconnected systems.

Suggested Actions

- Define the operating conditions and socio-technical context under which the AI system will be validated.
- Define and document processes to establish the system’s operational conditions and limits.
- Establish or identify, and document approaches to measure forms of validity, including:
  - construct validity (the test is measuring the concept it claims to measure)
  - internal validity (relationship being tested is not influenced by other factors or variables)
  - external validity (results are generalizable beyond the training condition) Standard approaches include the use of experimental design principles and statistical analyses and modeling.
- Assess and document system variance. Standard approaches include confidence intervals, standard deviation, standard error, bootstrapping, or cross-validation.

- Establish or identify, and document robustness measures.
- Establish or identify, and document reliability measures.
- Establish practices to specify and document the assumptions underlying measurement models to ensure proxies accurately reflect the concept being measured.
- Utilize standard statistical methods to test bias, inferential associations, correlation, and covariance in adopted measurement models.
- Utilize standard statistical methods to test variance and reliability of system outcomes.
- Monitor operating conditions for system performance outside of defined limits.
- Identify TEVV approaches for exploring AI system limitations, including testing scenarios that differ from the operational environment. Consult experts with knowledge of specific context of use.
- Define post-alert actions. Possible actions may include:
  - alerting other relevant AI actors before action,
  - requesting subsequent human review of action,
  - alerting downstream users and stakeholder that the system is operating outside it's defined validity limits,
  - tracking and mitigating possible error propagation
  - action logging
- Log input data and relevant system configuration information whenever there is an attempt to use the system beyond its well-defined range of system validity.
- Modify the system over time to extend its range of system validity to new operating conditions.

Transparency and Documentation

Organizations can document the following:

- What testing, if any, has the entity conducted on theAI system to identify errors and limitations (i.e.adversarial or stress testing)?
- Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
- How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?
- To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?
- What goals and objectives does the entity expect to achieve by designing, developing, and/or deploying the AI system?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL

References

Abigail Z. Jacobs and Hanna Wallach. “Measurement and Fairness.” FAccT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, 375–85. URL

Debugging Machine Learning Models. Proceedings of ICLR 2019 Workshop, May 6, 2019, New Orleans, Louisiana. URL

Patrick Hall. “Strategies for Model Debugging.” Towards Data Science, November 8, 2019. URL

Suchi Saria and Adarsh Subbaswamy. “Tutorial: Safe and Reliable Machine Learning.” arXiv preprint, submitted April 15, 2019. URL

Google Developers. “Overview of Debugging ML Models.” Google Developers Machine Learning Foundational Courses, n.d. URL

Software Resources

- Drifter library (performance assessment)

- Manifold library (performance assessment)
- MLextend library (performance assessment)
- PiML library (explainable models, performance assessment)
- SALib library (performance assessment)
- What-If Tool (performance assessment)

MEASURE 2.6

AI system is evaluated regularly for safety risks – as identified in the MAP function. The AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics implicate system reliability and robustness, real-time monitoring, and response times for AI system failures.

About

Many AI systems are being introduced into settings such as transportation, manufacturing or security, where failures may give rise to various physical or environmental harms. AI systems that may endanger human life, health, property or the environment are tested thoroughly prior to deployment, and are regularly evaluated to confirm the system is safe during normal operations, and in settings beyond its proposed use and knowledge limits.

Measuring activities for safety often relate to exhaustive testing in development and deployment contexts, understanding the limits of a system's reliable, robust, and safe behavior, and real-time monitoring of various aspects of system performance. These activities are typically conducted along with other risk mapping, management, and governance tasks such as avoiding past failed designs, establishing and rehearsing incident response plans that enable quick responses to system problems, the instantiation of redundant functionality to cover failures, and transparent and accountable governance. System safety incidents or failures are frequently reported to be related to organizational dynamics and culture. Independent auditors may bring important independent perspectives for reviewing evidence of AI system safety.

Suggested Actions

- Thoroughly measure system performance in development and deployment contexts, and under stress conditions.
  - Employ test data assessments and simulations before proceeding to production testing. Track multiple performance quality and error metrics.
  - Stress-test system performance under likely scenarios (e.g., concept drift, high load) and beyond known limitations, in consultation with domain experts.
  - Test the system under conditions similar to those related to past known incidents and measure system performance and safety characteristics.
  - Apply chaos engineering approaches to test systems in extreme conditions and gauge unexpected responses.
  - Document the range of conditions under which the system has been tested and demonstrated to fail safely.
- Measure and monitor system performance in real-time to enable rapid response when AI system incidents are detected.
- Collect pertinent safety statistics (e.g., out-of-range performance, incident response times, system down time, injuries, etc.) in anticipation of potential information sharing with impacted communities or as required by AI system oversight personnel.
- Align measurement to the goal of continuous improvement. Seek to increase the range of conditions under which the system is able to fail safely through system modifications in response to in-production testing and events.
- Document, practice and measure incident response plans for AI system incidents, including measuring response and down times.
- Compare documented safety testing and monitoring information with established risk tolerances on an on-going basis.

Transparency and Documentation

Organizations can document the following:

- What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e.adversarial or stress testing)?
- To what extent has the entity documented the AI system's development, testing methodology, metrics, and performance outcomes?
- Did you establish mechanisms that facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?
- Did you ensure that the AI system can be audited by independent third parties?
- Did you establish a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

AI Incident Database. 2022. URL

AIAAIC Repository. 2022. URL

Netflix. Chaos Monkey. URL

IBM. "IBM's Principles of Chaos Engineering." IBM, n.d. URL

Suchi Saria and Adarsh Subbaswamy. "Tutorial: Safe and Reliable Machine Learning." arXiv preprint, submitted April 15, 2019. URL

Daniel Kang, Deepti Raghavan, Peter Bailis, and Matei Zaharia. "Model assertions for monitoring and improving ML models." Proceedings of Machine Learning and Systems 2 (2020): 481-496. URL

Larysa Visengeriyeva, et al. "Awesome MLOps." GitHub. URL

MEASURE 2.7

AI system security and resilience – as identified in the MAP function – are evaluated and documented.

About

AI systems, as well as the ecosystems in which they are deployed, may be said to be resilient if they can withstand unexpected adverse events or unexpected changes in their environment or use – or if they can maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when this is necessary. Common security concerns relate to adversarial examples, data poisoning, and the exfiltration of models, training data, or other intellectual property through AI system endpoints. AI systems that can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use may be said to be secure.

Security and resilience are related but distinct characteristics. While resilience is the ability to return to normal function after an unexpected adverse event, security includes resilience but also encompasses protocols to avoid, protect against, respond to, or recover from attacks. Resilience relates to robustness and encompasses unexpected or adversarial use (or abuse or misuse) of the model or data.

Suggested Actions

- Establish and track AI system security tests and metrics (e.g., red-teaming activities, frequency and rate of anomalous events, system down-time, incident response times, time-to-bypass, etc.).

- Use red-team exercises to actively test the system under adversarial or stress conditions, measure system response, assess failure modes or determine if system can return to normal function after an unexpected adverse event.
- Document red-team exercise results as part of continuous improvement efforts, including the range of security test conditions and results.
- Use countermeasures (e.g, authentication, throttling, differential privacy, robust ML approaches) to increase the range of security conditions under which the system is able to return to normal function.
- Modify system security procedures and countermeasures to increase robustness and resilience to attacks in response to testing and events experienced in production.
- Verify that information about errors and attack patterns is shared with incident databases, other organizations with similar systems, and system users and stakeholders (MANAGE-4.1).
- Develop and maintain information sharing practices with AI actors from other organizations to learn from common attacks.
- Verify that third party AI resources and personnel undergo security audits and screenings. Risk indicators may include failure of third parties to provide relevant security information.

