NIST AI Risk Management Framework Playbook

– MANAGE

Abstract

The MANAGE function utilizes systematic documentation practices established in GOVERN, contextual information from MAP, and empirical information from MEASURE to treat identified risks and decrease the likelihood of system failures and negative impacts.

Risk treatment comprises plans to respond to, recover from, and communicate about incidents or events.

After completing the MANAGE function, plans for prioritizing risk and continuous monitoring and improvement will be in place.

Contents

MANAGE-1: AI risks based on assessments and other analytical output from the Map and Measure functions are prioritized, responded to, and managed.

MANAGE 1.1

A determination is as to whether the AI system achieves its intended purpose and stated objectives and whether its development or deployment should proceed.

About

AI systems may not necessarily be the right solution for a given business task or problem. A standard risk management practice is to formally weigh an AI system's negative risks against its benefits, and to determine if the AI system is an appropriate solution. Tradeoffs among trustworthiness characteristics —such as deciding to deploy a system based on system performance vs system transparency–may require regular assessment throughout the AI lifecycle.

Suggested Actions

- Consider trustworthiness characteristics when evaluating AI systems' negative risks and benefits.
- Utilize TEVV outputs from map and measure functions when considering risk treatment.
- Regularly track and monitor negative risks and benefits throughout the AI system lifecycle including in post-deployment monitoring.
- Regularly assess and document system performance relative to trustworthiness characteristics and tradeoffs between negative risks and opportunities.
- Evaluate tradeoffs in connection with real-world use cases and impacts and as enumerated in Map function outcomes.

Transparency and Documentation

Organizations can document the following:

- How do the technical specifications and requirements align with the AI system's goals and objectives?
- To what extent are the metrics consistent with system goals, objectives, and constraints, including ethical and compliance considerations?
- 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
- 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

Arvind Narayanan. How to recognize AI snake oil. Retrieved October 15, 2022. URL

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

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 29, 2021). URL

Fraser, Henry L and Bello y Villarino, Jose-Miguel, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation (September 30, 2021). LINK, URL

Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. (June 2022). URL

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. URL

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 695. URL

MANAGE 1.2

Treatment of documented AI risks is prioritized based on impact, likelihood, or available resources or methods.

About

Risk refers to the composite measure of an event's probability of occurring and the magnitude (or degree) of the consequences of the corresponding events. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or risks.

Organizational risk tolerances are often informed by several internal and external factors, including existing industry practices, organizational values, and legal or regulatory requirements. Since risk management resources are often limited, organizations usually assign them based on risk tolerance. AI risks that are deemed more serious receive more oversight attention and risk management resources.

Suggested Actions

- Assign risk management resources relative to established risk tolerance. AI systems with lower risk tolerances receive greater oversight, mitigation and management resources.
- Document AI risk tolerance determination practices and resource decisions.
- Regularly review risk tolerances and re-calibrate, as needed, in accordance with information from AI system monitoring and assessment .

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)?
- What assessments has the entity conducted on data security and privacy impacts associated with the AI system?
- Does your organization have an existing governance structure that can be leveraged to oversee the organization's use of AI?

AI Transparency Resources:

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

References

Arvind Narayanan. How to recognize AI snake oil. Retrieved October 15, 2022. URL

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

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 29, 2021). URL

Fraser, Henry L and Bello y Villarino, Jose-Miguel, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation (September 30, 2021). LINK, URL

Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. (June 2022). URL

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. URL

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 695. URL

MANAGE 1.3

Responses to the AI risks deemed high priority as identified by the Map function, are developed, planned, and documented. Risk response options can include mitigating, transferring, avoiding, or accepting.

About

Outcomes from GOVERN-1, MAP-5 and MEASURE-2, can be used to address and document identified risks based on established risk tolerances. Organizations can follow existing regulations and guidelines for risk criteria, tolerances and responses established by organizational, domain, discipline, sector, or professional requirements. In lieu of such guidance, organizations can develop risk response plans based on strategies such as accepted model risk management or enterprise risk management practices.

