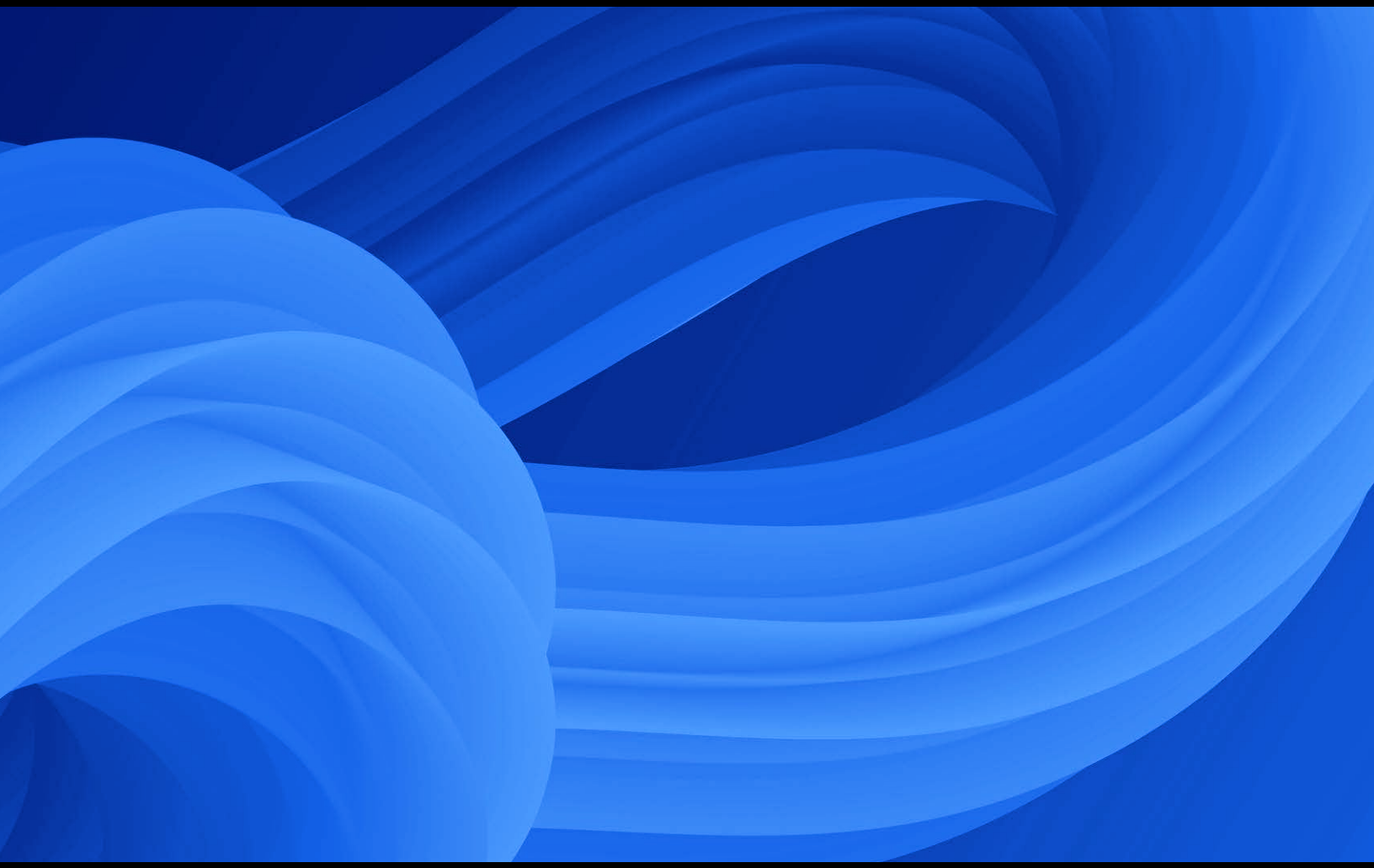
Best Practices in Generative AI

Responsible use and development in the modern workplace



Responsible Artificial Intelligence Institute

Executive Summary

Generative AI, a technology capable of producing realistic content in the form of text, images, sound, and more, presents significant opportunities and challenges for businesses today.

With generative AI (GenAI) applications ranging from customer service automation to content creation, the recent explosive adoption of LLM technologies like ChatGPT underscores the potential transformative scale of AI impact, both positive and negative. Potential risks and harms from generative AI impact human rights, privacy, security, labor, fairness, sustainability, and more. Without investing effort to comprehensively address these issues across the enterprise, businesses are exposed to the risks of compliance penalties, consumer harm, loss of trust, damages, and more.