Transparency and Documentation

Organizations can document the following:

- To what extent does the plan specifically address risks associated with acquisition, procurement of packaged software from vendors, cybersecurity controls, computational infrastructure, data, data science, deployment mechanics, and system failure?
- What assessments has the entity conducted on data security and privacy impacts associated with the AI system?
- What processes exist for data generation, acquisition/collection, security, maintenance, and dissemination?
- What testing, if any, has the entity conducted on the AI system to identify errors and limitations (i.e. adversarial or stress testing)?
- If a third party created the AI, how will you ensure a level of explainability or interpretability?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

Matthew P. Barrett. “Framework for Improving Critical Infrastructure Cybersecurity Version 1.1.” National Institute of Standards and Technology (NIST), April 16, 2018. URL

Nicolas Papernot. “A Marauder’s Map of Security and Privacy in Machine Learning.” arXiv preprint, submitted on November 3, 2018. URL

Gary McGraw, Harold Figueroa, Victor Shepardson, and Richie Bonett. “BIML Interactive Machine Learning Risk Framework.” Berryville Institute of Machine Learning (BIML), 2022. URL

Mitre Corporation. “Mitre/Advmlthreatmatrix: Adversarial Threat Landscape for AI Systems.” GitHub, 2023. URL

National Institute of Standards and Technology (NIST). “Cybersecurity Framework.” NIST, 2023. URL

Software Resources

- adversarial-robustness-toolbox
- counterfit
- foolbox
- ml_privacy_meter
- robustness

- tensorflow/privacy

MEASURE 2.8

Risks associated with transparency and accountability – as identified in the MAP function – are examined and documented.

About

Transparency enables meaningful visibility into entire AI pipelines, workflows, processes or organizations and decreases information asymmetry between AI developers and operators and other AI Actors and impacted communities. Transparency is a central element of effective AI risk management that enables insight into how an AI system is working, and the ability to address risks if and when they emerge. The ability for system users, individuals, or impacted communities to seek redress for incorrect or problematic AI system outcomes is one control for transparency and accountability. Higher level recourse processes are typically enabled by lower level implementation efforts directed at explainability and interpretability functionality. See Measure 2.9.

Transparency and accountability across organizations and processes is crucial to reducing AI risks. Accountable leadership – whether individuals or groups – and transparent roles, responsibilities, and lines of communication foster and incentivize quality assurance and risk management activities within organizations.

Lack of transparency complicates measurement of trustworthiness and whether AI systems or organizations are subject to effects of various individual and group biases and design blindspots and could lead to diminished user, organizational and community trust, and decreased overall system value. Enstating accountable and transparent organizational structures along with documenting system risks can enable system improvement and risk management efforts, allowing AI actors along the lifecycle to identify errors, suggest improvements, and figure out new ways to contextualize and generalize AI system features and outcomes.

Suggested Actions

- Instrument the system for measurement and tracking, e.g., by maintaining histories, audit logs and other information that can be used by AI actors to review and evaluate possible sources of error, bias, or vulnerability.
- Calibrate controls for users in close collaboration with experts in user interaction and user experience (UI/UX), human computer interaction (HCI), and/or human-AI teaming.
- Test provided explanations for calibration with different audiences including operators, end users, decision makers and decision subjects (individuals for whom decisions are being made), and to enable recourse for consequential system decisions that affect end users or subjects.
- Measure and document human oversight of AI systems:
  - Document the degree of oversight that is provided by specified AI actors regarding AI system output.
  - Maintain statistics about downstream actions by end users and operators such as system overrides.
  - Maintain statistics about and document reported errors or complaints, time to respond, and response types.
  - Maintain and report statistics about adjudication activities.
- Track, document, and measure organizational accountability regarding AI systems via policy exceptions and escalations, and document “go” and “no/go” decisions made by accountable parties.
- Track and audit the effectiveness of organizational mechanisms related to AI risk management, including:
  - Lines of communication between AI actors, executive leadership, users and impacted communities.
  - Roles and responsibilities for AI actors and executive leadership.
  - Organizational accountability roles, e.g., chief model risk officers, AI oversight committees, responsible or ethical AI directors, etc.

Transparency and Documentation

Organizations can document the following:

- To what extent has the entity clarified the roles, responsibilities, and delegated authorities to relevant stakeholders?
- What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?
- Who is accountable for the ethical considerations during all stages of the AI lifecycle?
- Who will be responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed?
- Are the responsibilities of the personnel involved in the various AI governance processes clearly defined?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

National Academies of Sciences, Engineering, and Medicine. Human-AI Teaming: State-of-the-Art and Research Needs. 2022. URL

Inioluwa Deborah Raji and Jingying Yang. “ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles.” arXiv preprint, submitted January 8, 2020. URL

Andrew Smith. “Using Artificial Intelligence and Algorithms.” Federal Trade Commission Business Blog, April 8, 2020. URL

Board of Governors of the Federal Reserve System. “SR 11-7: Guidance on Model Risk Management.” April 4, 2011. URL

Joshua A. Kroll. “Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems.” FAccT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 1, 2021, 758–71. URL

Jennifer Cobbe, Michelle Seng Lee, and Jatinder Singh. “Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems.” FAccT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 1, 2021, 598–609. URL

MEASURE 2.9

The AI model is explained, validated, and documented, and AI system output is interpreted within its context – as identified in the MAP function – and to inform responsible use and governance.

About

Explainability and interpretability assist those operating or overseeing an AI system, as well as users of an AI system, to gain deeper insights into the functionality and trustworthiness of the system, including its outputs.

Explainable and interpretable AI systems offer information that help end users understand the purposes and potential impact of an AI system. Risk from lack of explainability may be managed by describing how AI systems function, with descriptions tailored to individual differences such as the user’s role, knowledge, and skill level. Explainable systems can be debugged and monitored more easily, and they lend themselves to more thorough documentation, audit, and governance.

Risks to interpretability often can be addressed by communicating a description of why an AI system made a particular prediction or recommendation.

Transparency, explainability, and interpretability are distinct characteristics that support each other. Transparency can answer the question of “what happened”. Explainability can answer the question of “how” a decision was made in the system. Interpretability can answer the question of “why” a decision was made by the system and its meaning or context to the user.

Suggested Actions

- Verify systems are developed to produce explainable models, post-hoc explanations and audit logs.
- When possible or available, utilize approaches that are inherently explainable, such as traditional and penalized generalized linear models , decision trees, nearest-neighbor and prototype-based approaches, rule-based models, generalized additive models , explainable boosting machines and neural additive models.

- Test explanation methods and resulting explanations prior to deployment to gain feedback from relevant AI actors, end users, and potentially impacted individuals or groups about whether explanations are accurate, clear, and understandable.
- Document AI model details including model type (e.g., convolutional neural network, reinforcement learning, decision tree, random forest, etc.) data features, training algorithms, proposed uses, decision thresholds, training data, evaluation data, and ethical considerations.
- Establish, document, and report performance and error metrics across demographic groups and other segments relevant to the deployment context.
- Explain systems using a variety of methods, e.g., visualizations, model extraction, feature importance, and others. Since explanations may not accurately summarize complex systems, test explanations according to properties such as fidelity, consistency, robustness, and interpretability.
- Assess the characteristics of system explanations according to properties such as fidelity (local and global), ambiguity, interpretability, interactivity, consistency, and resilience to attack/manipulation.
- Test the quality of system explanations with end-users and other groups.
- Secure model development processes to avoid vulnerability to external manipulation such as gaming explanation processes.
- Test for changes in models over time, including for models that adjust in response to production data.
- Use transparency tools such as data statements and model cards to document explanatory and validation information.

