NIST AI Risk Management Framework Playbook
— MAP

Abstract

The Map function establishes the context and frames risks related to an AI system. Information gathered in this function informs decisions about model management, including an initial decision about appropriateness or the need for an AI solution.

Contents

MAP-1: Context is estabished and understood. 2
MAP 1.1 2
MAP 1.2 4
MAP 1.3 5
MAP 1.4 6
MAP 1.5 6
MAP 1.6 7
MAP 1.7 8
MAP-2: Classification of the AI system is performed. 11
MAP 2.1 11
MAP 2.2 11
MAP 2.3 13
MAP-3: AI capabilities, targeted usage, goals, and expected benefits and costs compared with the status quo are understood. 16
MAP 3.1 16
MAP 3.2 16
MAP 3.3 17
MAP-4: Risks and benefits are mapped for third-party software and data. 19
MAP 4.1 19
MAP 4.2 20
MAP-5: Impacts to individuals, groups, communities, organizational, or society are assessed. . . . 21
MAP 5.1 21
MAP 5.2 22
MAP 5.3 23

MAP-1: Context is estabished and understood.

MAP 1.1

Intended purpose, prospective settings in which the AI system will be deployed, the specific set or types of users along with their expectations, and impacts of system use are understood and documented. Assumptions and related limitations about AI system purpose and use are enumerated, documented and tied to TEVV considerations and system metrics.

About

It is not necessarily possible to have advanced knowledge about all potential settings in which a system will be deployed. To help delineate the bounds of acceptable deployment, context mapping may include examination of the following: * intended, prospective,and actual deployment setting. * specific set or types of users. * operator or subject expectations. * concept of operations. * intended purpose and impact of system use. * requirements for system deployment and operation. * potential negative impacts to individuals, groups, communities, organizations, and society – or context-specific impacts such as legal requirements or impacts to the environment. * unintended, downstream, or other unknown contextual factors.

Actions

- Pursue AI system design purposefully, after non-AI solutions are considered.
- Define and document the task, purpose, minimum functionality, and benefits of the AI system to inform considerations about whether the project is worth pursuing.
- Maintain awareness of industry, technical, and applicable legal standards.
- Collaboratively consider intended AI system design tasks along with unanticipated purposes.
- Determine the user and organizational requirements, including business and technical requirements.
- Determine and delineate the expected and acceptable AI system context of use, including:
  - operational environment
  - impacts to individuals, groups, communities, organizations, and society
  - user characteristics and tasks
  - social environment.
- Track and document existing AI systems held by the organization, and those maintained or supported by third-party entities.
- Gain and maintain awareness about evaluating scientific claims related to AI system performance and benefits before launching into system design.
- Identify human-AI interaction and/or roles, such as whether the application will support or replace human decision making.
- Plan for risks related to human-AI configurations, and document requirements, roles, and responsibilities for human oversight of deployed systems.

Transparency and Documentation

Organizations can document the following:

- Which AI actors are responsible for the decisions of the AI and is this person aware of the intended uses and limitations of the analytic?
- Which AI actors are responsible for maintaining, re-verifying, monitoring, and updating this AI once deployed?
- Who is the person(s) accountable for the ethical considerations across the AI lifecycle?

AI Transparency Resources:

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- “Stakeholders in Explainable AI,” Sep. 2018, URL.
- “Microsoft Responsible AI Standard, v2”, URL.

References

Socio-technical systems

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, et al. 2019. Fairness and Abstraction in Sociotechnical Systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT'19). Association for Computing Machinery, New York, NY, USA, 59–68. URL

Problem formulation

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices in artificial intelligence. Artificial Intelligence 300 (14 July 2021), 103555, ISSN 0004-3702. URL

Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT'19). Association for Computing Machinery, New York, NY, USA, 39–48. URL

Context mapping

Emilio Gómez-González and Emilia Gómez. 2020. Artificial intelligence in medicine and healthcare. Joint Research Centre (European Commission). URL

Sarah Spiekermann and Till Winkler. 2020. Value-based Engineering for Ethics by Design. arXiv:2004.13676. URL

Social Impact Lab. 2017. Framework for Context Analysis of Technologies in Social Change Projects (Draft v2.0). URL

Solon Barocas, Asia J. Biega, Margarita Boyarskaya, et al. 2021. Responsible computing during COVID-19 and beyond. Commun. ACM 64, 7 (July 2021), 30–32. URL

Identification of harms

Harini Suresh and John V. Guttag. 2020. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. arXiv:1901.10002. URL

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. arXiv:2011.13416. URL

Microsoft. Foundations of assessing harm. 2022. URL

Understanding and documenting limitations in ML

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

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

Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, et al. 2019. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv:1808.07261. URL

Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. URL

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

Bender, E. M., Friedman, B. & McMillan-Major, A., (2022). A Guide for Writing Data Statements for Natural Language Processing. University of Washington. Accessed July 14, 2022. URL

Meta AI. System Cards, a new resource for understanding how AI systems work, 2021. URL

When not to deploy

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

Statistical balance

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 6464 (25 Oct. 2019), 447-453. URL

Assessment of science in AI

Arvind Narayanan. How to recognize AI snake oil. URL

Emily M. Bender. 2022. On NYT Magazine on AI: Resist the Urge to be Impressed. (April 17, 2022). URL

MAP 1.2

Inter-disciplinary AI actors, competencies, skills and capacities for establishing context reflect demographic diversity and broad domain and user experience expertise, and their participation is documented. Opportunities for interdisciplinary collaboration are prioritized.