Suggested Actions

- Observe regulatory and established organizational, sector, discipline, or professional standards and requirements for applying risk tolerances within the organization.
- Document procedures for acting on AI system risks related to trustworthiness characteristics.
- Prioritize risks involving physical safety, legal liabilities, regulatory compliance, and negative impacts on individuals, groups, or society.
- Identify risk response plans and resources and organizational teams for carrying out response functions.
- Store risk management and system documentation in an organized, secure repository that is accessible by relevant AI Actors and appropriate personnel.

Transparency and Documentation

Organizations can document the following:

- Has the system been reviewed to ensure the AI system complies with relevant laws, regulations, standards, and guidance?
- To what extent has the entity defined and documented the regulatory environment—including minimum requirements in laws and regulations?
- 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)?

AI Transparency Resources:

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

References

Arvind Narayanan. How to recognize AI snake oil. Retrieved October 15, 2022. URL

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

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 29, 2021). URL

Fraser, Henry L and Bello y Villarino, Jose-Miguel, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation (September 30, 2021). LINK, URL

Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. (June 2022). URL

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. URL

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 695. URL

MANAGE 1.4

Negative residual risks (defined as the sum of all unmitigated risks) to both downstream acquirers of AI systems and end users are documented.

About

Organizations may choose to accept or transfer some of the documented risks from MAP and MANAGE 1.3 and 2.1. Such risks, known as residual risk, may affect downstream AI actors such as those engaged in system procurement or use. Transparent monitoring and managing residual risks enables cost benefit analysis and the examination of potential values of AI systems versus its potential negative impacts.

Suggested Actions

- Document residual risks within risk response plans, denoting risks that have been accepted, transferred, or subject to minimal mitigation.
- Establish procedures for disclosing residual risks to relevant downstream AI actors .
- Inform relevant downstream AI actors of requirements for safe operation, known limitations, and suggested warning labels as identified in MAP 3.4.

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?
- Who will be responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed?
- How will updates/revisions be documented and communicated? How often and by whom?
- How easily accessible and current is the information available to external stakeholders?

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

Arvind Narayanan. How to recognize AI snake oil. Retrieved October 15, 2022. URL

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

Emanuel Moss, Elizabeth Watkins, Ranjit Singh, Madeleine Clare Elish, Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. (June 29, 2021). URL

Fraser, Henry L and Bello y Villarino, Jose-Miguel, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation (September 30, 2021). LINK, URL

Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. (June 2022). URL

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. URL

Solon Barocas, Asia J. Biega, Benjamin Fish, et al. 2020. When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 695. URL

MANAGE-2: Strategies to maximize AI benefits and minimize negative impacts are planned, prepared, implemented, and documented, and informed by input from relevant AI actors.

MANAGE 2.1

Resources required to manage AI risks are taken into account, along with viable non-AI alternative systems, approaches, or methods – to reduce the magnitude or likelihood of potential impacts.

About

Organizational risk response may entail identifying and analyzing alternative approaches, methods, processes or systems, and balancing tradeoffs between trustworthiness characteristics and how they relate to organizational principles and societal values. Analysis of these tradeoffs is informed by consulting with interdisciplinary organizational teams, independent domain experts, and engaging with individuals or community groups. These processes require sufficient resource allocation.

Suggested Actions

- Plan and implement risk management practices in accordance with established organizational risk tolerances.
- Verify risk management teams are resourced to carry out functions, including
  - Establishing processes for considering methods that are not automated; semi-automated; or other procedural alternatives for AI functions.
  - Enhance AI system transparency mechanisms for AI teams.
  - Enable exploration of AI system limitations by AI teams.
  - Identify, assess, and catalog past failed designs and negative impacts or outcomes to avoid known failure modes.
- Identify resource allocation approaches for managing risks in systems:
  - deemed high-risk,
  - that self-update (adaptive, online, reinforcement self-supervised learning or similar),
  - trained without access to ground truth (unsupervised, semi-supervised, learning or similar),
  - with high uncertainty or where risk management is insufficient.
- Regularly seek and integrate external expertise and perspectives to supplement organizational diversity (e.g. demographic, disciplinary), equity, inclusion, and accessibility where internal capacity is lacking.
- Enable and encourage regular, open communication and feedback among AI actors and internal or external stakeholders related to system design or deployment decisions.
- Prepare and document plans for continuous monitoring and feedback mechanisms.

