NIST AI Risk Management Framework Playbook
– MAP

Abstract

The MAP function establishes the context to frame risks related to an AI system.

Without contextual knowledge, and awareness of risks within the identified contexts, risk management is difficult to perform. MAP is intended to enhance an organization's ability to identify risks and broader contributing factors.

Outcomes in the MAP function are the basis for the MEASURE and MANAGE functions.

Contents

MAP-1: Context is established and understood.

MAP 1.1

Intended purpose, potentially beneficial uses, context-specific laws, norms and expectations, and prospective settings in which the AI system will be deployed are understood and documented. Considerations include: specific set or types of users along with their expectations; potential positive and negative impacts of system uses to individuals, communities, organizations, society, and the planet; assumptions and related limitations about AI system purposes; uses and risks across the development or product AI lifecycle; TEVV and system metrics.

About

AI actors can work collaboratively, and with external parties such as community groups, to help delineate the bounds of acceptable deployment, consider preferable alternatives, and identify principles and strategies to manage likely risks. Context mapping is the first step in this effort, and may include examination of the following: * intended purpose and impact of system use. * concept of operations. * intended, prospective, and actual deployment setting. * requirements for system deployment and operation. * end user and operator expectations. * specific set or types of end users. * potential negative impacts to individuals, groups, communities, organizations, and society – or context-specific impacts such as legal requirements or impacts to the environment. * unanticipated, downstream, or other unknown contextual factors. * how AI system changes connect to impacts.

These types of processes can assist AI actors in understanding how limitations, constraints, and other realities associated with the deployment and use of AI technology can create impacts once they are deployed or operate in the real world. When coupled with the enhanced organizational culture resulting from the established policies and procedures in the Govern function, the Map function can provide opportunities to foster and instill new perspectives, activities, and skills for approaching risks and impacts.

Context mapping also includes discussion and consideration of non-AI or non-technology alternatives especially as related to whether the given context is narrow enough to manage AI and its potential negative impacts. Non-AI alternatives may include capturing and evaluating information using semi-autonomous or mostly-manual methods.

Suggested Actions

- Maintain awareness of industry, technical, and applicable legal standards.
- Examine trustworthiness of AI system design and consider, non-AI solutions
- Consider intended AI system design tasks along with unanticipated purposes in collaboration with human factors and socio-technical domain experts.
- Define and document the task, purpose, minimum functionality, and benefits of the AI system to inform considerations about whether the utility of the project or its lack of.
- Identify whether there are non-AI or non-technology alternatives that will lead to more trustworthy outcomes.
- Examine how changes in system performance affect downstream events such as decision-making (e.g: changes in an AI model objective function create what types of impacts in how many candidates do/do not get a job interview).

- Determine the end user and organizational requirements, including business and technical requirements.
- Determine and delineate the expected and acceptable AI system context of use, including:
  - social norms
  - Impacted individuals, groups, and communities
  - potential positive and negative impacts to individuals, groups, communities, organizations, and society
  - operational environment
- 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.
- 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:

- To what extent is the output of each component appropriate for the operational context?
- 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

Arvind Narayanan. "How to Recognize AI Snake Oil." Arthur Miller Lecture on Science and Ethics (2019). 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

Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul et al. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." Proceedings of the National Academy of Sciences 117, No. 15 (2020): 8398-8403. 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 broad and 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.

Suggested Actions

- Establish interdisciplinary teams to reflect a wide range of skills, competencies, and capabilities for AI efforts. Verify that team membership includes 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

Fernando Delgado, Stephen Yang, Michael Madaio, Qian Yang. (2021). Stakeholder Participation in AI: Beyond “Add Diverse Stakeholders and Stir”. URL

Kush Varshney, Tina Park, Inioluwa Deborah Raji, Gaurush Hiranandani, Narasimhan Harikrishna, Oluwasanmi Koyejo, Brianna Richardson, and Min Kyung Lee. Participatory specification of trustworthy machine learning, 2021.

Donald Martin, Vinodkumar Prabhakaran, Jill A. Kuhlberg, Andrew Smart and William S. Isaac. "Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics", ArXiv abs/2005.07572 (2020). URL

MAP 1.3

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

About

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.