Transparency and Documentation

Organizations can document the following:

- Given the purpose of the AI, what level of explainability or interpretability is required for how the AI made its determination?
- Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
- How has the entity documented the AI system's data provenance, including sources, origins, transformations, augmentations, labels, dependencies, constraints, and metadata?
- What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework- WEF - Companion to the Model AI Governance Framework, 2020. URL

References

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. "This Looks Like That: Deep Learning for Interpretable Image Recognition." arXiv preprint, submitted December 28, 2019. URL

Cynthia Rudin. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." arXiv preprint, submitted September 22, 2019. URL

David A. Broniatowski. “NISTIR 8367 Psychological Foundations of Explainability and Interpretability in Artificial Intelligence. National Institute of Standards and Technology (NIST), 2021. URL

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities, and Challenges Toward Responsible AI.” Information Fusion 58 (June 2020): 82–115. URL

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. “Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems.” IUI ’20: Proceedings of the 25th International Conference on Intelligent User Interfaces, March 17, 2020, 454–64. URL

P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, Amy N. Yates, Kristen Greene, David A. Broniatowski, and Mark A. Przybocki. “NISTIR 8312 Four Principles of Explainable Artificial Intelligence.” National Institute of Standards and Technology (NIST), September 2021. URL

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. “Model Cards for Model Reporting.” FAT *19: Proceedings of the Conference on Fairness, Accountability, and Transparency, January 2019, 220–29. URL

Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. “A Nutritional Label for Rankings.” SIGMOD ’18: Proceedings of the 2018 International Conference on Management of Data, May 27, 2018, 1773–76. URL

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “ ‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier.” arXiv preprint, submitted August 9, 2016. URL

Scott M. Lundberg and Su-In Lee. “A unified approach to interpreting model predictions.” NIPS’17: Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4, 2017, 4768-4777. URL

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. “Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods.” AIES ’20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, February 7, 2020, 180–86. URL

David Alvarez-Melis and Tommi S. Jaakkola. “Towards robust interpretability with self-explaining neural networks.” NIPS’18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, December 3, 2018, 7786-7795. URL

FinRegLab, Laura Biattner, and Jann Spiess. “Machine Learning Explainability & Fairness: Insights from Consumer Lending.” FinRegLab, April 2022. URL

Miguel Ferreira, Muhammad Bilal Zafar, and Krishna P. Gummadi. “The Case for Temporal Transparency: Detecting Policy Change Events in Black-Box Decision Making Systems.” arXiv preprint, submitted October 31, 2016. URL

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. “Interpretable & Explorable Approximations of Black Box Models.” arXiv preprint, July 4, 2017. URL

Software Resources

- SHAP
- LIME
- Interpret
- PiML
- Iml
- Dalex

MEASURE 2.10

Privacy risk of the AI system – as identified in the MAP function – is examined and documented.

About

Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. These norms and practices typically address freedom from intrusion, limiting observation, or individuals' agency to consent to disclosure or control of facets of their identities (e.g., body, data, reputation).

Privacy values such as anonymity, confidentiality, and control generally should guide choices for AI system design, development, and deployment. Privacy-related risks may influence security, bias, and transparency and come with tradeoffs with these other characteristics. Like safety and security, specific technical features of an AI system may promote or reduce privacy. AI systems can also present new risks to privacy by allowing inference to identify individuals or previously private information about individuals.

Privacy-enhancing technologies ("PETs") for AI, as well as data minimizing methods such as de-identification and aggregation for certain model outputs, can support design for privacy-enhanced AI systems. Under certain conditions such as data sparsity, privacy enhancing techniques can result in a loss in accuracy, impacting decisions about fairness and other values in certain domains.

Suggested Actions

- Specify privacy-related values, frameworks, and attributes that are applicable in the context of use through direct engagement with end users and potentially impacted groups and communities.
- Document collection, use, management, and disclosure of personally sensitive information in datasets, in accordance with privacy and data governance policies
- Quantify privacy-level data aspects such as the ability to identify individuals or groups (e.g. k-anonymity metrics, l-diversity, t-closeness).
- Establish and document protocols (authorization, duration, type) and access controls for training sets or production data containing personally sensitive information, in accordance with privacy and data governance policies.
- Monitor internal queries to production data for detecting patterns that isolate personal records.
- Monitor PSI disclosures and inference of sensitive or legally protected attributes
  - Assess the risk of manipulation from overly customized content. Evaluate information presented to representative users at various points along axes of difference between individuals (e.g. individuals of different ages, genders, races, political affiliation, etc.).
- Use privacy-enhancing techniques such as differential privacy, when publicly sharing dataset information.
- Collaborate with privacy experts, AI end users and operators, and other domain experts to determine optimal differential privacy metrics within contexts of use.

Transparency and Documentation

Organizations can document the following:

- Did your organization implement accountability-based practices in data management and protection (e.g. the PDPA and OECD Privacy Principles)?
- What assessments has the entity conducted on data security and privacy impacts associated with the AI system?
- Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g. personnel risk or changes to commercial objectives)?
- Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information)
- If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial, social or otherwise) What was done to mitigate or reduce the potential for harm?

AI Transparency Resources:

- WEF Companion to the Model AI Governance Framework- WEF - Companion to the Model AI Governance Framework, 2020. (URL
- Datasheets for Datasets. URL

References

Kaitlin R. Boeckl and Naomi B. Lefkovitz. “NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0.” National Institute of Standards and Technology (NIST), January 16, 2020. URL

Latanya Sweeney. “K-Anonymity: A Model for Protecting Privacy.” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, no. 5 (2002): 557–70. URL

Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. “L-Diversity: Privacy beyond K-Anonymity.” 22nd International Conference on Data Engineering (ICDE’06), 2006. URL

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. “CERIAS Tech Report 2007-78 t-Closeness: Privacy Beyond k-Anonymity and -Diversity.” Center for Education and Research, Information Assurance and Security, Purdue University, 2001. URL

J. Domingo-Ferrer and J. Soria-Comas. “From t-closeness to differential privacy and vice versa in data anonymization.” arXiv preprint, submitted December 21, 2015. URL

Joseph Near, David Darais, and Kaitlin Boeckly. “Differential Privacy for Privacy-Preserving Data Analysis: An Introduction to our Blog Series.” National Institute of Standards and Technology (NIST), July 27, 2020. URL

Cynthia Dwork. “Differential Privacy.” Automata, Languages and Programming, 2006, 1–12. URL

Zhanglong Ji, Zachary C. Lipton, and Charles Elkan. “Differential Privacy and Machine Learning: a Survey and Review.” arXiv preprint, submitted December 24,2014. URL

Michael B. Hawes. “Implementing Differential Privacy: Seven Lessons From the 2020 United States Census.” Harvard Data Science Review 2, no. 2 (2020). URL

Harvard University Privacy Tools Project. “Differential Privacy.” Harvard University, n.d. URL

John M. Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Matthew Spence Sexton and Pavel Zhuravlev. “The 2020 Census Disclosure Avoidance System TopDown Algorithm.” United States Census Bureau, April 7, 2022. URL

Nicolas Papernot and Abhradeep Guha Thakurta. “How to deploy machine learning with differential privacy.” National Institute of Standards and Technology (NIST), December 21, 2021. URL

Claire McKay Bowen. “Utility Metrics for Differential Privacy: No One-Size-Fits-All.” National Institute of Standards and Technology (NIST), November 29, 2021. URL

Helen Nissenbaum. “Contextual Integrity Up and Down the Data Food Chain.” Theoretical Inquiries in Law 20, L. 221 (2019): 221-256. URL

Sebastian Benthall, Seda Gürses, and Helen Nissenbaum. “Contextual Integrity through the Lens of Computer Science.” Foundations and Trends in Privacy and Security 2, no. 1 (December 22, 2017): 1–69. URL

Jenifer Sunrise Winter and Elizabeth Davidson. “Big Data Governance of Personal Health Information and Challenges to Contextual Integrity.” The Information Society: An International Journal 35, no. 1 (2019): 36–51. [URL](https://doi.org/10.1080/01972243.2018.1542648.