Transparency and Documentation

Organizations can document the following:

- Are mechanisms in place to evaluate whether internal teams are empowered and resourced to effectively carry out risk management functions?
- How will user and other forms of stakeholder engagement be integrated into risk management processes?

AI Transparency Resources:

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

References

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

David Wright. 2013. Making Privacy Impact Assessments More Effective. The Information Society, 29 (Oct 2013), 307-315. URL

Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. URL

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. URL

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al. 2021. Datasheets for Datasets. arXiv:1803.09010. URL

MANAGE 2.2

Mechanisms are in place and applied to sustain the value of deployed AI systems.

About

System performance and trustworthiness may evolve and shift over time, once an AI system is deployed and put into operation. This phenomenon, generally known as drift, can degrade the value of the AI system to the organization and increase the likelihood of negative impacts. Regular monitoring of AI systems' performance and trustworthiness enhances organizations' ability to detect and respond to drift, and thus sustain an AI system's value once deployed. Processes and mechanisms for regular monitoring address system functionality and behavior - as well as impacts and alignment with the values and norms within the specific context of use. For example, considerations regarding impacts on personal or public safety or privacy may include limiting high speeds when operating autonomous vehicles or restricting illicit content recommendations for minors.

Regular monitoring activities can enable organizations to systematically and proactively identify emergent risks and respond according to established protocols and metrics. Options for organizational responses include 1) avoiding the risk, 2)accepting the risk, 3) mitigating the risk, or 4) transferring the risk. Each of these actions require planning and resources. Organizations are encouraged to establish risk management protocols with consideration of the trustworthiness characteristics, the deployment context, and real world impacts.

Suggested Actions

- Establish risk controls considering trustworthiness characteristics, including:
  - Data management, quality, and privacy (e.g. minimization, rectification or deletion requests) controls as part of organizational data governance policies.
  - Machine learning and end-point security countermeasures (e.g., robust models, differential privacy, authentication, throttling).
  - Business rules that augment, limit or restrict AI system outputs within certain contexts
  - Utilizing domain expertise related to deployment context for continuous improvement and TEVV across the AI lifecycle.
  - Development and regular tracking of human-AI teaming configurations.
  - Model assessment and test, evaluation, validation and verification (TEVV) protocols.
  - Use of standardized documentation and transparency mechanisms.
  - Software quality assurance practices across AI lifecycle.
  - Mechanisms to explore system limitations and avoid past failed designs or deployments.
- Establish mechanisms to capture feedback from system end users and potentially impacted groups.
- Review insurance policies, warranties, or contracts for legal or oversight requirements for risk transfer procedures.
- Document risk tolerance decisions and risk acceptance procedures.

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?
- Could the AI system expose people to harm or negative impacts? What was done to mitigate or reduce the potential for harm?
- 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 or business 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

Safety, Validity and Reliability Risk Management Approaches and Resources

AI Incident Database. 2022. AI Incident Database. URL

AIAAIC Repository. 2022. AI, algorithmic and automation incidents collected, dissected, examined, and divulged. URL

Alexander D'Amour, Katherine Heller, Dan Moldovan, et al. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. arXiv:2011.03395. URL

Andrew L. Beam, Arjun K. Manrai, Marzyeh Ghassemi. 2020. Challenges to the Reproducibility of Machine Learning Models in Health Care. Jama 323, 4 (January 6, 2020), 305-306. URL

Anthony M. Barrett, Dan Hendrycks, Jessica Newman et al. 2022. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. arXiv:2206.08966. URL

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

Jessie J. Smith, Saleema Amershi, Solon Barocas, et al. 2022. REAL ML: Recognizing, Exploring, and Articulating Limitations of Machine Learning Research. arXiv:2205.08363. URL

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, et al. 2020. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program) arXiv:2003.12206. URL