Trustworthy AI technologies may present a demonstrable business benefit beyond implicit or explicit costs, provide added value, and don't lead to wasted resources. Organizations can feel confident in performing risk avoidance if the implicit or explicit risks outweigh the advantages of AI systems, and not implementing an AI solution whose risks surpass potential benefits.

For example, making AI systems more equitable can result in better managed risk, and can help enhance consideration of the business value of making inclusively designed, accessible and more equitable AI systems.

Suggested 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.
- Evaluate AI system purpose in consideration of 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

M.S. Ackerman (2000). The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. Human–Computer Interaction, 15, 179 - 203. URL

McKane Andrus, Sarah Dean, Thomas Gilbert, Nathan Lambert, Tom Zick (2021). AI Development for the Public Interest: From Abstraction Traps to Sociotechnical Risks. URL

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

Iason Gabriel, Artificial Intelligence, Values, and Alignment. Minds & Machines 30, 411–437 (2020). URL

PEAT “Business Case for Equitable AI”. URL

MAP 1.4

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

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 systems’ context of of business use and expectations, organizations can identify and manage these types of risks.

Suggested Actions

- Document business value or context of business use
- Reconcile documented concerns about the system’s purpose within the business context of use compared to 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 system 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

Ethical OS toolkit. 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 and documented.

About

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

Organizations can follow existing regulations and guidelines for risk criteria, tolerance and response established by organizational, domain, discipline, sector, or professional requirements. Some sectors or industries may have established definitions of harm or may have established documentation, reporting, and disclosure requirements.

Within sectors, risk management may depend on existing guidelines for specific applications and use case settings. Where established guidelines do not exist, organizations will want to define reasonable risk tolerance in consideration of different sources of risk (e.g., financial, operational, safety and wellbeing, business, reputational, and model risks) and different levels of risk (e.g., from negligible to critical).

Risk tolerances inform and support decisions about whether to continue with development or deployment - termed "go/no-go". Go/no-go decisions related to AI system risks can take stakeholder feedback into account, but remain independent from stakeholders' vested financial or reputational interests.

If mapping risk is prohibitively difficult, a "no-go" decision may be considered for the specific system.

Suggested Actions

- Utilize existing regulations and guidelines for risk criteria, tolerance and response established by organizational, domain, discipline, sector, or professional requirements.
- Establish risk tolerance levels for AI systems and allocate the appropriate oversight resources to each level.
- Establish risk criteria in consideration of different sources of risk, (e.g., financial, operational, safety and wellbeing, business, reputational, and model risks) and different levels of risk (e.g., from negligible to critical).

- Identify maximum allowable risk tolerance above which the system will not be deployed, or will need to be prematurely decommissioned, within the contextual or application setting.
- Articulate and analyze tradeoffs across trustworthiness characteristics as relevant to proposed context of use. When tradeoffs arise, document them and plan for traceable actions (e.g.: impact mitigation, removal of system from development or use) to inform management decisions.
- Review uses of AI systems for "off-label" purposes, 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:

- Which existing regulations and guidelines apply, and the entity has followed, in the development of system risk tolerances?
- What criteria and assumptions has the entity utilized when developing system risk tolerances?
- How has the entity identified maximum allowable risk tolerance?
- What conditions and purposes are considered "off-label" for system use?

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

Brenda Boultwood, How to Develop an Enterprise Risk-Rating Approach (Aug. 26, 2021). Global Association of Risk Professionals (garp.org). Accessed Jan. 4, 2023. URL

Virginia Eubanks, 1972-, Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor. New York, NY, St. Martin's Press, 2018.

GAO-17-63: Enterprise Risk Management: Selected Agencies' Experiences Illustrate Good Practices in Managing Risk. URL See Table 3.

NIST Risk Management Framework. URL

MAP 1.6

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

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, over-rely on implicit human biases such as confirmation bias and groupthink, and maintain exclusive focus on computational requirements.

Eliciting system requirements, designing for end users, and considering societal impacts early in the design phase is a priority that can enhance AI systems' trustworthiness.