MEASURE 2.11

Fairness and bias – as identified in the MAP function – is evaluated and results are documented.

About

Fairness in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination. Standards of fairness can be complex and difficult to define because perceptions of fairness differ among cultures and may shift depending on application. Organizations’ risk management efforts will be enhanced by recognizing and considering these differences. Systems in which harmful biases are mitigated are not necessarily fair. For example, systems in which predictions are somewhat balanced across demographic groups may still be inaccessible to individuals with disabilities or affected by the digital divide or may exacerbate existing disparities or systemic biases.

Bias is broader than demographic balance and data representativeness. NIST has identified three major categories of AI bias to be considered and managed: systemic, computational and statistical, and human-cognitive. Each of these can occur in the absence of prejudice, partiality, or discriminatory intent. - Systemic bias can be present in AI datasets, the organizational norms, practices, and processes across the AI lifecycle, and the broader society that uses AI systems. - Computational and statistical biases can be present in AI datasets and algorithmic processes, and often stem from systematic errors due to non-representative samples. - Human-cognitive biases relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about purposes and functions of an AI system. Human-cognitive biases are omnipresent in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI.

Bias exists in many forms and can become ingrained in the automated systems that help make decisions about our lives. While bias is not always a negative phenomenon, AI systems can potentially increase the speed and scale of biases and perpetuate and amplify harms to individuals, groups, communities, organizations, and society.

Suggested Actions

- Conduct fairness assessments to manage computational and statistical forms of bias which include the following steps:
  - Identify types of harms, including allocational, representational, quality of service, stereotyping, or erasure
  - Identify across, within, and intersecting groups that might be harmed
  - Quantify harms using both a general fairness metric, if appropriate (e.g. demographic parity, equalized odds, equal opportunity, statistical hypothesis tests), and custom, context-specific metrics developed in collaboration with affected communities
  - Analyze quantified harms for contextually significant differences across groups, within groups, and among intersecting groups
  - Refine identification of within-group and intersectional group disparities.
    - Evaluate underlying data distributions and employ sensitivity analysis during the analysis of quantified harms.
    - Evaluate quality metrics including false positive rates and false negative rates.
    - Consider biases affecting small groups, within-group or intersectional communities, or single individuals.
- Understand and consider sources of bias in training and TEVV data:
  - Differences in distributions of outcomes across and within groups, including intersecting groups.
  - Completeness, representativeness and balance of data sources.
  - Identify input data features that may serve as proxies for demographic group membership (i.e., credit score, ZIP code) or otherwise give rise to emergent bias within AI systems.
  - Forms of systemic bias in images, text (or word embeddings), audio or other complex or unstructured data.
- Leverage impact assessments to identify and classify system impacts and harms to end users, other individuals, and groups with input from potentially impacted communities.
- Identify the classes of individuals, groups, or environmental ecosystems which might be impacted through direct engagement with potentially impacted communities.
- Evaluate systems in regards to disability inclusion, including consideration of disability status in bias testing, and discriminatory screen out processes that may arise from non-inclusive design or deployment decisions.
- Develop objective functions in consideration of systemic biases, in-group/out-group dynamics.
- Use context-specific fairness metrics to examine how system performance varies across groups, within

groups, and/or for intersecting groups. Metrics may include statistical parity, error-rate equality, statistical parity difference, equal opportunity difference, average absolute odds difference, standardized mean difference, percentage point differences.

- Customize fairness metrics to specific context of use to examine how system performance and potential harms vary within contextual norms.
- Define acceptable levels of difference in performance in accordance with established organizational governance policies, business requirements, regulatory compliance, legal frameworks, and ethical standards within the context of use
- Define the actions to be taken if disparity levels rise above acceptable levels.
- Identify groups within the expected population that may require disaggregated analysis, in collaboration with impacted communities.
- Leverage experts with knowledge in the specific context of use to investigate substantial measurement differences and identify root causes for those differences.
- Monitor system outputs for performance or bias issues that exceed established tolerance levels.
- Ensure periodic model updates; test and recalibrate with updated and more representative data to stay within acceptable levels of difference.
- Apply pre-processing data transformations to address factors related to demographic balance and data representativeness.
- Apply in-processing to balance model performance quality with bias considerations.
- Apply post-processing mathematical/computational techniques to model results in close collaboration with impact assessors, socio-technical experts, and other AI actors with expertise in the context of use.
- Apply model selection approaches with transparent and deliberate consideration of bias management and other trustworthy characteristics.
- Collect and share information about differences in outcomes for the identified groups.
- Consider mediations to mitigate differences, especially those that can be traced to past patterns of unfair or biased human decision making.
- Utilize human-centered design practices to generate deeper focus on societal impacts and counter human-cognitive biases within the AI lifecycle.
- Evaluate practices along the lifecycle to identify potential sources of human-cognitive bias such as availability, observational, and confirmation bias, and to make implicit decision making processes more explicit and open to investigation.
- Work with human factors experts to evaluate biases in the presentation of system output to end users, operators and practitioners.
- Utilize processes to enhance contextual awareness, such as diverse internal staff and stakeholder engagement.

Transparency and Documentation

Organizations can document the following:

- To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?
- If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?
- Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
- How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?
- To what extent has the entity identified and mitigated potential bias—statistical, contextual, and historical—in the data?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

- WEF Companion to the Model AI Governance Framework- WEF - Companion to the Model AI Governance Framework, 2020. URL
- Datasheets for Datasets. URL

References

Ali Hasan, Shea Brown, Jovana Davidovic, Benjamin Lange, and Mitt Regan. “Algorithmic Bias and Risk Assessments: Lessons from Practice.” Digital Society 1 (2022). URL

Richard N. Landers and Tara S. Behrend. “Auditing the AI Auditors: A Framework for Evaluating Fairness and Bias in High Stakes AI Predictive Models.” American Psychologist 78, no. 1 (2023): 36–49. URL

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. “A Survey on Bias and Fairness in Machine Learning.” ACM Computing Surveys 54, no. 6 (July 2021): 1–35. URL

Michele Loi and Christoph Heitz. “Is Calibration a Fairness Requirement?” FAccT ’22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 2022, 2026–34. URL

Shea Brown, Ryan Carrier, Merve Hickok, and Adam Leon Smith. “Bias Mitigation in Data Sets.” SocArXiv, July 8, 2021. URL

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. “NIST Special Publication 1270 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.” National Institute of Standards and Technology (NIST), 2022. URL

Microsoft Research. “AI Fairness Checklist.” Microsoft, February 7, 2022. URL

Samir Passi and Solon Barocas. “Problem Formulation and Fairness.” FAT* ’19: Proceedings of the Conference on Fairness, Accountability, and Transparency, January 2019, 39–48. URL

Jade S. Franklin, Karan Bhanot, Mohamed Ghalwash, Kristin P. Bennett, Jamie McCusker, and Deborah L. McGuinness. “An Ontology for Fairness Metrics.” AIES ’22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, July 2022, 265–75. URL

Arvind Narayanan. “Tl;DS - 21 Fairness Definition and Their Politics by Arvind Narayanan.” Dora’s world, July 19, 2019. URL

Ben Green. “Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness.” Philosophy and Technology 35, no. 90 (October 8, 2022). URL