To position themselves to responsibly capitalize on this potential, organizations must implement governance to pave the way for trustworthy AI deployment, procurement, sale, and use, as applicable. Applying Responsible AI (RAI) frameworks to generative and other forms of AI across the organization can mitigate pressing risks and threats, allowing the technology's potential to be maximized.

The RAI Institute offers the following set of best practices for responsible generative AI use to guide AI practitioners, executive, and other professionals. These guidelines include recommendations related to gathering the right teams and tools, tracking legal requirements, evolving the workforce, and implementing clear objectives and requirements for generative AI.

These best practices are grouped into five categories of Responsible Generative AI:

1. Strategy: Planning, Policies, and Governance
2. Workforce: Training, Education, and Upskilling
3. Capacity: Resourcing and Tools
4. Practice: Development, Procurement, and Use
5. Proactivity: Ongoing Enhancement and Monitoring

Background

Generative AI is a type of artificial intelligence (AI) that creates realistic content like images, text, and videos. It works by using a neural network to learn from a dataset and then generate new content based on what it "learned." However, generative AI can cause serious harm if not used responsibly, such as privacy risks, bias, security threats, lack of transparency, environmental costs, and more. To optimize returns while mitigating risks, businesses must implement Responsible AI frameworks grounded in leading standards and best practices.

Generative AI Today

The launch of ChatGPT in late 2022 precipitated widespread interest in harnessing the capabilities of generative AI, especially of large language models (LLMs). The use of GenAI to enable better customer service will be widespread–according to Salesforce's new generative AI in IT Survey, 77% of senior IT leaders believe that generative AI will help their organization serve their customers faster. For example, generative AI can transform customer service by automating responses to inquiries and providing personalized support, significantly reducing wait times and improving customer satisfaction. It also helps analyze customer feedback in real-time, enabling businesses to swiftly address concerns and tailor services to meet evolving needs.

Beyond customer service, common uses of GenAI include content creation, where it generates articles, reports, and creative writing, enhancing productivity across various sectors. In design and development, GenAI assists in creating software code, architectural plans, and new product concepts. It also plays a crucial role in data analysis, automating the extraction of insights from large data sets. Moreover, personalized education and training solutions benefit from GenAI's ability to adapt learning materials to the user's needs.

As organizations adopt GenAI today, considerations and guidelines for responsible use become increasingly important to address potential biases and ensure privacy and security.

Risks of Generative AI

Generative AI technology can cause harm if not used responsibly. One risk is that it can be used to mislead people with fake videos and images, such as deepfakes used to scam or misrepresent individuals. Generative AI can also cause harm with biased outputs if it's only trained on information from certain groups, which can lead to unfair and unrepresentative outcomes. For example, when prompted to describe or depict a “professional person in the workplace,” an AI system that was trained on biased data might omit photos of women, particularly women of color.

Furthermore, AI systems can generate factually inaccurate outputs, even making up or “hallucinating” research reports, laws, or historical events in their outputs. This hallucination occurs when an AI system learns from data and produces its own new, plausible-seeming but fabricated information. This can occur due to data quality and mitigation issues such as biased or limited training data and model overfitting in response to the data.

There are also other risks associated with generative AI, such as security problems. For example, generative AI can rely on large-scale datasets that hold private information about individuals that can be elicited through prompts. This also poses intellectual property concerns in terms of both the inputs and outputs to the AI. For example, a recent study of over 10,000 employees found that 15 percent of employees input company data into ChatGPT, putting their company at risk of a security breach. Security breaches can happen due to a generative AI system’s vulnerability to threats such as model theft, data poisoning, and adversarial attacks. Another risk is a lack of transparency in the AI’s decision-making processes, which can confuse both users and developers on a system’s outputs and blur the lines of legal liability.

According to a Salesforce survey, many IT leaders share a variety of concerns related to implementing GenAI in their organizations. 79% of leaders were concerned about potential security risks, 59% believed generative AI outputs are inaccurate, and 63% believed that there is bias in generative AI outputs, including misinformation and hate speech. 71% of leaders also believed that generative AI would increase their carbon footprint through increased IT energy use.