Suggested Actions

- Proactively incorporate trustworthy characteristics into system requirements.
- Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders related to system design or deployment decisions.
- Develop and standardize practices to assess potential impacts at all stages of the AI lifecycle, and in collaboration with interdisciplinary experts, actors external to the team that developed or deployed the AI system, and potentially impacted communities .
- Include potentially impacted groups, communities and external entities (e.g. civil society organizations, research institutes, local community groups, and trade associations) in the formulation of priorities, definitions and outcomes during impact assessment activities.
- Conduct qualitative interviews with end user(s) to regularly evaluate expectations and design plans related to Human-AI configurations and tasks.
- Analyze dependencies between contextual factors and system requirements. List potential 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.)
- How will the relevant 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 the operational/business environment, which may impact the accuracy of the AI system?
- 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

Abeba Birhane, William S. Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel and Shakir Mohamed. "Power to the People? Opportunities and Challenges for Participatory AI." Equity and Access in Algorithms, Mechanisms, and Optimization (2022). 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

Victor Udoewa, (2022). An introduction to radical participatory design: decolonising participatory design processes. Design Science. 8. 10.1017/dsj.2022.24. URL

MAP-2: Categorization 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).

About

AI actors define the technical learning or decision-making task(s) an AI system is designed to accomplish, or 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.

Suggested Actions

- Define and document AI system's 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 about the AI system's knowledge limits and how system output may be utilized and overseen by humans is documented. Documentation provides sufficient information to assist relevant AI actors when making informed decisions and taking subsequent actions.

About

An AI lifecycle consists of many interdependent activities involving a diverse set of actors that often do not have full visibility or control over other parts of the lifecycle and its associated contexts or risks. The interdependencies between these activities, and among the relevant AI actors and organizations, can make it difficult to reliably anticipate potential impacts of AI systems. For example, early decisions in identifying the purpose and objective of an AI system can alter its behavior and capabilities, and the dynamics of deployment setting (such as end users or impacted individuals) can shape the positive or negative impacts of AI system decisions. As a result, the best intentions within one dimension of the AI lifecycle can be undermined via interactions with decisions and conditions in other, later activities. This complexity and varying levels of visibility can introduce uncertainty. And, 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. For example, downstream decisions can be influenced by end user over-trust or under-trust, and other complexities related to AI-supported decision-making.

Anticipating, articulating, assessing and documenting AI systems' knowledge limits and how system output may be utilized and overseen by humans can help mitigate the uncertainty associated with the realities of AI system deployments.

Suggested Actions

- Document settings, environments and conditions that are outside the AI system's intended use.
- Design for end user workflows and toolsets, concept of operations, and explainability and interpretability criteria in conjunction with end user(s) and associated qualitative feedback.
- Plan and test human-AI configurations under close to real-world conditions and document results.
- Follow stakeholder feedback processes to determine whether a system achieved its documented purpose within a given use context, and whether end 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 system provide sufficient information to assist the personnel to make an informed decision and take actions accordingly?
- 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)

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

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

Human Readiness Level Scale in the System Development Process, American National Standards Institute and Human Factors and Ergonomics Society, ANSI/HFES 400-2021

Microsoft Responsible AI Standard, v2. URL

Saar Alon-Barkat, Madalina Busuioc, Human–AI Interactions in Public Sector Decision Making: “Automation Bias” and “Selective Adherence” to Algorithmic Advice, Journal of Public Administration Research and Theory, 2022;, muac007. 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

Mary L. Cummings. 2006 Automation and accountability in decision support system interface design.The Journal of Technology Studies 32(1): 23–31. URL

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M. F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. NYU School of Law, Public Law Research Paper, (20-54). 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

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

Grgić-Hlača, N., Engel, C., & Gummadi, K. P. (2019). Human decision making with machine assistance: An experiment on bailing and jailing. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-25. 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

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

T. Warden, P. Carayon, 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. URL

MAP 2.3

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

About

Standard testing and evaluation protocols provide a basis to confirm assurance in a system that it is operating as designed and claimed. AI systems’ complexities create challenges for traditional testing and evaluation methodologies, which tend to be designed for static or isolated system performance. Opportunities for risk continue well beyond design and deployment, into system operation and application of system-enabled decisions. Testing and evaluation methodologies and metrics therefore address a continuum of activities. TEVV is enhanced when key metrics for performance, safety, and reliability are interpreted in a socio-technical context and not confined to the boundaries of the AI system pipeline.