Alexandra Chouldechova. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” Big Data 5, no. 2 (June 1, 2017): 153–63. URL

Sina Fazelpour and Zachary C. Lipton. “Algorithmic Fairness from a Non-Ideal Perspective.” AIES ’20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, February 7, 2020, 57–63. URL

Hemank Lamba, Kit T. Rodolfa, and Rayid Ghani. “An Empirical Comparison of Bias Reduction Methods on Real-World Problems in High-Stakes Policy Settings.” ACM SIGKDD Explorations Newsletter 23, no. 1 (May 29, 2021): 69–85. URL

ISO. “ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making.” ISO Standards, November 2021. URL

Shari Trewin. “AI Fairness for People with Disabilities: Point of View.” arXiv preprint, submitted November 26, 2018. URL

MathWorks. “Explore Fairness Metrics for Credit Scoring Model.” MATLAB & Simulink, 2023. URL

Abigail Z. Jacobs and Hanna Wallach. “Measurement and Fairness.” FAccT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021, 375–85. URL

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. “Quantifying and Reducing Stereotypes in Word Embeddings.” arXiv preprint, submitted June 20, 2016. URL

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. “Semantics Derived Automatically from Language Corpora Contain Human-Like Biases.” Science 356, no. 6334 (April 14, 2017): 183–86. URL

Sina Fazelpour and Maria De-Arteaga. “Diversity in Sociotechnical Machine Learning Systems.” Big Data and Society 9, no. 1 (2022). URL

Fairlearn. “Fairness in Machine Learning.” Fairlearn 0.8.0 Documentation, n.d. URL

Safiya Umoja Noble. Algorithms of Oppression: How Search Engines Reinforce Racism. New York, NY: New York University Press, 2018. URL

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” Science 366, no. 6464 (October 25, 2019): 447–53. URL

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. “A Reductions Approach to Fair Classification.” arXiv preprint, submitted July 16, 2018. URL

Moritz Hardt, Eric Price, and Nathan Srebro. “Equality of Opportunity in Supervised Learning.” arXiv preprint, submitted October 7, 2016. URL

Alekh Agarwal, Miroslav Dudik, Zhiwei Steven Wu. “Fair Regression: Quantitative Definitions and Reduction-Based Algorithms.” Proceedings of the 36th International Conference on Machine Learning, PMLR 97:120-129, 2019. URL

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. “Fairness and Abstraction in Sociotechnical Systems.” FAT* ’19: Proceedings of the Conference on Fairness, Accountability, and Transparency, January 29, 2019, 59–68. URL

Matthew Kay, Cynthia Matuszek, and Sean A. Munson. “Unequal Representation and Gender Stereotypes in Image Search Results for Occupations.” CHI ’15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, April 18, 2015, 3819–28. URL

Software Resources

- aequitas
- AI Fairness 360:
  - Python
  - R
- algofairness
- fairlearn
- fairml
- fairmodels
- fairness
- solas-ai-disparity
- tensorflow/fairness-indicators
- Themis

MEASURE 2.12

TEnvironmental impact and sustainability of AI model training and management activities – as identified in the MAP function – are assessed and documented.

About

Large-scale, high-performance computational resources used by AI systems for training and operation can contribute to environmental impacts. Direct negative impacts to the environment from these processes are related to energy consumption, water consumption, and greenhouse gas (GHG) emissions. The OECD has identified metrics for each type of negative direct impact.

Indirect negative impacts to the environment reflect the complexity of interactions between human behavior, socio-economic systems, and the environment and can include induced consumption and "rebound effects", where efficiency gains are offset by accelerated resource consumption.

Other AI related environmental impacts can arise from the production of computational equipment and networks (e.g. mining and extraction of raw materials), transporting hardware, and electronic waste recycling or disposal.

Suggested Actions

- Include environmental impact indicators in AI system design and development plans, including reducing consumption and improving efficiencies.
- Identify and implement key indicators of AI system energy and water consumption and efficiency, and/or GHG emissions.
- Establish measurable baselines for sustainable AI system operation in accordance with organizational policies, regulatory compliance, legal frameworks, and environmental protection and sustainability norms.
- Assess tradeoffs between AI system performance and sustainable operations in accordance with organizational principles and policies, regulatory compliance, legal frameworks, and environmental protection and sustainability norms.
- Identify and establish acceptable resource consumption and efficiency, and GHG emissions levels, along with actions to be taken if indicators rise above acceptable levels.
- Estimate AI system emissions levels throughout the AI lifecycle via carbon calculators or similar process.

Transparency and Documentation

Organizations can document the following:

- How will the accuracy or appropriate performance metrics be assessed?
- How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed? How much distributional shift or model drift from baseline performance is acceptable?
- What corrective actions has the entity taken to enhance the quality, accuracy, reliability, and representativeness of the data?
- Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- Datasheets for Datasets. URL

References

Organisation for Economic Co-operation and Development (OECD). "Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint." OECD Digital Economy Papers, No. 341, OECD Publishing, Paris. URL

Victor Schmidt, Alexandra Luccioni, Alexandre Lacoste, and Thomas Dandres. "Machine Learning CO2 Impact Calculator." ML CO2 Impact, n.d. URL

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. "Quantifying the Carbon Emissions of Machine Learning." arXiv preprint, submitted November 4, 2019. URL

Matthew Hutson. "Measuring AI's Carbon Footprint: New Tools Track and Reduce Emissions from Machine Learning." IEEE Spectrum, November 22, 2022. URL

Association for Computing Machinery (ACM). "TechBriefs: Computing and Climate Change." ACM Technology Policy Council, November 2021. URL

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. "Green AI." Communications of the ACM 63, no. 12 (December 2020): 54–63. URL

MEASURE 2.13

Effectiveness of the employed TEVV metrics and processes in the MEASURE function are evaluated and documented.

About

The development of metrics is a process often considered to be objective but, as a human and organization driven endeavor, can reflect implicit and systemic biases, and may inadvertently reflect factors unrelated to the target function. Measurement approaches can be oversimplified, gamed, lack critical nuance, become used and relied upon in unexpected ways, fail to account for differences in affected groups and contexts.

Revisiting the metrics chosen in Measure 2.1 through 2.12 in a process of continual improvement can help AI actors to evaluate and document metric effectiveness and make necessary course corrections.

Suggested Actions

- Review selected system metrics and associated TEVV processes to determine if they are able to sustain system improvements, including the identification and removal of errors.
- Regularly evaluate system metrics for utility, and consider descriptive approaches in place of overly complex methods.
- Review selected system metrics for acceptability within the end user and impacted community of interest.
- Assess effectiveness of metrics for identifying and measuring risks.

Transparency and Documentation

Organizations can document the following:

- To what extent does the system/entity consistently measure progress towards stated goals and objectives?
- Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
- What corrective actions has the entity taken to enhance the quality, accuracy, reliability, and representativeness of the data?
- To what extent are the model outputs consistent with the entity's values and principles to foster public trust and equity?
- How will the accuracy or appropriate performance metrics be assessed?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

Arvind Narayanan. "The limits of the quantitative approach to discrimination." 2022 James Baldwin lecture, Princeton University, October 11, 2022. URL

Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. "A Human-Centered Review of Algorithms Used within the U.S. Child Welfare System." CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, April 23, 2020, 1–15. URL

Rachel Thomas and David Uminsky. "Reliance on Metrics Is a Fundamental Challenge for AI." Patterns 3, no. 5 (May 13, 2022): 100476. URL

Momin M. Malik. "A Hierarchy of Limitations in Machine Learning." arXiv preprint, submitted February 29, 2020. [URL](https://arxiv.org/abs/2002.05193

MEASURE-3: Mechanisms for tracking identified AI risks over time are in place.