About

Successfully mapping context requires a team of AI actors with a diversity of experience, expertise, abilities and backgrounds, and with the resources and independence to engage in critical inquiry.

Having a diverse team contributes to more open sharing of ideas and assumptions about the purpose and function of the technology being designed and developed – making these implicit aspects more explicit. The benefit of a diverse staff in managing AI risks is not the beliefs or presumed beliefs of individual workers, but the behavior that results from a collective perspective. An environment which fosters critical inquiry creates opportunities to surface problems and identify existing and emergent risks.

Actions

- Establish interdisciplinary teams to reflect a wide range of skills, competencies, and capacity for AI efforts. Verify that team membership includes both demographic diversity, broad domain expertise, and lived experiences. Document team composition.
- Create and empower interdisciplinary expert teams to capture, learn, and engage the interdependencies of deployed AI systems and related terminologies and concepts from disciplines outside of AI practice such as law, sociology, psychology, anthropology, public policy, systems design, and engineering.

Transparency and Documentation

Organizations can document the following:

- To what extent do the teams responsible for developing and maintaining the AI system reflect diverse opinions, backgrounds, experiences, and perspectives?
- Did the entity document the demographics of those involved in the design and development of the AI system to capture and communicate potential biases inherent to the development process, according to forum participants?
- What specific perspectives did stakeholders share, and how were they integrated across the design, development, deployment, assessment, and monitoring of the AI system?
- To what extent has the entity addressed stakeholder perspectives on the potential negative impacts of the AI system on end users and impacted populations?

- 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?
- Did your organization address usability problems and test whether user interfaces served their intended purposes? Consulting the community or end users at the earliest stages of development to ensure there is transparency on the technology used and how it is deployed.

AI Transparency Resources:

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- WEF Model AI Governance Framework Assessment 2020, URL.
- WEF Companion to the Model AI Governance Framework- 2020, URL
- AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019, URL.

References

Sina Fazelpour and Maria De-Arteaga. 2022. Diversity in sociotechnical machine learning systems. Big Data & Society 9, 1 (Jan. 2022). URL

Microsoft Community Jury , Azure Application Architecture Guide. URL

MAP 1.3

The business value or context of business use has been clearly defined or – in the case of assessing existing AI systems – re-evaluated.

About

AI systems should present a business benefit beyond the status quo when considering inherent risks and implicit or explicit costs. Defining and documenting the specific business purpose of an AI system in a broader context of societal values helps teams to evaluate risks and increases the clarity of “go/no-go” decisions about whether to deploy.

Actions

- Build transparent practices into AI system development processes.
- Review the documented system purpose from a socio-technical perspective and in consideration of societal values.
- Determine possible misalignment between societal values and stated organizational principles and code of ethics.
- Flag latent incentives that may contribute to negative impacts.
- Balance AI system purpose with potential risks, societal values, and stated organizational principles.

Transparency and Documentation

Organizations can document the following:

- How does the AI system help the entity meet its goals and objectives?
- How do the technical specifications and requirements align with the AI system’s goals and objectives?
- To what extent is the output appropriate for the operational context?

AI Transparency Resources:

- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI – 2019, LINK, URL.
- Including Insights from the Comptroller General’s Forum on the Oversight of Artificial Intelligence An Accountability Framework for Federal Agencies and Other Entities, 2021, URL, PDF.

References

Abeba Birhane, Pratyusha Kalluri, Dallas Card, et al. 2022. The Values Encoded in Machine Learning Research. arXiv:2106.15590. URL

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

MAP 1.4

The organization's mission and relevant goals for the AI technology are understood.

About

Socio-technical AI risks emerge from the interplay between technical development decisions and how a system is used, who operates it, and the social context into which it is deployed. Addressing these risks is complex and requires a commitment to understanding how contextual factors may interact with AI lifecycle actions. One such contextual factor is how organizational mission and identified system purpose create incentives within AI system design, development, and deployment tasks that may result in positive and negative impacts. By establishing comprehensive and explicit enumeration of AI system purpose and expectations, organizations can identify and manage these types of risks and benefits.

Actions

- Reconcile documented concerns about system context of use or purpose against the organization's stated values, mission statements, social responsibility commitments, and AI principles.
- Reconsider the design, implementation strategy, or deployment of AI systems with potential impacts that do not reflect institutional values.

Transparency and Documentation

Organizations can document the following:

- What goals and objectives does the entity expect to achieve by designing, developing, and/or deploying the AI system?
- To what extent are the model outputs consistent with the entity's values and principles to foster public trust and equity?
- To what extent are the metrics consistent with system goals, objectives, and constraints, including ethical and compliance considerations?

AI Transparency Resources:

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020, URL.
- WEF Model AI Governance Framework Assessment 2020, URL.

References

Algorithm Watch. AI Ethics Guidelines Global Inventory. URL

Emanuel Moss and Jacob Metcalf. 2020. Ethics Owners: A New Model of Organizational Responsibility in Data-Driven Technology Companies. Data & Society Research Institute. URL

Future of Life Institute. Asilomar AI Principles. URL

Leonard Haas, Sebastian Gießler, and Veronika Thiel. 2020. In the realm of paper tigers – exploring the failings of AI ethics guidelines. (April 28, 2020). URL

MAP 1.5

Organizational risk tolerances are determined.

About

Risk tolerance reflects the level and type of risk the organization will accept while conducting its mission and carrying out its strategy.

Deployment should not be pre-determined. Rather, it should result from a clearly defined process based on organizational risk tolerances.