Other challenges for managing AI risks relate to dependence on large scale datasets, which can impact data quality and validity concerns. The difficulty of finding the "right" data may lead AI actors to select datasets based more on accessibility and availability than on suitability for operationalizing the phenomenon that the AI system intends to support or inform. Such decisions could contribute to an environment where the data used in processes is not fully representative of the populations or phenomena that are being modeled, introducing downstream risks. Practices such as dataset reuse may also lead to disconnect from the social contexts and time periods of their creation. This contributes to issues of validity of the underlying dataset for providing proxies, measures, or predictors within the model.

Suggested Actions

- Identify and document experiment design and statistical techniques that are valid for testing complex socio-technical systems like AI, which involve human factors, emergent properties, and dynamic context(s) of use.
- Develop and apply TEVV protocols for models, system and its subcomponents, deployment, and operation.
- Demonstrate and document that AI system performance and validation metrics are interpretable and unambiguous for downstream decision making tasks, and take socio-technical factors such as context of use into consideration.
- Identify and document assumptions, techniques, and metrics used for testing and evaluation throughout the AI lifecycle including experimental design techniques for data collection, selection, and management practices in accordance with data governance policies established in GOVERN.
- Identify testing modules that can be incorporated throughout the AI lifecycle, and verify that processes enable corroboration by independent evaluators.
- Establish mechanisms for regular communication and feedback among relevant AI actors and internal or external stakeholders related to the validity of design and deployment assumptions.
- Establish mechanisms for regular communication and feedback between relevant AI actors and internal or external stakeholders related to the development of TEVV approaches throughout the lifecycle to detect and assess potentially harmful impacts
- Document assumptions made and techniques used in data selection, curation, preparation and analysis, including:
  - identification of constructs and proxy targets,
  - development of indices – especially those operationalizing 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.
- Identify and document transparent methods (e.g. causal discovery methods) for inferring causal relationships between constructs being modeled and dataset attributes or proxies.
- Identify and document processes to understand and trace test and training data lineage and its metadata resources 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 and other external AI actors to:
  - Gain and maintain contextual awareness and knowledge about how human behavior, organizational factors and dynamics, and society influence, and are represented in, datasets, processes, models, and system output.
  - 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..

- Investigate and document potential negative impacts due related to the full product lifecycle and associated processes 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

Laura Freeman, "Test and evaluation for artificial intelligence." Insight 23.1 (2020): 27-30. URL

Existing frameworks

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

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

MAP-3: AI capabilities, targeted usage, goals, and expected benefits and costs compared with appropriate benchmarks are understood.

MAP 3.1

Potential benefits of intended AI system functionality and performance are examined and documented.

About

AI systems have enormous potential to improve quality of life, enhance economic prosperity and security costs. Organizations are encouraged to define and document system purpose and utility, and its potential positive impacts. benefits beyond current known performance benchmarks.

It is encouraged that risk management and assessment of benefits and impacts include processes for regular and meaningful communication with potentially affected groups and communities. These stakeholders can provide valuable input related to systems' benefits and possible limitations. Organizations may differ in the types and number of stakeholders with which they engage.

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

Suggested Actions

- Utilize participatory approaches and engage with system end users to understand and document AI systems' potential benefits, efficacy and interpretability of AI task output.
- 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 participatory activities such as eliciting, capturing, and synthesizing user, operator and external 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.
- Consider performance to human baseline metrics or other standard benchmarks.
- Incorporate feedback from end users, and potentially impacted individuals and communities about perceived system benefits .

Transparency and Documentation

Organizations can document the following:

- Have the benefits of the AI system been communicated to end users?
- Have the appropriate training material and disclaimers about how to adequately use the AI system been provided to end users?
- Has your organization implemented a risk management system to address risks involved in deploying the identified AI system (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

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 3.2

Potential costs, including non-monetary costs, which result from expected or realized AI errors or system functionality and trustworthiness - as connected to organizational risk tolerance - 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 system non-functionality or use outside of its operational limits, 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 understanding systems' potential impacts – and subsequently – systems' risks.