MEASURE 3.1

Approaches, personnel, and documentation are in place to regularly identify and track existing, unanticipated, and emergent AI risks based on factors such as intended and actual performance in deployed contexts.

About

For trustworthy AI systems, regular system monitoring is carried out in accordance with organizational governance policies, AI actor roles and responsibilities, and within a culture of continual improvement. If and when emergent or complex risks arise, it may be necessary to adapt internal risk management procedures, such as regular monitoring, to stay on course. Documentation, resources, and training are part of an overall strategy to support AI actors as they investigate and respond to AI system errors, incidents or negative impacts.

Suggested Actions

- Compare AI system risks with:
  - simpler or traditional models
  - human baseline performance
  - other manual performance benchmarks
- Compare end user and community feedback about deployed AI systems to internal measures of system performance.
- Assess effectiveness of metrics for identifying and measuring emergent risks.
- Measure error response times and track response quality.
- Elicit and track feedback from AI actors in user support roles about the type of metrics, explanations and other system information required for fulsome resolution of system issues. Consider:
  - Instances where explanations are insufficient for investigating possible error sources or identifying responses.
  - System metrics, including system logs and explanations, for identifying and diagnosing sources of system error.
- Elicit and track feedback from AI actors in incident response and support roles about the adequacy of staffing and resources to perform their duties in an effective and timely manner.

Transparency and Documentation

Organizations can document the following:

- Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g. personnel risk or changes to commercial objectives)?
- To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?
- What metrics has the entity developed to measure performance of the AI system, including error logging?
- To what extent do the metrics provide accurate and useful measure of performance?
- What testing, if any, has the entity conducted o

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations URL

References

ISO. "ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems.” 2nd ed. ISO Standards, July 2019. URL

Larysa Visengeriyeva, et al. “Awesome MLOps.“ GitHub. URL

MEASURE 3.2

Risk tracking approaches are considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available.

About

Risks identified in the Map function may be complex, emerge over time, or difficult to measure. Systematic methods for risk tracking, including novel measurement approaches, can be established as part of regular monitoring and improvement processes.

Suggested Actions

- Establish processes for tracking emergent risks that may not be measurable with current approaches. Some processes may include:
  - Recourse mechanisms for faulty AI system outputs.
  - Bug bounties.
  - Human-centered design approaches.
  - User-interaction and experience research.
  - Participatory stakeholder engagement with affected or potentially impacted individuals and communities.
- Identify AI actors responsible for tracking emergent risks and inventory methods.
- Determine and document the rate of occurrence and severity level for complex or difficult-to-measure risks when:
  - Prioritizing new measurement approaches for deployment tasks.
  - Allocating AI system risk management resources.
  - Evaluating AI system improvements.
  - Making go/no-go decisions for subsequent system iterations.

Transparency and Documentation

Organizations can document the following:

- Who is ultimately responsible for the decisions of the AI and is this person aware of the intended uses and limitations of the analytic?
- Who will be responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed?
- To what extent does the entity communicate its AI strategic goals and objectives to the community of stakeholders?
- Given the purpose of this AI, what is an appropriate interval for checking whether it is still accurate, unbiased, explainable, etc.? What are the checks for this model?
- If anyone believes that the AI no longer meets this ethical framework, who will be responsible for receiving the concern and as appropriate investigating and remediating the issue? Do they have authority to modify, limit, or stop the use of the AI?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

ISO. “ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems.” 2nd ed. ISO Standards, July 2019. URL

Mark C. Paulk, Bill Curtis, Mary Beth Chrissis, and Charles V. Weber. "Capability Maturity Model, Version 1.1." IEEE Software 10, no. 4 (1993): 18–27. URL

Jeff Patton, Peter Economy, Martin Fowler, Alan Cooper, and Marty Cagan. User Story Mapping: Discover the Whole Story, Build the Right Product. O'Reilly, 2014. URL

Rumman Chowdhury and Jutta Williams. "Introducing Twitter's first algorithmic bias bounty challenge." Twitter Engineering Blog, July 30, 2021. URL

HackerOne. "Twitter Algorithmic Bias." HackerOne, August 8, 2021. URL

Josh Kenway, Camille François, Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. "Bug Bounties for Algorithmic Harms?" Algorithmic Justice League, January 2022. URL

Microsoft. "Community Jury." Microsoft Learn's Azure Application Architecture Guide, 2023. URL

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. "Overcoming Failures of Imagination in AI Infused System Development and Deployment." arXiv preprint, submitted December 10, 2020. URL

MEASURE 3.3

Feedback processes for end users and impacted communities to report problems and appeal system outcomes are established and integrated into AI system evaluation metrics.

About

Assessing impact is a two-way effort. Many AI system outcomes and impacts may not be visible or recognizable to AI actors across the development and deployment dimensions of the AI lifecycle, and may require direct feedback about system outcomes from the perspective of end users and impacted groups.

Feedback can be collected indirectly, via systems that are mechanized to collect errors and other feedback from end users and operators

Metrics and insights developed in this sub-category feed into Manage 4.1 and 4.2.

Suggested Actions

- Measure efficacy of end user and operator error reporting processes.
- Categorize and analyze type and rate of end user appeal requests and results.
- Measure feedback activity participation rates and awareness of feedback activity availability.
- Utilize feedback to analyze measurement approaches and determine subsequent courses of action.
- Evaluate measurement approaches to determine efficacy for enhancing organizational understanding of real world impacts.
- Analyze end user and community feedback in close collab

Transparency and Documentation

Organizations can document the following:

- To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?
- Did your organization address usability problems and test whether user interfaces served their intended purposes?
- How easily accessible and current is the information available to external stakeholders?
- What type of information is accessible on the design, operations, and limitations of the AI system to external stakeholders, including end users, consumers, regulators, and individuals impacted by use of the AI system?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL

- WEF Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations URL

References

Sasha Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: The MIT Press, 2020. URL

David G. Robinson. Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. New York: Russell Sage Foundation, 2022. URL

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir.' " arXiv preprint, submitted November 1, 2021. URL

George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. "Human-Centered Design of Artificial Intelligence." In Handbook of Human Factors and Ergonomics, edited by Gavriel Salvendy and Waldemar Karwowski, 5th ed., 1085–1106. John Wiley & Sons, 2021. URL

Ben Shneiderman. Human-Centered AI. Oxford: Oxford University Press, 2022

Batya Friedman, David G. Hendry, and Alan Borning. "A Survey of Value Sensitive Design Methods." Foundations and Trends in Human-Computer Interaction 11, no. 2 (November 22, 2017): 63–125. URL

Batya Friedman, Peter H. Kahn, Jr., and Alan Borning. "Value Sensitive Design: Theory and Methods." University of Washington Department of Computer Science & Engineering Technical Report 02-12-01, December 2002. URL

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest." SSRN, July 8, 2021. URL

Alexandra Reeve Givens, and Meredith Ringel Morris. "Centering Disability Perspectives in Algorithmic Fairness, Accountability, & Transparency." FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27, 2020, 684-84. URL

MEASURE-4: Feedback about efficacy of measurement is gathered and assessed.

MEASURE 4.1

Measurement approaches for identifying AI risks are connected to deployment context(s) and informed through consultation with domain experts and other end users. Approaches are documented.

About

AI Actors carrying out TEVV tasks may have difficulty evaluating impacts within the system context of use. AI system risks and impacts are often best described by end users and others who may be affected by output and subsequent decisions. AI Actors can elicit feedback from impacted individuals and communities via participatory engagement processes established in Govern 5.1 and 5.2, and carried out in Map 1.6, 5.1, and 5.2.