Go/no-go decisions should be incorporated throughout the AI system's lifecycle. For systems deemed "higher risk," such decisions should include approval from relevant technical or risk-focused executives.

Go/no-go decisions related to AI system risks should take stakeholder feedback into account, but remain independent from stakeholders' vested financial or reputational interests

Actions

- Establish risk tolerance levels for AI systems and allocate the appropriate oversight resources to each level.
- Identify maximum allowable risk thresholds above which the system will not be deployed, or will need to be prematurely decommissioned, within the contextual or application setting.
- Attempts to use a system for "off-label" purposes should be approached with caution, especially in settings that organizations have deemed as high-risk. Document decisions, risk-related trade-offs, and system limitations.

Transparency and Documentation

Organizations can document the following:

- What justifications, if any, has the entity provided for the assumptions, boundaries, and limitations of the AI system?
- 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?

AI Transparency Resources:

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- WEF Model AI Governance Framework Assessment 2020, URL.
- WEF Companion to the Model AI Governance Framework- 2020, URL.

References

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

The Office of the Comptroller of the Currency. Enterprise Risk Appetite Statement. (Nov. 20, 2019). URL

MAP 1.6

Practices and personnel for design activities enable regular engagement with stakeholders, and integrate actionable user and community feedback about unanticipated negative impacts

About

Risk management should include processes for regular and meaningful communication with stakeholder groups. Stakeholders can provide valuable input related to system gaps and limitations. Organizations may differ in the types and number of stakeholders with which they engage.

Participatory approaches such as human-centered design (HCD) and value-sensitive design (VSD) can help AI teams to engage broadly with stakeholder communities. This type of engagement can enable AI teams to learn about how a given technology may cause impacts, both positive and negative, that were not originally considered or intended.

Actions

- Maintain awareness and documentation of the individuals, groups, or communities who make up the system's internal and external stakeholders.
- Verify that appropriate skills and practices are available in-house for carrying out stakeholder engagement activities such as eliciting, capturing, and synthesizing stakeholder feedback, and translating it for AI design and development functions.
- Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders related to system design or deployment decisions.
- Define which AI actors, beyond AI design and development teams, will review system design, implementation, and operation tasks. Define which AI actors will administer and implement test, evaluation, verification, and validation (TEVV) tasks across the AI lifecycle.

Transparency and Documentation

Organizations can document the following:

- 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?
- To what extent is this information sufficient and appropriate to promote transparency? Promote transparency by enabling external stakeholders to access information on the design, operation, and limitations of the AI system.
- To what extent has relevant information been disclosed regarding the use of AI systems, such as (a) what the system is for, (b) what it is not for, (c) how it was designed, and (d) what its limitations are? (Documentation and external communication can offer a way for entities to provide transparency.)
- What metrics has the entity developed to measure performance of the AI system?
- What justifications, if any, has the entity provided for the assumptions, boundaries, and limitations of the AI system?

AI Transparency Resources:

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- Stakeholders in Explainable AI, Sep. 2018, URL.
- High-Level Expert Group on Artificial Intelligence set up by the European Commission, Ethics Guidelines for Trustworthy AI, URL, PDF.

References

Vincent T. Covello. 2021. Stakeholder Engagement and Empowerment. In Communicating in Risk, Crisis, and High Stress Situations (Vincent T. Covello, ed.), 87-109. URL

Yilin Huang, Giacomo Poderi, Sanja Šćepanović, et al. 2019. Embedding Internet-of-Things in Large-Scale Socio-technical Systems: A Community-Oriented Design in Future Smart Grids. In The Internet of Things for Smart Urban Ecosystems (2019), 125-150. Springer, Cham. URL

Eloise Taysom and Nathan Crilly. 2017. Resilience in Sociotechnical Systems: The Perspectives of Multiple Stakeholders. She Ji: The Journal of Design, Economics, and Innovation, 3, 3 (2017), 165-182, ISSN 2405-8726. URL

MAP 1.7

System requirements (e.g., "the system shall respect the privacy of its users") are elicited and understood from stakeholders. Design decisions take socio-technical implications into account to address AI risk.

About

AI system development requirements may outpace documentation processes for traditional software. When written requirements are unavailable or incomplete, AI actors may inadvertently overlook business and stakeholder needs, or over-rely on implicit human biases such as confirmation bias and groupthink. To mitigate the influence of these implicit factors, AI actors can seek input from, and develop transparent and actionable recourse mechanisms for, end-users and operators. Engaging external stakeholders in this process integrates broader perspectives on socio-technical risk factors. Incorporating trustworthy characteristics early in the design phase should be a priority – instead of forcing a solution onto existing systems.

Actions

- Proactively incorporate trustworthy characteristics into system requirements.
- Consider risk factors related to Human-AI configurations and tasks.
- Analyze dependencies between contextual factors and system requirements. List impacts that may arise from not fully considering the importance of trustworthiness characteristics in any decision making.
- Follow responsible design techniques in tasks such as software engineering, product management, and participatory engagement. Some examples for eliciting and documenting stakeholder requirements include product requirement documents (PRDs), user stories, user interaction/user experience (UI/UX) research, systems engineering, ethnography and related field methods.
- Conduct user research to understand individuals, groups and communities that will be impacted by the AI, their values & context, and the role of systemic and historical biases. Integrate learnings into decisions about data selection and representation.