Kirstie Whitaker. 2017. Showing your working: a how to guide to reproducible research. (August 2017). LINK, URL

Netflix. Chaos Monkey. URL

Peter Henderson, Riashat Islam, Philip Bachman, et al. 2018. Deep reinforcement learning that matters. Proceedings of the AAAI Conference on Artificial Intelligence. 32, 1 (Apr. 2018). URL

Suchi Saria, Adarsh Subbaswamy. 2019. Tutorial: Safe and Reliable Machine Learning. arXiv:1904.07204. URL

Kang, Daniel, 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

Managing Risk Bias

National Institute of Standards and Technology (NIST), Reva Schwartz, Apostol Vassilev, et al. 2022. NIST Special Publication 1270 Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. URL

Bias Testing and Remediation Approaches

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, et al. 2018. A Reductions Approach to Fair Classification. arXiv:1803.02453. URL

Brian Hu Zhang, Blake Lemoine, Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. arXiv:1801.07593. URL

Drago Plečko, Nicolas Bennett, Nicolai Meinshausen. 2021. Fairadapt: Causal Reasoning for Fair Data Pre-processing. arXiv:2110.10200. URL

Faisal Kamiran, Toon Calders. 2012. Data Preprocessing Techniques for Classification without Discrimination. Knowledge and Information Systems 33 (2012), 1–33. URL

Faisal Kamiran; Asim Karim; Xiangliang Zhang. 2012. Decision Theory for Discrimination-Aware Classification. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, December 10-13, 2012, Brussels, Belgium. IEEE, 924-929. URL

Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, et al. 2017. Optimized Data Pre-Processing for Discrimination Prevention. arXiv:1704.03354. URL

Geoff Pleiss, Manish Raghavan, Felix Wu, et al. 2017. On Fairness and Calibration. arXiv:1709.02012. URL

L. Elisa Celis, Lingxiao Huang, Vijay Keswani, et al. 2020. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. arXiv:1806.06055. URL

Michael Feldman, Sorelle Friedler, John Moeller, et al. 2014. Certifying and Removing Disparate Impact. arXiv:1412.3756. URL

Michael Kearns, Seth Neel, Aaron Roth, et al. 2017. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. arXiv:1711.05144. URL

Michael Kearns, Seth Neel, Aaron Roth, et al. 2018. An Empirical Study of Rich Subgroup Fairness for Machine Learning. arXiv:1808.08166. URL

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), 2016, Barcelona, Spain. URL

Rich Zemel, Yu Wu, Kevin Swersky, et al. 2013. Learning Fair Representations. In Proceedings of the 30th International Conference on Machine Learning 2013, PMLR 28, 3, 325-333. URL

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh & Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In Peter A. Flach, Tijl De Bie, Nello Cristianini (eds) Machine Learning and Knowledge Discovery in Databases. European Conference ECML PKDD 2012, Proceedings Part II, September 24-28, 2012, Bristol, UK. Lecture Notes in Computer Science 7524. Springer, Berlin, Heidelberg. URL

Security and Resilience Resources

FTC Start With Security Guidelines. 2015. URL

Gary McGraw et al. 2022. BIML Interactive Machine Learning Risk Framework. Berryville Institute for Machine Learning. URL

Ilia Shumailov, Yiren Zhao, Daniel Bates, et al. 2021. Sponge Examples: Energy-Latency Attacks on Neural Networks. arXiv:2006.03463. URL

Marco Barreno, Blaine Nelson, Anthony D. Joseph, et al. 2010. The Security of Machine Learning. Machine Learning 81 (2010), 121-148. URL

Matt Fredrikson, Somesh Jha, Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15), October 2015. Association for Computing Machinery, New York, NY, USA, 1322–1333. URL

National Institute for Standards and Technology (NIST). 2022. Cybersecurity Framework. URL

Nicolas Papernot. 2018. A Marauder's Map of Security and Privacy in Machine Learning. arXiv:1811.01134. URL

Reza Shokri, Marco Stronati, Congzheng Song, et al. 2017. Membership Inference Attacks against Machine Learning Models. arXiv:1610.05820. URL