Activities described in the Measure function enable AI actors to evaluate feedback from impacted individuals and communities. To increase awareness of insights, feedback can be evaluated in close collaboration with AI actors responsible for impact assessment, human-factors, and governance and oversight tasks, as well as with other socio-technical domain experts and researchers. To gain broader expertise for interpreting evaluation outcomes, organizations may consider collaborating with advocacy groups and civil society organizations.

Insights based on this type of analysis can inform TEVV-based decisions about metrics and related courses of action.

Suggested Actions

- Support mechanisms for capturing feedback from system end users (including domain experts, operators, and practitioners). Successful approaches are:
  - conducted in settings where end users are able to openly share their doubts and insights about AI system output, and in connection to their specific context of use (including setting and task-specific lines of inquiry)
  - developed and implemented by human-factors and socio-technical domain experts and researchers
  - designed to ensure control of interviewer and end user subjectivity and biases
- Identify and document approaches
  - for evaluating and integrating elicited feedback from system end users
  - in collaboration with human-factors and socio-technical domain experts,
  - to actively inform a process of continual improvement.
- Evaluate feedback from end users alongside evaluated feedback from impacted communities (MEASURE 3.3).
- Utilize end user feedback to investigate how selected metrics and measurement approaches interact with organizational and operational contexts.
- Analyze and document system-internal measurement processes in comparison to collected end user feedback.
- Identify and implement approaches to measure effectiveness and satisfaction with end user elicitation techniques, and document results.

Transparency and Documentation

Organizations can document the following:

- Did your organization address usability problems and test whether user interfaces served their intended purposes?
- How will user and peer engagement be integrated into the model development process and periodic performance review once deployed?
- To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?
- To what extent are the established procedures effective in mitigating bias, inequity, and other concerns resulting from the system?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL
- WEF Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations URL

References

Batya Friedman, and David G. Hendry. Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: The MIT Press, 2019. URL

Batya Friedman, David G. Hendry, and Alan Borning. "A Survey of Value Sensitive Design Methods." Foundations and Trends in Human-Computer Interaction 11, no. 2 (November 22, 2017): 63–125. URL

Steven Umbrello, and Ibo van de Poel. "Mapping Value Sensitive Design onto AI for Social Good Principles." AI and Ethics 1, no. 3 (February 1, 2021): 283–96. URL

Karen Boyd. "Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development." FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 20, 2022, 2069–82. URL

Janet Davis and Lisa P. Nathan. "Value Sensitive Design: Applications, Adaptations, and Critiques." In Handbook of Ethics, Values, and Technological Design, edited by Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel, January 1, 2015, 11–40. URL

Ben Shneiderman. Human-Centered AI. Oxford: Oxford University Press, 2022.

Shneiderman, Ben. "Human-Centered AI." Issues in Science and Technology 37, no. 2 (2021): 56–61. URL

Shneiderman, Ben. "Tutorial: Human-Centered AI: Reliable, Safe and Trustworthy." IUI '21 Companion: 26th International Conference on Intelligent User Interfaces - Companion, April 14, 2021, 7–8. URL

George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. "Human-Centered Design of Artificial Intelligence." In Handbook of Human Factors and Ergonomics, edited by Gavriel Salvendy and Waldemar Karwowski, 5th ed., 1085–1106. John Wiley & Sons, 2021. URL

Caitlin Thompson. "Who's Homeless Enough for Housing? In San Francisco, an Algorithm Decides." Coda, September 21, 2021. URL

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. "Algorithmic Decision-Making and the Control Problem." Minds and Machines 29, no. 4 (December 11, 2019): 555–78. URL

Fry, Hannah. Hello World: Being Human in the Age of Algorithms. New York: W.W. Norton & Company, 2018. URL

Sasha Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: The MIT Press, 2020. URL

David G. Robinson. Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. New York: Russell Sage Foundation, 2022. URL

Diane Hart, Gabi Diercks-O'Brien, and Adrian Powell. "Exploring Stakeholder Engagement in Impact Evaluation Planning in Educational Development Work." Evaluation 15, no. 3 (2009): 285–306. URL

Asit Bhattacharyya and Lorne Cummings. "Measuring Corporate Environmental Performance – Stakeholder Engagement Evaluation." Business Strategy and the Environment 24, no. 5 (2013): 309–25. URL

Hendricks, Sharief, Nailah Conrad, Tania S. Douglas, and Tinashe Mutsvangwa. "A Modified Stakeholder Participation Assessment Framework for Design Thinking in Health Innovation." Healthcare 6, no. 3 (September 2018): 191–96. URL

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir.' " arXiv preprint, submitted November 1, 2021. URL

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest." SSRN, July 8, 2021. URL

Alexandra Reeve Givens, and Meredith Ringel Morris. "Centering Disability Perspectives in Algorithmic Fairness, Accountability, & Transparency." FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27, 2020, 684-84. URL

MEASURE 4.2

Measurement results regarding AI system trustworthiness in deployment context(s) and across AI lifecycle are informed by input from domain experts and other relevant AI actors to validate whether the system is performing consistently as intended. Results are documented.

About

Feedback captured from relevant AI Actors can be evaluated in combination with output from Measure 2.5 to 2.11 to determine if the AI system is performing within pre-defined operational limits for validity and reliability, safety, security and resilience, privacy, bias and fairness, explainability and interpretability, and transparency and accountability. This feedback provides an additional layer of insight about AI system performance, including potential misuse or reuse outside of intended settings.

Insights based on this type of analysis can inform TEVV-based decisions about metrics and related courses of action.

Suggested Actions

- Integrate feedback from end users, operators, and affected individuals and communities from Map function as inputs to assess AI system trustworthiness characteristics. Ensure both positive and negative feedback is being assessed.
- Evaluate feedback in connection with AI system trustworthiness characteristics from Measure 2.5 to 2.11.
- Evaluate feedback regarding end user satisfaction with, and confidence in, AI system performance including whether output is considered valid and reliable, and explainable and interpretable.
- Identify mechanisms to confirm/support AI system output (e.g. recommendations), and end user perspectives about that output.
- Measure frequency of AI systems' override decisions, evaluate and document results, and feed insights back into continual improvement processes.
- Consult AI actors in impact assessment, human factors and socio-technical tasks to assist with analysis and interpretation of results.

Transparency and Documentation

Organizations can document the following:

- To what extent does the system/entity consistently measure progress towards stated goals and objectives?
- What policies has the entity developed to ensure the use of the AI system is consistent with its stated values and principles?
- To what extent are the model outputs consistent with the entity's values and principles to foster public trust and equity?
- Given the purpose of the AI, what level of explainability or interpretability is required for how the AI made its determination?
- To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

Batya Friedman, and David G. Hendry. Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: The MIT Press, 2019. URL

Batya Friedman, David G. Hendry, and Alan Borning. “A Survey of Value Sensitive Design Methods.” Foundations and Trends in Human-Computer Interaction 11, no. 2 (November 22, 2017): 63–125. URL

Steven Umbrello, and Ibo van de Poel. “Mapping Value Sensitive Design onto AI for Social Good Principles.” AI and Ethics 1, no. 3 (February 1, 2021): 283–96. URL

Karen Boyd. “Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development.” FAccT ’22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 20, 2022, 2069–82. URL

Janet Davis and Lisa P. Nathan. “Value Sensitive Design: Applications, Adaptations, and Critiques.” In Handbook of Ethics, Values, and Technological Design, edited by Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel, January 1, 2015, 11–40. URL

Ben Shneiderman. Human-Centered AI. Oxford: Oxford University Press, 2022.