Transparency and Documentation

Organizations can document the following:

- 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?
- To what extent is this information sufficient and appropriate to promote transparency? Promote transparency by enabling external stakeholders to access information on the design, operation, and limitations of the AI system.
- To what extent has relevant information been disclosed regarding the use of AI systems, such as (a) what the system is for, (b) what it is not for, (c) how it was designed, and (d) what its limitations are? (Documentation and external communication can offer a way for entities to provide transparency.)
- What metrics has the entity developed to measure performance of the AI system?
- What justifications, if any, has the entity provided for the assumptions, boundaries, and limitations of the AI system?

AI Transparency Resources:

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- Stakeholders in Explainable AI, Sep. 2018, URL
- High-Level Expert Group on Artificial Intelligence set up by the European Commission, Ethics Guidelines for Trustworthy AI, URL, PDF.

References

National Academies of Sciences, Engineering, and Medicine 2022. Fostering Responsible Computing Research: Foundations and Practices. Washington, DC: The National Academies Press. URL

Amit K. Chopra, Fabiano Dalpiaz, F. Başak Aydemir, et al. 2014. Protos: Foundations for engineering innovative sociotechnical systems. In 2014 IEEE 22nd International Requirements Engineering Conference (RE) (2014), 53-62. URL

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, et al. 2019. Fairness and Abstraction in Sociotechnical Systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 59–68. URL

Gordon Baxter and Ian Sommerville. 2011. Socio-technical systems: From design methods to systems engineering. Interacting with Computers, 23, 1 (Jan. 2011), 4–17. URL

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices in artificial intelligence. Artificial Intelligence 300 (14 July 2021), 103555, ISSN 0004-3702. URL

Yilin Huang, Giacomo Poderi, Sanja Šćepanović, et al. 2019. Embedding Internet-of-Things in Large-Scale Socio-technical Systems: A Community-Oriented Design in Future Smart Grids. In The Internet of Things for Smart Urban Ecosystems (2019), 125-150. Springer, Cham. URL

MAP-2: Classification of the AI system is performed.

MAP 2.1

The specific task, and methods used to implement the task, that the AI system will support is defined (e.g., classifiers, generative models, recommenders, etc.).

About

AI actors should define the technical learning or decision-making task an AI system is designed to accomplish, along with the benefits that the system will provide. The clearer and narrower the task definition, the easier it is to map its benefits and risks, leading to more fulsome risk management.

Actions

- Define and document AI system existing and potential learning task(s) along with known assumptions and limitations.

Transparency and Documentation

Organizations can document the following:

- To what extent has the entity clearly defined technical specifications and requirements for the AI system?
- To what extent has the entity documented the AI system's development, testing methodology, metrics, and performance outcomes?
- How do the technical specifications and requirements align with the AI system's goals and objectives?
- Did your organization implement accountability-based practices in data management and protection (e.g. the PDPA and OECD Privacy Principles)?
- How are outputs marked to clearly show that they came from an AI?

AI Transparency Resources:

- Datasheets for Datasets, URL.
- WEF Model AI Governance Framework Assessment 2020, URL.
- WEF Companion to the Model AI Governance Framework- 2020, URL.
- ATARC Model Transparency Assessment (WD) – 2020, URL.
- Transparency in Artificial Intelligence - S. Larsson and F. Heintz – 2020, URL.

References

Leong, Brenda (2020). The Spectrum of Artificial Intelligence - An Infographic Tool. Future of Privacy Forum. URL

Brownlee, Jason (2020). A Tour of Machine Learning Algorithms. Machine Learning Mastery. URL

MAP 2.2

Information is documented about the system's knowledge limits, and how output will be utilized and overseen by humans.

About

Once deployed and in use, AI systems may sometimes perform poorly, manifest unanticipated negative impacts, or violate legal or ethical norms. These risks and incidents can result from a variety of factors, including developing systems in highly-controlled environments that differ considerably from the deployment context. Regular stakeholder engagement and feedback can provide enhanced contextual awareness about how an AI system may interact in its real-world setting. Example practices include broad stakeholder engagement with potentially impacted community groups, consideration of user interaction and user experience (UI/UX) factors, and regular system testing and evaluation in non-optimized conditions.

Actions

- Extend documentation beyond system and task requirements to include possible risks due to deployment contexts and human-AI configurations.
- Follow stakeholder feedback processes to determine whether a system achieved its documented purpose within a given use context, and whether users can correctly comprehend system outputs or results.
- Document dependencies on upstream data and other AI systems, including if the specified system is an upstream dependency for another AI system or other data.
- Document connections the AI system or data will have to external networks (including the internet), financial markets, and critical infrastructure that have potential for negative externalities. Identify and document negative impacts as part of considering the broader risk thresholds and subsequent go/no-go deployment as well as post-deployment decommissioning decisions.

Transparency and Documentation

Organizations can document the following:

- Does the AI solution provides sufficient information to assist the personnel to make an informed decision and take actions accordingly?
- To what extent is the output of each component appropriate for the operational context?
- 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?
- Based on the assessment, did your organization implement the appropriate level of human involvement in AI-augmented decision-making? (WEF Assessment)
- How will the accountable AI actor(s) address changes in accuracy and precision due to either an adversary's attempts to disrupt the AI system or unrelated changes in operational/business environment, which may impact the accuracy of the AI system?

AI Transparency Resources:

- Datasheets for Datasets, URL.
- WEF Model AI Governance Framework Assessment 2020, URL.
- WEF Companion to the Model AI Governance Framework- 2020, URL.
- ATARC Model Transparency Assessment (WD) – 2020, URL.
- Transparency in Artificial Intelligence - S. Larsson and F. Heintz – 2020, URL.