LLMs also pose new risks due to their training style and propensity to be procured for third-party use and or integrated into downstream applications. LLMs are notoriously obscure across explainability and interpretability dimensions, leading to unpredictable behaviors and vulnerabilities. There are also privacy and copyright concerns related to the use of users’ and data subjects’ data for training, which may lead to legal challenges and penalties in near future.

Why a Responsible Approach Generative AI

In response to the myriad risks that generative AI poses to both the financial and reputational standing of organizations, as well as to the safety and rights of the public, organizations are urged to incorporate Responsible AI principles into their operations and product development. Responsible AI represents a comprehensive, stakeholder-driven methodology for the design, deployment, and implementation of technology. It emphasizes adherence to regulations, laws, and organizational values as central to AI creation and decision-making processes, applicable at both organizational and product levels.

AI technologies, by their nature, carry inherent risks. The decisions shaping the design, development, deployment, evaluation, and utilization of AI systems often reflect systemic biases and human cognitive limitations. These risks extend beyond individual developers, affecting entire organizations and potentially leading to widespread societal impacts. Traditional corporate governance structures are ill-equipped to keep pace with AI's rapid development, and existing risk management frameworks fail to address the unique challenges AI systems introduce.

AI technologies present both new and amplified risks compared to traditional software. For instance, the data fueling AI systems might not accurately reflect the intended context or use, potentially lacking a reliable ground truth and introducing harmful biases. The reliance on extensive, complex data for training, alongside the potential for significant changes during this process, underscores the uniqueness of AI-specific risks. These include the detachment of training datasets from their intended context, the enormous scale and complexity of AI systems, and the challenges associated with managing pre-trained models.

Furthermore, AI systems face unique challenges such as increased statistical uncertainty, bias management issues, privacy risks due to enhanced data aggregation capabilities, and the necessity for more frequent maintenance. The opacity of AI systems and concerns over reproducibility, coupled with underdeveloped standards for software testing and documentation, highlight the need for AI-specific risk management and strategic planning. These strategies must clearly articulate objectives and delineate human roles and responsibilities in overseeing AI systems, ensuring that organizations can navigate the complexities of generative AI responsibly and effectively.

Furthermore, demand for responsible trustworthy AI is only growing. A study by Edelman showed that 81% of consumers prefer purchasing from companies that prioritize data privacy and security. Furthermore, research by PwC found that 60% of consumers are more likely to trust companies that are transparent about their AI use. According to a report by the Deloitte AI Institute and US Chamber of Commerce, a trustworthy AI approach “can mitigate risks that

might otherwise reduce confidence in AI systems and stifle innovation in this critical sector while focusing investment on beneficial applications of AI that can lead to economic growth and improved health, safety, and well-being." Investment in developing RAI maturity saves organizations by improving trust, market advantage, and more.

Becoming a Leader in Responsible Generative AI

Leaders in the generative AI field have implemented a variety of strategies to ensure the responsible use and development of AI technologies within their organizations. They have disseminated enterprise-wide guidance on best practices, regulations, and intellectual property considerations related to generative AI. Additionally, they've fostered a culture of continuous learning and knowledge sharing through initiatives such as lunch-and-learns, dedicated Slack channels, intranet pages, and repositories. These platforms allow employees to discuss and deepen their understanding of generative AI.

To further commit to trustworthy AI use, these leaders have established internal Responsible AI review boards, incorporating external subject matter experts to align the company's AI applications with its principles and objectives. They've developed comprehensive procedures for tracking and researching leading AI technologies, alongside relevant regulations and mitigation strategies. This includes creating AI-specific procurement and vendor evaluation processes to responsibly acquire third-party AI systems. Employees are encouraged to engage in professional development opportunities related to AI, including training, pilot projects, and contributions to standards and open-source technology research. Moreover, pursuing Responsible AI certification for their systems, in line with national accreditation standards and leading practices, underscores their dedication to responsible AI deployment.

Best Practices

This RAI Institute guide serves as a resource for organizations aiming to foster responsible use of generative AI. It outlines a non-exhaustive list of best practices in GenAI use, procurement, and development. These best practices should be implemented in parallel where appropriate by diverse, inclusive, and cross-functional teams.
These best practices are grouped into five categories of Responsible Generative AI:

1. Strategy: This encompasses Planning, Policies, and Governance, ensuring that the organization's GenAI initiatives are well-aligned with its overall goals and compliant with relevant regulations.
2. Workforce: Focuses on Training, Education, and Upskilling, equipping employees with the necessary knowledge and skills to effectively engage with GenAI technologies.
3. Capacity: Relates to Resourcing and Tools, addressing the need for adequate resources and tools to support GenAI development and deployment.
4. Practice: Covers Development, Procurement, and Use, guiding organizations on responsible GenAI creation, acquisition, and application.
5. Proactivity: Emphasizes Ongoing Enhancement and Monitoring, advocating for continuous improvement and vigilance in GenAI practices.