Suggested Actions

- Perform context analysis to map potential negative impacts arising from not integrating trustworthiness characteristics. When negative impacts are not direct or obvious, AI actors can engage with stakeholders external to the team that developed or deployed the AI system, and potentially impacted communities, to examine and document:
  - Who could be harmed?
  - What could be harmed?
  - When could harm arise?
  - How could harm arise?
- Identify and 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 and documented based on the system's capability, established context, and AI system categorization.

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 TEVV functions and related resources within an organization.

For example, large language models or 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.

Suggested Actions

- Consider narrowing contexts for system deployment, including factors related to:
  - How outcomes may directly or indirectly affect users, groups, communities and the environment.
  - Length of time the system is deployed in between re-trainings.
  - Geographical regions in which the system operates.
  - Dynamics related to community standards or likelihood of system misuse or abuses (either purposeful or unanticipated).
  - How AI system features and capabilities can be utilized within other applications, or in place of other existing processes .
- 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 3.4

Processes for operator and practitioner proficiency with AI system performance and trustworthiness – and relevant technical standards and certifications – are defined, assessed and documented.

About

Human-AI configurations can span from fully autonomous to fully manual. AI systems can autonomously make decisions, defer decision-making to a human expert, or be used by a human decision-maker as an additional opinion. In some scenarios, professionals with expertise in a specific domain work in conjunction with an AI system towards a specific end goal—for example, a decision about another individual(s). Depending on the purpose of the system, the expert may interact with the AI system but is rarely part of the design or development of the system itself. These experts are not necessarily familiar with machine learning, data science, computer science, or other fields traditionally associated with AI design or development and - depending on the application - will likely not require such familiarity. For example, for AI systems that are deployed in health care delivery the experts are the physicians and bring their expertise about medicine—not data science, data modeling and engineering, or other computational factors. The challenge in these settings is not educating the end user about AI system capabilities, but rather leveraging, and not replacing, practitioner domain expertise.

Questions remain about how to configure humans and automation for managing AI risks. Risk management is enhanced when organizations that design, develop or deploy AI systems for use by professional operators and practitioners: - are aware of these knowledge limitations and strive to identify risks in human-AI interactions and configurations across all contexts, and the potential resulting impacts, - define and differentiate the various human roles and responsibilities when using or interacting with AI systems, and - determine proficiency standards for AI system operation in proposed context of use, as enumerated in MAP-1 and established in GOVERN-3.2.

Suggested Actions

- Identify and declare AI system features and capabilities that may affect downstream AI actors' decision-making in deployment and operational settings for example how system features and capabilities may activate known risks in various human-AI configurations, such as selective adherence.
- Identify skills and proficiency requirements for operators, practitioners and other domain experts that interact with AI systems,Develop AI system operational documentation for AI actors in deployed and operational environments, including information about known risks, mitigation criteria, and trustworthy characteristics enumerated in Map-1.
- Define and develop training materials for proposed end users, practitioners and operators about AI system use and known limitations.
- Define and develop certification procedures for operating AI systems within defined contexts of use, and information about what exceeds operational boundaries.

- Include operators, practitioners and end users in AI system prototyping and testing activities to help inform operational boundaries and acceptable performance. Conduct testing activities under scenarios similar to deployment conditions.
- Verify model output provided to AI system operators, practitioners and end users is interactive, and specified to context and user requirements defined in MAP-1.

- Verify AI system output is interpretable and unambiguous for downstream decision making tasks.
- Design AI system explanation complexity to match the level of problem and context complexity.
- Verify that design principles are in place for safe operation by AI actors in decision-making environments.
- Develop approaches to track human-AI configurations, operator, and practitioner outcomes for integration into continual improvement.

Transparency and Documentation

Organizations can document the following:

- What policies has the entity developed to ensure the use of the AI system is consistent with its stated values and principles?
- 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 operational/business environment, which may impact the accuracy of the AI?
- How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?
- Are the relevant staff dealing with AI systems properly trained to interpret AI model output and decisions as well as to detect and manage bias in data?
- What metrics has the entity developed to measure performance of various components?

AI Transparency Resources:

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

References

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

Human Readiness Level Scale in the System Development Process, American National Standards Institute and Human Factors and Ergonomics Society, ANSI/HFES 400-2021.