References

Context of use

International Standards Organization (ISO). 2019. ISO 9241-210:2019 Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems. URL

National Institute of Standards and Technology (NIST), Mary Theofanos, Yee-Yin Choong, et al. 2017. NIST Handbook 161 Usability Handbook for Public Safety Communications: Ensuring Successful Systems for First Responders. URL

Human-AI interaction

Smith, C. J. (2019). Designing trustworthy AI: A human-machine teaming framework to guide development. arXiv preprint arXiv:1910.03515.

Warden T, Carayon P, Roth EM, et al. The National Academies Board on Human System Integration (BOHSI) Panel: Explainable AI, System Transparency, and Human Machine Teaming. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2019;63(1):631-635. doi:10.1177/1071181319631100

Committee on Human-System Integration Research Topics for the 711th Human Performance Wing of the Air Force Research Laboratory and the National Academies of Sciences, Engineering, and Medicine. 2022. Human-AI Teaming: State-of-the-Art and Research Needs. Washington, D.C. National Academies Press. URL

Ben Green. 2021. The Flaws of Policies Requiring Human Oversight of Government Algorithms. Computer Law & Security Review 45 (26 Apr. 2021). URL

Ben Green and Amba Kak. 2021. The False Comfort of Human Oversight as an Antidote to A.I. Harm. (June 15, 2021). URL

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, et al. 2021. Manipulating and Measuring Model Interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 237, 1–52. 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

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 188 (April 2021), 21 pages. URL

Microsoft Responsible AI Standard, v2. URL

MAP 2.3

Scientific integrity and TEVV considerations are identified and documented including related to experimental design, data collection and selection (e.g., availability, representativeness, suitability), and construct validation.

About

Many AI system risks can be traced to insufficient testing and evaluation processes. For example, machine learning requires large scale datasets. The difficulty of finding the "right" data may lead AI actors to select datasets based more on accessibility and availability than on suitability. Such decisions may contribute to an environment where the data used in processes is not fully representative of the populations or phenomena that are being modeled, inserting or introducing downstream risks.

Other risks arise when selected datasets and/or attributes within datasets are not good proxies, measures, or predictors for operationalizing the phenomenon that the AI system intends to support or inform. Practices such as dataset reuse may also lead to data becoming disconnected from the social contexts and time periods of their creation. Datasets may also present security concerns or be polluted by bad actors in an attempt to alter system outcomes.

Collected data may differ significantly from what occurs in the real world. Large scale datasets used in AI systems often do not include representation of people who have been historically excluded. This may have a disproportionately negative impact on black, indigenous, and people of color, women, LGBTQ+ individuals, people with disabilities, or people with limited access to computer network technologies.

Actions

- Document assumptions made and techniques used during the selection, curation, preparation, and analysis of data, and when identifying constructs and proxy targets, and developing indices – especially when seeking to measure concepts that are inherently unobservable (e.g. "hireability," "criminality." "lendability").
- Map adherence to policies that address data and construct validity, bias, privacy and security for AI systems and verify documentation, oversight,and processes.
- Establish processes and practices that employ experimental design techniques for data collection, selection, and management practices.
- Establish practices to ensure data used in AI systems is linked to the documented purpose of the AI system (e.g., by causal discovery methods).
- Establish and document processes to ensure that test and training data lineage is well understood, traceable, and metadata resources are available for mapping risks.

- Document known limitations, risk mitigation efforts associated with, and methods used for, training data collection, selection, labeling, cleaning, and analysis (e.g. treatment of missing, spurious, or outlier data; biased estimators).
- Establish and document practices to check for capabilities that are in excess of those that are planned for, such as emergent properties, and to revisit prior risk management steps in light of any new capabilities.
- Establish processes to test and verify that design assumptions about the set of deployment contexts continue to be accurate and sufficiently complete.
- Work with domain experts to:
  - Gain and maintain contextual awareness and knowledge about how human behavior is reflected in datasets, organizational factors and dynamics, and society.
  - Identify participatory approaches for responsible Human-AI configurations and oversight tasks, taking into account sources of cognitive bias.
  - Identify techniques to manage and mitigate sources of bias (systemic, computational, human-cognitive) in computational models and systems, and the assumptions and decisions in their development.
- Follow standard statistical principles and document the extent to which the proposed technology does not meet standard validation criteria.
- Investigate and document potential negative impacts due to supply chain issues that may conflict with organizational values and principles.

Transparency and Documentation

Organizations can document the following:

- Are there any known errors, sources of noise, or redundancies in the data?
- Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame
- What is the variable selection and evaluation process?
- How was the data collected? Who was involved in the data collection process? If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)
- As time passes and conditions change, is the training data still representative of the operational environment?
- Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?
- How does the entity ensure that the data collected are adequate, relevant, and not excessive in relation to the intended purpose?

AI Transparency Resources:

- Datasheets for Datasets, URL.
- WEF Model AI Governance Framework Assessment 2020, URL.
- WEF Companion to the Model AI Governance Framework- 2020, URL.
- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- ATARC Model Transparency Assessment (WD) – 2020, URL.
- Transparency in Artificial Intelligence - S. Larsson and F. Heintz – 2020, URL.