By adhering to these best practices, organizations can more effectively navigate the complexities of GenAI responsibly, ensuring that their efforts are both fruitful and aligned with broader responsible AI and regulatory standards. Given the rapid evolution of GenAI technology and the continuous development of best practices, this document is regularly updated to remain current and relevant.

Responsible Generative AI Strategy: Planning, Policies, and Governance

The strategic deployment of Generative AI necessitates an integrated tiered approach. Developing a responsible GenAI strategy will involve assembling a cross-functional team to centralize AI expertise and capabilities for enhanced collaboration, governance, and standardization across the organization. In strategic planning, organizations should involve external experts and stakeholders to ensure comprehensive AI development and deployment, alongside clarifying legal requirements to navigate the evolving GenAI landscape and defining clear objectives, use cases, and requirements for GenAI solutions to align with business goals and responsible AI standards.

Assemble a cross-functional team to further AI activities

Centralizing existing AI experience and capabilities, including GenAI in a cross-functional team, allows for efficient knowledge sharing, collaboration, and standardization across different departments and business units. This team, potentially crystallized into a Responsible GenAI Committee, Council, or other formal body, should align ongoing efforts and be accountable for AI governance, standardization of development processes, resource pooling, and internal expertise for GenAI projects. In its scope, this team should be tasked with steering upskilling initiatives, developing a long-term AI roadmap aligned with company objectives, and promoting innovation through challenges and competitions.

- Consider the following roles of functions when assembling this team: AI Product Owner and/or Business Lead, Machine Learning / AI analyst Engineer or Developer, Responsible AI, Data Governance, Legal, Risk, Security, Privacy, Procurement, Internal Audit.
- Additionally, engage external actors, including subject matter experts, users, and at-risk groups, to ensure a comprehensive approach to AI development and deployment within the organization.

Clarify and map legal requirements

Legal and regulatory requirements involving AI and generative AI must be well understood, managed, and documented. Since GenAI is a rapidly evolving field, input from a specialized legal team is needed to not only maintain already well-documented IP and user privacy controls, but to also monitor the legal landscape with respect to upcoming GenAI rulings, laws, and regulations.

- The legal team should have a real-time research function that monitors the legal landscape for GenAI, with the help of external resources and tools. The scope of this function will depend on the organization's functional and geographical needs.
- For example, on intellectual property issues, contracts and licensing agreements of each component of a GenAI model (training data, the model itself, model outputs, and model users) should be reviewed continuously to ensure clear ownership of AI assets.
- The legal team should also work closely with the development and governance teams to minimize the risk of third party information being used without consent. When GenAI models are incorporated into internal or external products, a higher level of legal care is required to continuously review the contracts and/or license agreements related to each model to ensure that models are being used within the remit of their contracts and to determine liability exposure for unintended outputs.
- Standards are a category of "soft law" mechanisms that are a useful complement to the "hard law" of AI laws and regulation that can serve as guidance for an organization's practices, such as ISO/IEC 42001 AI Management Standard, the world's first AI management system standard.

Check out the RAI Institute AI Regulatory Tracker on Airtable here.

Clearly define objectives, use cases, and requirements of GenAI solutions

Articulating the goals and vision driving investment in GenAI solutions helps the organization align efforts with business goals and support efficient resource allocation. Involving key stakeholders across the company helps gather diverse perspectives on current and potential use cases, prioritizing high-impact opportunities.

- Involve key stakeholders to gather diverse perspectives and prioritize high-impact opportunities.
- Adopt an enterprise-wide definition of 'AI System' including GenAI terminology to clarify scope, aid compliance, and prepare for audits.
- Articulate clear objectives to justify the investment of time, money, expertise, and environmental resources in developing GenAI capacity.
- Define use cases to inform the deployment of generative models and enhance return on investment (ROI) calculations.
- Establish comprehensive requirements for selecting appropriate GenAI models across the organization, ensuring alignment with the company's objectives, resources, and technical constraints. Consider the following in standard GenAI solution requirement lists:

- Value alignment: The model type's appropriate prioritization of end-to-end RAI considerations and alignment with your organization
- Input-output combination: Text-to-text, text-to-image, image-to-image, text-to-audio, and others
- Performance: Often given through the size of model parameters
- Interpretability: Open source or closed-source (proprietary services)
- Training data size: Ensure that the model can perform effectively with the training data set
- Flexibility and adaptability: Determine if the model can be fine-tuned or adapted to for the specific tasks required of the company
- Legal considerations: The limitations of a model's usage license
- Timing: Time requirements to train the model vs. project timelines
- Support: If the model is supported by an open source community, a private company, or elsewhere

Understand data flow and consider data quality, diversity, and privacy

Data serves as the cornerstone for all AI models, including GenAI models, where the focus on data quality, diversity, privacy, and other RAI considerations is crucial for creating effective, reliable, and responsible outputs. Ensuring high-quality input and training data enhances the accuracy and robustness of GenAI models so it is crucial to consider considerations for acquiring, using, providing, and managing data related to GenAI.

- Quality: Strong data quality aligns data used for the development and enhancement of AI system with objectives and RAI best practices. Consider articulating and documenting requirements for acquiring, selecting, using, and providing data. Encourage organization-wide data provenance by standardizing processes for verifying and recording the provenance of data used in AI systems over the AI and data lifecycles. Guarantee that GenAI training data is of high quality and directly relevant to the initially defined use cases.
- Diversity and Representativeness: Ensure the training dataset is diverse and representative of various real-world scenarios to minimize bias and enhance the model's generalizability. Encourage tracking of fairness, bias, interpretability, explainability, and other related metrics throughout the lifecycle.
- Privacy: Stay abreast of privacy, human rights, intellectual property, and consumer protection best practices and comply with all applicable AI privacy regulations, This includes efforts such as anonymizing sensitive data to the fullest extent possible and utilizing design options for users to opt-in or use incognito mode to limit data use, where applicable.

Ensure executive buy-in for GenAI efforts

Securing executive buy-in is crucial for ensuring the necessary support, resources, and commitment for the effective deployment of GenAI models and the implementation of supporting organizational policies, governance, frameworks, tools, and training for ethical and responsible AI development. It bridges the gap between short-term tactical decision-making around model deployment and the long-term strategic vision of the organization, emphasizing the importance of executive leadership in managing risks associated with AI system development and deployment.

- As a best practice, organizations are recommended to identify an executive to be accountable for Responsible AI and GenAI initiatives, ensuring they have the authority and resources to lead development efforts and enforce ethical AI practices.
- Executive leadership should actively manage and take responsibility for risk assessments in AI system development and deployment, promoting a balanced and informed approach to mitigating potential issues.

Document a tiered policy for the responsible development or use of AI systems

GenAI introduces complexities that require explicit governance and framework elements beyond those needed for traditional AI, data analytics, or software. Establishing comprehensive organization-wide GenAI policies and guidelines helps ensure that stakeholder expectations are clear and fosters collaboration towards successful outcomes. This approach not only improves governance but also ensures that AI systems are developed and used responsibly.

- Develop Responsible AI policies and practices to enhance governance and ensure ethical use of AI systems. Establish organization-wide GenAI policies and guidelines to set clear expectations and foster collaboration. This can be done in parallel based on context and risk priorities.
- Define terms and objectives within the GenAI policy and management plan to ensure AI use is valid, reliable, safe, secure, resilient, accountable, transparent, explainable, interpretable, privacy-enhanced, and fair. Incorporate mandatory governance gates for AI system implementation to maintain their suitability, adequacy, effectiveness, and responsibility throughout all lifecycle stages.
- Align GenAI policy with existing organizational RAI, AI, and non-AI policies for coherence and comprehensive governance.
- Periodically review and adjust the policy and framework to ensure its ongoing relevance, adequacy, and effectiveness, adapting to the pace of GenAI development.

- Enhance transparency and accountability by communicating externally about the organization's responsible GenAI approach through reporting and publications. Provide accessible information to stakeholders, both external and internal, about AI systems including risk levels, incidents, and other relevant details.
- As an example of foundational RAI frameworks for a policy, the RAI Institute aligns with the NIST AI Risk Management Framework in its Model Corporate AI Policy.

Prioritize a proactive and robust GenAI risk management approach

A comprehensive approach to GenAI risk and harm allows for the early identification and mitigation of potential risks specific to GenAI, ensuring that AI systems are developed, used, and operated securely and effectively. By empowering the risk management function within an organization, it is possible to enable oversight over the evolving landscape of AI technologies, thereby safeguarding intellectual property, data, and the integrity of AI models.