Shneiderman, Ben. “Human-Centered AI.” Issues in Science and Technology 37, no. 2 (2021): 56–61. URL

Shneiderman, Ben. “Tutorial: Human-Centered AI: Reliable, Safe and Trustworthy.” IUI ’21 Companion: 26th International Conference on Intelligent User Interfaces - Companion, April 14, 2021, 7–8. URL

George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. “Human-Centered Design of Artificial Intelligence.” In Handbook of Human Factors and Ergonomics, edited by Gavriel Salvendy and Waldemar Karwowski, 5th ed., 1085–1106. John Wiley & Sons, 2021. URL

Caitlin Thompson. “Who’s Homeless Enough for Housing? In San Francisco, an Algorithm Decides.” Coda, September 21, 2021. URL

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. “Algorithmic Decision-Making and the Control Problem.” Minds and Machines 29, no. 4 (December 11, 2019): 555–78. URL

Fry, Hannah. Hello World: Being Human in the Age of Algorithms. New York: W.W. Norton & Company, 2018. URL

Sasha Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: The MIT Press, 2020. URL

David G. Robinson. Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. New York: Russell Sage Foundation, 2022. URL

Diane Hart, Gabi Diercks-O’Brien, and Adrian Powell. “Exploring Stakeholder Engagement in Impact Evaluation Planning in Educational Development Work.” Evaluation 15, no. 3 (2009): 285–306. URL

Asit Bhattacharyya and Lorne Cummings. “Measuring Corporate Environmental Performance – Stakeholder Engagement Evaluation.” Business Strategy and the Environment 24, no. 5 (2013): 309–25. URL

Hendricks, Sharief, Nailah Conrad, Tania S. Douglas, and Tinashe Mutsvangwa. “A Modified Stakeholder Participation Assessment Framework for Design Thinking in Health Innovation.” Healthcare 6, no. 3 (September 2018): 191–96. URL

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. “Stakeholder Participation in AI: Beyond ‘Add Diverse Stakeholders and Stir.’ ” arXiv preprint, submitted November 1, 2021. URL

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. “Assembling Accountability: Algorithmic Impact Assessment for the Public Interest.” SSRN, July 8, 2021. URL

Alexandra Reeve Givens, and Meredith Ringel Morris. “Centering Disability Perspectives in Algorithmic Fairness, Accountability, & Transparency.” FAT* ’20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27, 2020, 684-84. URL

MEASURE 4.3

Measurable performance improvements or declines based on consultations with relevant AI actors including affected communities, and field data about context-relevant risks and trustworthiness characteristics, are identified and documented.

About

TEVV activities conducted throughout the AI system lifecycle can provide baseline quantitative measures for trustworthy characteristics. When combined with results from Measure 2.5 to 2.11 and Measure 4.1 and 4.2, TEVV actors can maintain a comprehensive view of system performance. These measures can be augmented through participatory engagement with potentially impacted communities or other forms of stakeholder elicitation about AI systems’ impacts. These sources of information can allow AI actors to explore potential adjustments to system components, adapt operating conditions, or institute performance improvements.

Suggested Actions

- Develop baseline quantitative measures for trustworthy characteristics.
- Delimit and characterize baseline operation values and states.
- Utilize qualitative approaches to augment and complement quantitative baseline measures, in close coordination with impact assessment, human factors and socio-technical AI actors.
- Monitor and assess measurements as part of continual improvement to identify potential system adjustments or modifications
- Perform and document sensitivity analysis to characterize actual and expected variance in performance after applying system or procedural updates.
- Document decisions related to the sensitivity analysis and record expected influence on system performance and identified risks.

Transparency and Documentation

Organizations can document the following:

- To what extent are the model outputs consistent with the entity’s values and principles to foster public trust and equity?
- How were sensitive variables (e.g., demographic and socioeconomic categories) that may be subject to regulatory compliance specifically selected or not selected for modeling purposes?
- Did your organization implement a risk management system to address risks involved in deploying the identified AI solution (e.g. personnel risk or changes to commercial objectives)?
- How will the accountable human(s) address changes in accuracy and precision due to either an adversary’s attempts to disrupt the AI or unrelated changes in the operational/business environment?
- How will user and peer engagement be integrated into the model development process and periodic performance review once deployed?

AI Transparency Resources:

- GAO-21-519SP - Artificial Intelligence: An Accountability Framework for Federal Agencies & Other Entities. URL
- Artificial Intelligence Ethics Framework For The Intelligence Community. URL

References

Batya Friedman, and David G. Hendry. Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: The MIT Press, 2019. URL

Batya Friedman, David G. Hendry, and Alan Borning. "A Survey of Value Sensitive Design Methods." Foundations and Trends in Human-Computer Interaction 11, no. 2 (November 22, 2017): 63–125. URL

Steven Umbrello, and Ibo van de Poel. "Mapping Value Sensitive Design onto AI for Social Good Principles." AI and Ethics 1, no. 3 (February 1, 2021): 283–96. URL

Karen Boyd. "Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development." FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, June 20, 2022, 2069–82. URL

Janet Davis and Lisa P. Nathan. "Value Sensitive Design: Applications, Adaptations, and Critiques." In Handbook of Ethics, Values, and Technological Design, edited by Jeroen van den Hoven, Pieter E. Vermaas, and Ibo van de Poel, January 1, 2015, 11–40. URL

Ben Shneiderman. Human-Centered AI. Oxford: Oxford University Press, 2022.

Shneiderman, Ben. "Human-Centered AI." Issues in Science and Technology 37, no. 2 (2021): 56–61. URL

Shneiderman, Ben. "Tutorial: Human-Centered AI: Reliable, Safe and Trustworthy." IUI '21 Companion: 26th International Conference on Intelligent User Interfaces - Companion, April 14, 2021, 7–8. URL

George Margetis, Stavroula Ntoa, Margherita Antona, and Constantine Stephanidis. "Human-Centered Design of Artificial Intelligence." In Handbook of Human Factors and Ergonomics, edited by Gavriel Salvendy and Waldemar Karwowski, 5th ed., 1085–1106. John Wiley & Sons, 2021. URL

Caitlin Thompson. "Who's Homeless Enough for Housing? In San Francisco, an Algorithm Decides." Coda, September 21, 2021. URL

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. "Algorithmic Decision-Making and the Control Problem." Minds and Machines 29, no. 4 (December 11, 2019): 555–78. URL

Fry, Hannah. Hello World: Being Human in the Age of Algorithms. New York: W.W. Norton & Company, 2018. URL

Sasha Costanza-Chock. Design Justice: Community-Led Practices to Build the Worlds We Need. Cambridge: The MIT Press, 2020. URL

David G. Robinson. Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. New York: Russell Sage Foundation, 2022. URL

Diane Hart, Gabi Diercks-O'Brien, and Adrian Powell. "Exploring Stakeholder Engagement in Impact Evaluation Planning in Educational Development Work." Evaluation 15, no. 3 (2009): 285–306. URL

Asit Bhattacharyya and Lorne Cummings. "Measuring Corporate Environmental Performance – Stakeholder Engagement Evaluation." Business Strategy and the Environment 24, no. 5 (2013): 309–25. URL

Hendricks, Sharief, Nailah Conrad, Tania S. Douglas, and Tinashe Mutsvangwa. "A Modified Stakeholder Participation Assessment Framework for Design Thinking in Health Innovation." Healthcare 6, no. 3 (September 2018): 191–96. URL

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. "Stakeholder Participation in AI: Beyond 'Add Diverse Stakeholders and Stir.' " arXiv preprint, submitted November 1, 2021. URL

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest." SSRN, July 8, 2021. URL

Alexandra Reeve Givens, and Meredith Ringel Morris. "Centering Disability Perspectives in Algorithmic Fairness, Accountability, & Transparency." FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27, 2020, 684-84. URL