References

Challenges with dataset selection

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Front. Big Data 2, 13 (11 July 2019). URL

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, et al. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. arXiv:2012.05345. URL

Catherine D'Ignazio and Lauren F. Klein. 2020. Data Feminism. The MIT Press, Cambridge, MA. URL

Miceli, M., & Posada, J. (2022). The Data-Production Dispositif. ArXiv, abs/2205.11963.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. arXiv:1608.07836. URL

Dataset and test, evaluation, validation and verification (TEVV) processes in AI system development

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

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, et al. 2021. AI and the Everything in the Whole Wide World Benchmark. arXiv:2111.15366. URL

Statistical balance

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 6464 (25 Oct. 2019), 447-453. URL

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, et al. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research. arXiv:2012.05345. URL

Solon Barocas, Anhong Guo, Ece Kamar, et al. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, New York, NY, USA, 368–378. URL

Measurement and evaluation

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 375–385. URL

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, et al. 2022. Evaluation Gaps in Machine Learning Practice. arXiv:2205.05256. URL

Existing frameworks

National Institute of Standards and Technology. (2018). Framework for improving critical infrastructure cybersecurity. URL

Boeckl, K. R., & Lefkovitz, N. B. (2020). NIST privacy framework: A tool for improving privacy through enterprise risk management, version 1.0. URL

MAP-3: AI capabilities, targeted usage, goals, and expected benefits and costs compared with the status quo are understood.

MAP 3.1

Benefits of intended system functionality and performance are examined and documented.

About

AI system benefits should counterbalance the inherent risks and implicit and explicit costs. To identify system benefits, organizations should define and document system purpose and utility, along with foreseeable costs, risks, and negative impacts. Credible justification for anticipated benefits beyond the status quo should be clarified and documented.

Actions

- Utilize participatory approaches and engage with system end users to evaluate system efficacy and interpretability of AI task output.
- Incorporate stakeholder feedback about perceived system benefits beyond the status quo.
- Align system requirements with intended purpose and document decisions.
- Perform context analysis related to time frame, safety concerns, geographic area, physical environment, ecosystems, social environment, and cultural norms within the intended setting (or conditions that closely approximate the intended setting).

Transparency and Documentation

Organizations can document the following:

- Have the benefits of the AI system been communicated to users?
- Have the appropriate training material and disclaimers about how to adequately use the AI system been provided to users?
- Has your organization implemented 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:

- Intel.gov: AI Ethics Framework for Intelligence Community - 2020, URL.
- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI – 2019, LINK, URL.

References

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2021. Hard choices in artificial intelligence. Artificial Intelligence 300 (14 July 2021), 103555, ISSN 0004-3702. URL

Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 39–48. URL

MAP 3.2

Potential costs, including non-monetary costs, which result from expected or realized errors or system performance are examined and documented.

About

Anticipating negative impacts of AI systems is a difficult task. Negative impacts can be due to many factors, such as poor system performance, and may range from minor annoyance to serious injury, financial losses, or regulatory enforcement actions. AI actors can work with a broad set of stakeholders to improve their capacity for assessing system impacts – and subsequently – system risks. Hasty or non-thorough impact assessments may result in erroneous determinations of no-risk for more complex or higher risk systems.

Actions

- Perform a context analysis to map negative impacts arising from not integrating trustworthiness characteristics. When negative impacts are not direct or obvious, AI actors should engage with external stakeholders to investigate and document:
  - Who could be harmed?
  - What could be harmed?
  - When could harm arise?
  - How could harm arise?
- Implement procedures for regularly evaluating the qualitative and quantitative costs of internal and external AI system failures. Develop actions to prevent, detect, and/or correct potential risks and related impacts. Regularly evaluate failure costs to inform go/no-go deployment decisions throughout the AI system lifecycle.

Transparency and Documentation

Organizations can document the following:

- To what extent does the system/entity consistently measure progress towards stated goals and objectives?
- To what extent can users or parties affected by the outputs of the AI system test the AI system and provide feedback?
- Have you documented and explained that machine errors may differ from human errors?

AI Transparency Resources:

- Intel.gov: AI Ethics Framework for Intelligence Community - 2020, URL.
- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI – 2019, LINK, URL.

References

Abagayle Lee Blank. 2019. Computer vision machine learning and future-oriented ethics. Honors Project. Seattle Pacific University (SPU), Seattle, WA. URL

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. arXiv:2011.13416. URL

Jeff Patton. 2014. User Story Mapping. O'Reilly, Sebastopol, CA. URL

Margarita Boenig-Liptsin, Anissa Tanweer & Ari Edmundson (2022) Data Science Ethos Lifecycle: Interplay of ethical thinking and data science practice, Journal of Statistics and Data Science Education, DOI: 10.1080/26939169.2022.2089411

J. Cohen, D. S. Katz, M. Barker, N. Chue Hong, R. Haines and C. Jay, "The Four Pillars of Research Software Engineering," in IEEE Software, vol. 38, no. 1, pp. 97-105, Jan.-Feb. 2021, doi: 10.1109/MS.2020.2973362.

National Academies of Sciences, Engineering, and Medicine 2022. Fostering Responsible Computing Research: Foundations and Practices. Washington, DC: The National Academies Press. URL

MAP 3.3

Targeted application scope is specified, narrowed, and documented based on established context and AI system classification.

About

Systems that function in a narrow scope tend to enable better mapping, measurement, and management of risks in the learning or decision-making tasks and the system context. A narrow application scope also helps ease oversight functions and related resources within an organization.

For example, open-ended chatbot systems that interact with the public on the internet have a large number of risks that may be difficult to map, measure, and manage due to the variability from both the decision-making task and the operational context. Instead, a task-specific chatbot utilizing templated responses that follow a defined "user journey" is a scope that can be more easily mapped, measured and managed.