Human-Machine Teaming Systems Engineering Guide. P McDermott, C Dominguez, N Kasdaglis, M Ryan, I Trahan, A Nelson. MITRE Corporation, 2018.

Saar Alon-Barkat, Madalina Busuioc, Human–AI Interactions in Public Sector Decision Making: “Automation Bias” and “Selective Adherence” to Algorithmic Advice, Journal of Public Administration Research and Theory, 2022;, muac007. URL

Breana M. Carter-Browne, Susannah B. F. Paletz, Susan G. Campbell , Melissa J. Carraway, Sarah H. Vahlkamp, Jana Schwartz , Polly O'Rourke, “There is No “AI” in Teams: A Multidisciplinary Framework for AIs to Work in Human Teams; Applied Research Laboratory for Intelligence and Security (ARLIS) Report, June 2021. URL

R Crootof, ME Kaminski, and WN Price II. Humans in the Loop (March 25, 2022). Vanderbilt Law Review, Forthcoming 2023, U of Colorado Law Legal Studies Research Paper No. 22-10, U of Michigan Public Law Research Paper No. 22-011. URL

S Mo Jones-Jang, Yong Jin Park, How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability, Journal of Computer-Mediated Communication, Volume 28, Issue 1, January 2023, zmac029. URL

A Knack, R Carter and A Babuta, “Human-Machine Teaming in Intelligence Analysis: Requirements for developing trust in machine learning systems,” CETaS Research Reports (December 2022). URL

SD Ramchurn, S Stein , NR Jennings. Trustworthy human-AI partnerships. iScience. 2021;24(8):102891. Published 2021 Jul 24. doi:10.1016/j.isci.2021.102891. URL

M. Veale, M. Van Kleek, and R. Binns, "Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making," in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. Montreal QC, Canada: ACM Press, 2018, pp. 1–14. URL

MAP 3.5

Processes for human oversight are defined, assessed, and documented in accordance with organizational policies from GOVERN function.

About

As AI systems have evolved in accuracy and precision, computational systems have moved from being used purely for decision support—or for explicit use by and under the control of a human operator—to automated decision making with limited input from humans. Computational decision support systems augment another, typically human, system in making decisions.These types of configurations increase the likelihood of outputs being produced with little human involvement.

Defining and differentiating various human roles and responsibilities for AI systems' governance, and differentiating AI system overseers and those using or interacting with AI systems can enhance AI risk management activities.

In critical systems, high-stakes settings, and systems deemed high-risk it is of vital importance to evaluate risks and effectiveness of oversight procedures before an AI system is deployed.

Ultimately, AI system oversight is a shared responsibility, and attempts to properly authorize or govern oversight practices will not be effective without organizational buy-in and accountability mechanisms, for example those suggested in the GOVERN function.

Suggested Actions

- Identify and document AI systems' features and capabilities that require human oversight, in relation to operational and societal contexts, trustworthy characteristics, and risks identified in MAP-1.
- Establish practices for AI systems' oversight in accordance with policies developed in GOVERN-1.
- Define and develop training materials for relevant AI Actors about AI system performance, context of use, known limitations and negative impacts, and suggested warning labels.
- Include relevant AI Actors in AI system prototyping and testing activities. Conduct testing activities under scenarios similar to deployment conditions.
- Evaluate AI system oversight practices for validity and reliability. When oversight practices undergo extensive updates or adaptations, retest, evaluate results, and course correct as necessary.
- Verify that model documents contain interpretable descriptions of system mechanisms, enabling oversight personnel to make informed, risk-based decisions about system risks.

Transparency and Documentation

Organizations can document the following:

- What are the roles, responsibilities, and delegation of authorities of personnel involved in the design, development, deployment, assessment and monitoring of the AI system?
- How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?
- Are the relevant staff dealing with AI systems properly trained to interpret AI model output and decisions as well as to detect and manage bias in data?
- To what extent has the entity documented the AI system's development, testing methodology, metrics, and performance outcomes?