Adversarial Threat Matrix (MITRE). 2021. URL

Interpretability and Explainability Approaches

Chaofan Chen, Oscar Li, Chaofan Tao, et al. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. arXiv:1806.10574. URL

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. arXiv:1811.10154. URL

Daniel W. Apley, Jingyu Zhu. 2019. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. arXiv:1612.08468. URL

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

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, et al. 2021. Manipulating and Measuring Model Interpretability. arXiv:1802.07810. URL

Hongyu Yang, Cynthia Rudin, Margo Seltzer. 2017. Scalable Bayesian Rule Lists. arXiv:1602.08610. URL

P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, et al. 2021. Four Principles of Explainable Artificial Intelligence. National Institute of Standards and Technology (NIST) IR 8312. National Institute of Standards and Technology, Gaithersburg, MD. URL

Scott Lundberg, Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874. URL

Susanne Gaube, Harini Suresh, Martina Raue, et al. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. npj Digital Medicine 4, Article 31 (2021). URL

Yin Lou, Rich Caruana, Johannes Gehrke, et al. 2013. Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13), August 2013. Association for Computing Machinery, New York, NY, USA, 623–631. URL

Privacy Resources

National Institute for Standards and Technology (NIST). 2022. Privacy Framework. URL

Data Governance

Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, Tomasz Janowski, Data governance: Organizing data for trustworthy Artificial Intelligence, Government Information Quarterly, Volume 37, Issue 3, 2020, 101493, ISSN 0740-624X. URL

Software Resources

- PiML (explainable models, performance assessment)
- Interpret (explainable models)
- Iml (explainable models)
- Drifter library (performance assessment)
- Manifold library (performance assessment)
- SALib library (performance assessment)
- What-If Tool (performance assessment)
- MLextend (performance assessment)
- AI Fairness 360: - Python (bias testing and mitigation) - R (bias testing and mitigation)

- Adversarial-robustness-toolbox (ML security)
- Robustness (ML security)
- tensorflow/privacy (ML security)
- NIST De-identification Tools (Privacy and ML security)
- Dvc (MLops, deployment)
- Gigantum (MLops, deployment)
- Mlflow (MLops, deployment)
- Mlmd (MLops, deployment)
- Modeldb (MLops, deployment)

MANAGE 2.3

Procedures are followed to respond to and recover from a previously unknown risk when it is identified.

About

AI systems – like any technology – can demonstrate non-functionality or failure or unexpected and unusual behavior. They also can be subject to attacks, incidents, or other misuse of abuses – which their sources are not always known apriori. Organizations can establish, document, communicate and maintain treatment procedures to recognize and counter, mitigate and manage risks that were not previously identified.

Suggested Actions

- Protocols, resources, and metrics are in place for continual monitoring of AI systems' performance, trustworthiness, and alignment with contextual norms and values
- Establish and regularly review treatment and response plans for incidents, negative impacts, or outcomes.
- Establish and maintain procedures to regularly monitor system components for drift, decontextualization, or other AI system behavior factors,
- Establish and maintain procedures for capturing feedback about negative impacts.
- Verify contingency processes to handle any negative impacts associated with mission-critical AI systems, and to deactivate systems.
- Enable preventive and post-hoc exploration of AI system limitations by relevant AI actor groups.
- Decommission systems that exceed risk tolerances.

Transparency and Documentation

Organizations can document the following:

- 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? (Including responsibilities to decommission the AI system.)
- What processes exist for data generation, acquisition/collection, ingestion, staging/storage, transformations, security, maintenance, and dissemination?
- How will the appropriate performance metrics, such as accuracy, of the AI be monitored after the AI is deployed?