Actions

- Consider narrowing contexts for system deployment, including factors related to:
  - How outcomes may directly or indirectly impact users and stakeholders.
  - Length of time the system is deployed in between re-trainings.
  - Geographical regions in which the system operates.
- Engage AI actors from legal and procurement functions when specifying target application scope.

Transparency and Documentation

Organizations can document the following:

- To what extent has the entity clearly defined technical specifications and requirements for the AI system?
- How do the technical specifications and requirements align with the AI system's goals and objectives?

AI Transparency Resources:

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI – 2019, LINK, URL.

References

Mark J. Van der Laan and Sherri Rose (2018). Targeted Learning in Data Science. Cham: Springer International Publishing, 2018.

Alice Zheng. 2015. Evaluating Machine Learning Models (2015). O'Reilly. URL

Brenda Leong and Patrick Hall (2021). 5 things lawyers should know about artificial intelligence. ABA Journal. URL

UK Centre for Data Ethics and Innovation, "The roadmap to an effective AI assurance ecosystem". URL

MAP-4: Risks and benefits are mapped for third-party software and data.

MAP 4.1

Approaches for mapping third-party technology risks are in place and documented.

About

Technologies and personnel from third-parties are another source of risk to consider during AI risk management activities. Such risks may be difficult to map since third-party provider risk tolerances may not be the same as the contracting institution.

For example, the use of pre-trained models, which tend to rely on large uncurated web dataset or often have undisclosed origins, has raised concerns about privacy, bias, and unintended effects along with possible introduction of increased levels of statistical uncertainty, difficulty with reproducibility, and issues with scientific validity.

Actions

- Review audit reports, testing results, product roadmaps, warranties, terms of service, end-user license agreements, contracts, and other documentation related to third-party entities to assist in value assessment and risk management activities.
- Review third-party software release schedules and software change management plans (hotfixes, patches, updates, forward- and backward- compatibility guarantees) for irregularities that may contribute to AI system risks.
- Inventory third-party material (hardware, open-source software, foundation models, open source data, proprietary software, proprietary data, etc.) required for system implementation and maintenance.
- Review redundancies related to third-party technology and personnel to assess potential risks due to lack of adequate support.

Transparency and Documentation

Organizations can document the following:

- 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?
- If your organization obtained datasets from a third party, did your organization assess and manage the risks of using such datasets?
- How will the results be independently verified?

AI Transparency Resources:

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020, URL.
- WEF Model AI Governance Framework Assessment 2020, URL.

References

Language models

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. URL

Julia Kreutzer, Isaac Caswell, Lisa Wang, et al. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. Transactions of the Association for Computational Linguistics 10 (2022), 50–72. URL

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, et al. 2022. Taxonomy of Risks posed by Language Models. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 214–229. URL

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

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. 2021. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258. URL

MAP 4.2

Internal risk controls for third-party technology risks are in place and documented.

About

In the course of their work, AI actors often utilize open-source, or otherwise freely available, third-party technologies – some of which have been reported to have privacy, bias, and security risks. Organizations may consider tightening up internal risk controls for these technology sources.

Actions

- Supply resources such as model documentation templates and software safelists to assist in third-party technology inventory and approval activities.
- Review third-party material (including data and models) for risks related to bias, data privacy, and security vulnerabilities.
- Apply controls – such as procurement, security, and data privacy controls – to all acquired third-party technologies.

Transparency and Documentation

Organizations can document the following:

- Did you ensure that the AI system can be audited by independent third parties?
- To what extent do these policies foster public trust and confidence in the use of the AI system?
- 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)?

AI Transparency Resources:

- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020, URL.
- WEF Model AI Governance Framework Assessment 2020, URL.
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI - 2019, LINK, URL.

References

Office of the Comptroller of the Currency. 2021. Comptroller's Handbook: Model Risk Management, Version 1.0, August 2021. Retrieved on July 7, 2022. URL

Proposed Interagency Guidance on Third-Party Relationships: Risk Management, 2021. URL

MAP-5: Impacts to individuals, groups, communities, organizational, or society are assessed.

MAP 5.1

Potential positive or negative impacts to individuals, groups, communities, organizations, or society are regularly identified and documented.

About

AI systems are socio-technical in nature and can have positive, neutral, or negative implications that extend beyond their stated purpose. Negative impacts can be wide- ranging and affect individuals, groups, communities, organizations, and society, as well as the environment and national security.

The Map function provides an opportunity for organizations to assess potential AI system impacts based on identified risks. This enables organizations to create a baseline for system monitoring and to increase opportunities for detecting emergent risks. Impact assessments also help to identify new benefits and purposes which may arise from AI system use. After an AI system is deployed, engaging different stakeholder groups – who may be aware of, or experience, benefits or negative impacts that are unknown to AI actors – allows organizations to understand and monitor system benefits and impacts more readily.

Actions

- Establish and document stakeholder engagement processes at the earliest stages of system formulation to identify potential impacts from the AI system on individuals, groups, communities, organizations, and society.
- Employ methods such as value sensitive design (VSD) to identify misalignments between organizational and societal values, and system implementation and impact.
- Identify approaches to engage, capture, and incorporate input from system users and other key stakeholders to assist with continuous monitoring for impacts and emergent risks. Incorporate quantitative, qualitative, and mixed methods in the assessment and documentation of potential impacts to individuals, groups, communities, organizations, and society.
- Identify a team (internal or external) that is independent of AI design and development functions to assess AI system benefits, positive and negative impacts and their likelihood.
- Develop impact assessment procedures that incorporate socio-technical elements and methods and plan to normalize across organizational culture. Regularly review and refine impact assessment processes.