AI Transparency Resources:

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

References

Ben Green, “The Flaws of Policies Requiring Human Oversight of Government Algorithms,” SSRN Journal, 2021. URL

Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn Jonker, Jeroen van den Hoven, Deborah Forster, & Reginald Lagendijk (2021). Meaningful human control: actionable properties for AI system development. AI and Ethics. URL

Mary Cummings, (2014). Automation and Accountability in Decision Support System Interface Design. The Journal of Technology Studies. 32. 10.21061/jots.v32i1.a.4. URL

Madeleine Elish, M. (2016). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction (WeRobot 2016). SSRN Electronic Journal. 10.2139/ssrn.2757236. URL

R Crootof, ME Kaminski, and WN Price II. Humans in the Loop (March 25, 2022). Vanderbilt Law Review, Forthcoming 2023, U of Colorado Law Legal Studies Research Paper No. 22-10, U of Michigan Public Law Research Paper No. 22-011. LINK, URL

Bogdana Rakova, Jingying Yang, Henriette Cramer, & Rumman Chowdhury (2020). Where Responsible AI meets Reality. Proceedings of the ACM on Human-Computer Interaction, 5, 1 - 23. URL

MAP-4: Risks and benefits are mapped for all components of the AI system including third-party software and data.

MAP 4.1

Approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – are in place, followed, and documented, as are risks of infringement of a third-party's intellectual property or other rights.

About

Technologies and personnel from third-parties are another potential sources of risk to consider during AI risk management activities. Such risks may be difficult to map since risk priorities or tolerances may not be the same as the deployer organization.

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

Suggested 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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, William Fedus. "Emergent Abilities of Large Language Models." ArXiv abs/2206.07682 (2022). URL

MAP 4.2

Internal risk controls for components of the AI system including third-party AI technologies are identified 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 may have privacy, bias, and security risks. Organizations may consider internal risk controls for these technology sources and build up practices for evaluating third-party material prior to deployment.

Suggested Actions

- Track third-parties preventing or hampering risk-mapping as indications of increased risk.

- 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 traditional technology risk controls – such as procurement, security, and data privacy controls – to all acquired third-party technologies.

Transparency and Documentation

Organizations can document the following:

- Can the AI system be audited by independent third parties?
- To what extent do these policies foster public trust and confidence in the use of the AI system?
- Are mechanisms established to 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

Kang, D., Raghavan, D., Bailis, P.D., & Zaharia, M.A. (2020). Model Assertions for Monitoring and Improving ML Models. ArXiv, abs/2003.01668. URL

MAP-5: Impacts to individuals, groups, communities, organizations, and society are characterized.

MAP 5.1

Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented.

About

AI actors can evaluate, document and triage the likelihood of AI system impacts identified in Map 5.1 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 TEVV resources appropriate for the risk level.

Suggested Actions

- Establish assessment scales for measuring AI systems' 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 TEVV regularly at key stages in the AI lifecycle, connected to system impacts and frequency of system updates.
- Identify and document likelihood and magnitude of system benefits and negative impacts in relation to trustworthiness characteristics.

Transparency and Documentation

Organizations can document the following:

- Which population(s) does the AI system impact?
- What assessments has the entity conducted on trustworthiness characteristics for example data security and privacy impacts associated with the AI system?
- Can the AI system be tested 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 5.2

Practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, and unanticipated impacts are in place 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.

Organizations can create a baseline for system monitoring to increase opportunities for detecting emergent risks. 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 involved in the design, development and deployment activities – allows organizations to understand and monitor system benefits and potential negative impacts more readily.

Suggested 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 end users and other key stakeholders to assist with continuous monitoring for potential 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.
- Evaluate and document stakeholder feedback to assess potential impacts for actionable insights regarding trustworthiness characteristics and changes in design approaches and principles.
- Develop TEVV procedures that incorporate socio-technical elements and methods and plan to normalize across organizational culture. Regularly review and refine TEVV 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 managed?
- If the AI system relates to other ethically protected groups, 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

Shari Trewin (2018). AI Fairness for People with Disabilities: Point of View. ArXiv, abs/1811.10670. 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 Responsible AI Standard, v2. URL

Microsoft Research AI Fairness Checklist. URL

PEAT AI & Disability Inclusion Toolkit – Risks of Bias and Discrimination in AI Hiring Tools. URL