AI Transparency Resources:

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

References

AI Incident Database. 2022. AI Incident Database. URL

AIAAIC Repository. 2022. AI, algorithmic and automation incidents collected, dissected, examined, and divulged. URL

Andrew Burt and Patrick Hall. 2018. What to Do When AI Fails. O'Reilly Media, Inc. (May 18, 2020). Retrieved October 17, 2022. URL

National Institute for Standards and Technology (NIST). 2022. Cybersecurity Framework. URL

SANS Institute. 2022. Security Consensus Operational Readiness Evaluation (SCORE) Security Checklist [or Advanced Persistent Threat (APT) Handling Checklist]. URL

Suchi Saria, Adarsh Subbaswamy. 2019. Tutorial: Safe and Reliable Machine Learning. arXiv:1904.07204. URL

MANAGE 2.4

Mechanisms are in place and applied, responsibilities are assigned and understood to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.

About

Superseding (bypassing), disengaging, or deactivating/decommissioning a model, AI system component(s), or the entire AI system may be necessary when: - a system reaches the end of its lifetime - detected or identified risks exceed tolerance thresholds - adequate system mitigation actions are beyond the organization's capacity - feasible system mitigation actions do not meet regulatory, legal, norms or standards. - impending risk is detected during continual monitoring, for which feasible mitigation cannot be identified or implemented in a timely fashion.

Safely removing AI systems from operation, either temporarily or permanently, under these scenarios requires standard protocols that minimize operational disruption and downstream negative impacts. Protocols can involve redundant or backup systems that are developed in alignment with established system governance policies (see GOVERN 1.7), regulatory compliance, legal frameworks, business requirements and norms and l standards within the application context of use. Decision thresholds and metrics for actions to bypass or deactivate system components are part of continual monitoring procedures. Incidents that result in a bypass/deactivate decision require documentation and review to understand root causes, impacts, and potential opportunities for mitigation and redeployment. Organizations are encouraged to develop risk and change management protocols that consider and anticipate upstream and downstream consequences of both temporary and/or permanent decommissioning, and provide contingency options.

Suggested Actions

- Regularly review established procedures for AI system bypass actions, including plans for redundant or backup systems to ensure continuity of operational and/or business functionality.
- Regularly review Identify system incident thresholds for activating bypass or deactivation responses.
- Apply change management processes to understand the upstream and downstream consequences of bypassing or deactivating an AI system or AI system components.
- Apply protocols, resources and metrics for decisions to supersede, bypass or deactivate AI systems or AI system components.
- Preserve materials for forensic, regulatory, and legal review.
- Conduct internal root cause analysis and process reviews of bypass or deactivation events.
- Decommission and preserve system components that cannot be updated to meet criteria for redeployment.
- Establish criteria for redeploying updated system components, in consideration of 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?
- 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)?
- 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 does the entity have established procedures for retiring the AI system, if it is no longer needed?
- How did the entity use assessments and/or evaluations to determine if the system can be scaled up, continue, or be decommissioned?

AI Transparency Resources:

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

References

Decommissioning Template. Application Lifecycle And Supporting Docs. Cloud and Infrastructure Community of Practice. URL

Develop a Decommission Plan. M3 Playbook. Office of Shared Services and Solutions and Performance Improvement. General Services Administration. URL

MANAGE-3: AI risks and benefits from third-party entities are managed.

MANAGE 3.1

AI risks and benefits from third-party resources are regularly monitored, and risk controls are applied and documented.

About

AI systems may depend on external resources and associated processes, including third-party data, software or hardware systems. Third parties' supplying organizations with components and services, including tools, software, and expertise for AI system design, development, deployment or use can improve efficiency and scalability. It can also increase complexity and opacity, and, in-turn, risk. Documenting third-party technologies, personnel, and resources that were employed can help manage risks. Focusing first and foremost on risks involving physical safety, legal liabilities, regulatory compliance, and negative impacts on individuals, groups, or society is recommended.

Suggested Actions

- Have legal requirements been addressed?
- Apply organizational risk tolerance to third-party AI systems.
- Apply and document organizational risk management plans and practices to third-party AI technology, personnel, or other resources.
- Identify and maintain documentation for third-party AI systems and components.
- Establish testing, evaluation, validation and verification processes for third-party AI systems which address the needs for transparency without exposing proprietary algorithms .
- Establish processes to identify beneficial use and risk indicators in third-party systems or components, such as inconsistent software release schedule, sparse documentation, and incomplete software change management (e.g., lack of forward or backward compatibility).
- Organizations can establish processes for third parties to report known and potential vulnerabilities, risks or biases in supplied resources.
- Verify contingency processes for handling negative impacts associated with mission-critical third-party AI systems.
- Monitor third-party AI systems for potential negative impacts and risks associated with trustworthiness characteristics.
- Decommission third-party systems that exceed risk tolerances.