Transparency and Documentation

Organizations can document the following:

- If the AI system relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?
- If the AI system relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals)
- If the AI system 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:

- Datasheets for Datasets, URL.
- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019, URL.
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020, URL.
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI - 2019, LINK, URL.

References

Susanne Vernim, Harald Bauer, Erwin Rauch, et al. 2022. A value sensitive design approach for designing AI-based worker assistance systems in manufacturing. Procedia Comput. Sci. 200, C (2022), 505–516. URL

Harini Suresh and John V. Guttag. 2020. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. arXiv:1901.10002. Retrieved from URL

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. arXiv:2011.13416. URL

Konstantinia Charitoudi and Andrew Blyth. A Socio-Technical Approach to Cyber Risk Management and Impact Assessment. Journal of Information Security 4, 1 (2013), 33-41. URL

Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.

Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, & Jacob Metcalf. 2021. Assemlbing Accountability: Algorithmic Impact Assessment for the Public Interest. Data & Society. Accessed 7/14/2022 at URL

Ada Lovelace Institute. 2022. Algorithmic Impact Assessment: A Case Study in Healthcare. Accessed July 14, 2022. URL

Microsoft. Responsible AI Impact Assessment Template. 2022. Accessed July 14, 2022. URL

Microsoft. Responsible AI Impact Assessment Guide. 2022. Accessed July 14, 2022. URL

Microsoft. Foundations of assessing harm. 2022. URL

Microsoft Responsible AI Standard, v2. URL

MAP 5.2

Likelihood and magnitude of each identified impact based on expected use, past uses of AI systems in similar contexts, public incident reports, stakeholder feedback, or other data are identified and documented.

About

The likelihood of AI system impacts identified in Map 5.1 should be evaluated. Potential impacts should be documented and triaged.

Likelihood estimates may then be assessed and judged for go/no-go decisions about deploying an AI system. If an organization decides to proceed with deploying the system, the likelihood estimate can be used to assign oversight resources appropriate for the risk level.

Actions

- Establish assessment scales for measuring AI system impact. Scales may be qualitative, such as red-amber-green (RAG), or may entail simulations or econometric approaches. Document and apply scales uniformly across the organization's AI portfolio.
- Apply impact assessments regularly at key stages in the AI lifecycle, connected to system impacts and frequency of system updates.
- Assess system benefits and negative impacts in relation to trustworthy characteristics.

Transparency and Documentation

Organizations can document the following:

- Which population(s) does the AI system impact?
- What assessments has the entity conducted on data security and privacy impacts associated with the AI system?

- Can the AI system be audited by independent third parties?

AI Transparency Resources:

- Datasheets for Datasets, URL.
- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019, URL.
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020, URL.
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI - 2019, LINK, URL.

References

Emilio Gómez-González and Emilia Gómez. 2020. Artificial intelligence in medicine and healthcare. Joint Research Centre (European Commission). URL

Artificial Intelligence Incident Database. 2022. URL

Anthony M. Barrett, Dan Hendrycks, Jessica Newman and Brandie Nonnecke. "Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks". ArXiv abs/2206.08966 (2022) URL

MAP 5.3

Assessments of benefits vs impacts are based on analyses of impact, magnitude, and likelihood of risk.

About

The final output of the Map function is the go/no-go decision for deploying the AI system. This decision should take into account the risks mapped from previous steps and the organizational capacity for their management.

Risk mapping should also list system benefits beyond the status quo. Go/no-go decisions to deploy may be made by an independent third-party or organizational management. For higher risk systems, it is often appropriate – and may well be critical – for technical or risk executives to be involved in the approval of go/no-go decisions to deploy.

The decision to deploy should not be made by AI actors carrying out design and development functions, whose objective judgment may be hindered by the incentive to deploy systems in which they were closely involved.

Actions

- Review and examine documentation, including system purpose and benefits, and mapped potential impacts with associated likelihoods.
- Document the system's estimated risk.
- Make a go/no-go determination based on magnitude, and likelihood of impact. Do not deploy (no-go) or decommission the system if estimated risk surpasses organizational tolerances or thresholds. If a decision is made to proceed with deployment, assign the system to an appropriate risk tolerance and align oversight resources with the assessed risk.

Transparency and Documentation

Organizations can document the following:

- To what extent do these policies foster public trust and confidence in the use 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?
- How has the entity identified and mitigated potential impacts of bias in the data, including inequitable or discriminatory outcomes?

AI Transparency Resources:

- Datasheets for Datasets, URL.
- GAO-21-519SP: AI Accountability Framework for Federal Agencies & Other Entities, URL.
- AI policies and initiatives, in Artificial Intelligence in Society, OECD, 2019, URL.
- Intel.gov: AI Ethics Framework for Intelligence Community - 2020, URL.
- Assessment List for Trustworthy AI (ALTAI) - The High-Level Expert Group on AI - 2019, LINK, URL.

References

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

Elisa Jillson. 2021. Aiming for truth, fairness, and equity in your company's use of AI (April 19, 2021). URL

Sarah Spiekermann and Till Winkler. 2020. Value-based Engineering for Ethics by Design. arXiv:2004.13676. URL

Sri Krishnamurthy. Quantifying Model Risk: Issues and approaches to measure and assess model risk when building quant models. QuantUniversity, Charlestown, MA. URL