Transparency and Documentation

Organizations can document the following:

- If a third party created the AI system or some of its components, how will you ensure a level of explainability or interpretability? Is there documentation?
- If your organization obtained datasets from a third party, did your organization assess and manage the risks of using such datasets?
- 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?
- Have legal requirements been addressed?

AI Transparency Resources:

- 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
- Datasheets for Datasets. URL

References

Office of the Comptroller of the Currency. 2021. Proposed Interagency Guidance on Third-Party Relation-

ships: Risk Management. July 12, 2021. URL

MANAGE 3.2

Pre-trained models which are used for development are monitored as part of AI system regular monitoring and maintenance.

About

A common approach in AI development is transfer learning, whereby an existing pre-trained model is adapted for use in a different, but related application. AI actors in development tasks often use pre-trained models from third-party entities for tasks such as image classification, language prediction, and entity recognition, because the resources to build such models may not be readily available to most organizations. Pre-trained models are typically trained to address various classification or prediction problems, using exceedingly large datasets and computationally intensive resources. The use of pre-trained models can make it difficult to anticipate negative system outcomes or impacts. Lack of documentation or transparency tools increases the difficulty and general complexity when deploying pre-trained models and hinders root cause analyses.

Suggested Actions

- Identify pre-trained models within AI system inventory for risk tracking.
- Establish processes to independently and continually monitor performance and trustworthiness of pre-trained models, and as part of third-party risk tracking.
- Monitor performance and trustworthiness of AI system components connected to pre-trained models, and as part of third-party risk tracking.
- Identify, document and remediate risks arising from AI system components and pre-trained models per organizational risk management procedures, and as part of third-party risk tracking.
- Decommission AI system components and pre-trained models which exceed risk tolerances, and as part of third-party risk tracking.

Transparency and Documentation

Organizations can document the following:

- How has the entity documented the AI system's data provenance, including sources, origins, transformations, augmentations, labels, dependencies, constraints, and metadata?
- Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?
- How does the entity ensure that the data collected are adequate, relevant, and not excessive in relation to the intended purpose?
- If the dataset becomes obsolete how will this be communicated?

AI Transparency Resources:

- 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
- Datasheets for Datasets. URL

References

Larysa Visengeriyeva et al. "Awesome MLOps," GitHub. Accessed January 9, 2023. URL

MANAGE-4: Risk treatments including response and recovery, and communication plans to the identified and measured AI risks are documented and monitored regularly.

MANAGE 4.1

Post-deployment AI system monitoring plans are implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management.

About

AI system performance and trustworthiness can change due to a variety of factors. Regular AI system monitoring can help deployers identify performance degradations, adversarial attacks, unexpected and unusual behavior, and impacts. Including pre- and post-deployment external feedback about AI system performance can enhance organizational awareness about positive and negative impacts, and reduce the time to respond to risks and harms.

Suggested Actions

- Establish and maintain procedures to monitor AI system performance for risks and negative and positive impacts associated with trustworthiness characteristics.
- Perform post-deployment TEVV tasks to evaluate AI system validity and reliability, bias and fairness, privacy, and security and resilience.
- Evaluate AI system trustworthiness in conditions similar to deployment context of use, and prior to deployment.
- Establish and implement red-teaming exercises at a prescribed cadence, and evaluate their efficacy.
- Establish procedures for tracking dataset modifications such as data deletion or rectification requests.
- Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders to capture information about system performance, trustworthiness and impact.
- Share information about errors and attack patterns with incident databases, other organizations with similar systems, and system users and stakeholders.
- Respond to and document detected or reported negative impacts or issues in AI system performance and trustworthiness.
- Decommission systems that exceed establish risk tolerances.

Transparency and Documentation

Organizations can document the following:

- To what extent has the entity documented the post-deployment AI system's testing methodology, metrics, and performance outcomes?
- How easily accessible and current is the information available to external stakeholders?

AI Transparency Resources:

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

References

Navdeep Gill, Patrick Hall, Kim Montgomery, and Nicholas Schmidt. "A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing." Information 11, no. 3 (2020): 137. URL

MANAGE 4.2

Measurable activities for continual improvements are integrated into AI system updates and include regular engagement with interested parties, including relevant AI actors.

About

Regular monitoring processes enable system updates to enhance performance and functionality in accordance with regulatory and legal frameworks, and organizational and contextual values and norms. These processes also facilitate analyses of root causes, system degradation, drift, and failure, and incident response and documentation.

AI actors across the lifecycle have many opportunities to capture and incorporate external feedback about system performance, limitations, and impacts, and implement continuous improvements. Improvements may not always be to model pipeline or system processes, and may instead be based on metrics beyond accuracy or other quality performance measures. In these cases, improvements may entail adaptations to business or organizational procedures or practices. Organizations are encouraged to develop improvements that will maintain traceability and transparency for developers, end users, auditors, and relevant AI actors.

Suggested Actions

- Integrate trustworthiness characteristics into protocols and metrics used for continual improvement.
- Establish processes for evaluating and integrating feedback into AI system improvements.
- Assess and evaluate alignment of proposed improvements with relevant regulatory and legal frameworks
- Assess and evaluate alignment of proposed improvements connected to the values and norms within the context of use.
- Document the basis for decisions made relative to tradeoffs between trustworthy characteristics, system risks, and system opportunities

Transparency and Documentation

Organizations can document the following:

- How will user and other forms of stakeholder engagement be integrated into the model development process and regular 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 has the entity defined and documented the regulatory environment—including minimum requirements in laws and regulations?

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
** Yen, Po-Yin, et al. "Development and Evaluation of Socio-Technical Metrics to Inform HIT Adaptation." URL

Carayon, Pascale, and Megan E. Salwei. "Moving toward a sociotechnical systems approach to continuous health information technology design: the path forward for improving electronic health record usability and reducing clinician burnout." Journal of the American Medical Informatics Association 28.5 (2021): 1026-1028. URL

Mishra, Deepa, et al. "Organizational capabilities that enable big data and predictive analytics diffusion and organizational performance: A resource-based perspective." Management Decision (2018).

MANAGE 4.3

Incidents and errors are communicated to relevant AI actors including affected communities. Processes for tracking, responding to, and recovering from incidents and errors are followed and documented.

About

Regularly documenting an accurate and transparent account of identified and reported errors can enhance AI risk management activities., Examples include: - how errors were identified, - incidents related to the error, - whether the error has been repaired, and - how repairs can be distributed to all impacted stakeholders and users.

Suggested Actions

- Establish procedures to regularly share information about errors, incidents and negative impacts with relevant stakeholders, operators, practitioners and users, and impacted parties.
- Maintain a database of reported errors, incidents and negative impacts including date reported, number of reports, assessment of impact and severity, and responses.
- Maintain a database of system changes, reason for change, and details of how the change was made, tested and deployed.
- Maintain version history information and metadata to enable continuous improvement processes.
- Verify that relevant AI actors responsible for identifying complex or emergent risks are properly resourced and empowered.

Transparency and Documentation

Organizations can document the following:

- What corrective actions has the entity taken to enhance the quality, accuracy, reliability, and representativeness of the data?
- To what extent does the entity communicate its AI strategic goals and objectives to the community of stakeholders? 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

References

Wei, M., & Zhou, Z. (2022). AI Ethics Issues in Real World: Evidence from AI Incident Database. ArXiv, abs/2206.07635. URL

McGregor, Sean. "Preventing repeated real world AI failures by cataloging incidents: The AI incident database." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 17. 2021. URL

Macrae, Carl. "Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk." Risk analysis 42.9 (2022): 1999-2025. URL