- Empower the risk management function to regularly develop and conduct comprehensive risk assessments, prioritizing risks based on severity and potential organizational impact.
- Formulate clear, actionable risk mitigation strategies and collaborate with development teams for rigorous testing and validation of models in every iteration.
- Carry out periodic audits of AI systems and implement robust security measures to protect AI models, intellectual property, and data.
- Set up continuous monitoring of AI systems in real-world settings to quickly identify and mitigate emerging risks.
- Establish an AI governance framework to oversee risk assessment and management processes, ensuring senior management is informed about training needs to better understand GenAI risks.

Responsible Generative AI Workforce: Training, Education, and Upskilling

To foster a responsible AI workforce, organizations must offer clear, all-encompassing upskilling resources that span legal, privacy, cybersecurity, and responsible AI practices. Organizations should develop a multifaceted approach to this effort, including providing GenAI and RAI guidance, introducing RAI training at all levels, and supporting relevant professional development, repositories, forums, and collaboration avenues. Programs should seek to empower employees to use and develop GenAI as a tool for good by clarifying what "good" looks like in their specific roles. Regularly communicating with stakeholders regarding AI considerations and incorporating feedback from employees, users, and experts is crucial to refine upskilling plans, keeping efforts in sync with rapid technological progress and evolving standards of responsible AI.

Bolster role-relevant and application-specific training

Organization should develop or procure responsible AI training resources for specific roles and determine what kinds of training should be mandatory or encouraged.

- Training programs should include risk, compliance, and RAI training across various functions. Training materials should include case studies and should be engaging, accessible, and effective.
- For example, HR training should cover appropriate uses, risks, and requirements specific to relevant use cases. Technical users should be trained in model assessment frameworks, research reviews, and analytical techniques on a mandatory basis. Compliance professionals should be trained in regulatory requirements related to AI. Application users should have current knowledge in their fields. Executives should be trained on the potential risks of AI, the regulatory environment, responsible AI KPIs, dashboards, and scoresheets. New hires in AI-related roles and people interested in AI should be able to attend general AI training, access cheat sheets, attend lunch-and-learns, and connect with internal and external experts via online resources.
- Staff in many other roles should be trained to act as translators between the business, analytics, and compliance worlds.

Train senior leadership in compliance matters

Organizational leadership should complete executive-level training on responsible AI compliance. Though executives need not be experts on global regulation, it is important to get

an idea of the kinds of objectives and requirements that may be required or expected from AI systems in the future.

- One area of compliance focus recommended for many organizations is the EU's proposed Artificial Intelligence Act, a critical piece of emerging regulation positioned to inform the development of AI systems globally.
- Executive training on legal compliance issues will also build buy-in and help leaders determine approaches to future training and partnerships with research institutes, law firms, consulting firms, information services firm, or other experts.

Upskill legal function

Legal and regulatory requirements involving AI and generative AI must be well understood, managed, and documented. Since GenAI is a rapidly evolving field, a specialized legal team is needed to not only maintain already well-documented IP and user privacy controls, but to also monitor the legal landscape with respect to upcoming GenAI rulings, laws, and regulations.

- On intellectual property issues, contracts and licensing agreements of each component of a GenAI model (training data, the model itself, model outputs, and model users) should be reviewed continuously to ensure clear ownership of AI assets. The legal team should also work closely with the development and governance teams to minimize the risk of third party information being used without consent.
- When GenAI models are incorporated into internal or external products, a higher level of legal care is required to continuously review the contracts and/or license agreements related to each model to ensure that models are being used within the remit of their contracts and to determine liability exposure for unintended outputs.
- The legal team should have a real-time research function that monitors the legal landscape for GenAI, with the help of external resources and tools. The scope of this function will depend on the organization's functional and geographical needs.

Upskill cybersecurity function

Ensuring the confidentiality, integrity, and availability of critical AI assets is imperative to maintaining an organization's competitive advantage. For example, the cybersecurity function needs to be closely involved when GenAI models are trained or fine-tuned on an organization's proprietary information, to ensure that all components of each model are secured based on likely vulnerabilities.

- The cybersecurity function should develop capacity and specialization in securing AI assets, conduct regular threat modeling exercises, develop comprehensive security policies, provide regular cybersecurity training to employees, develop incident response plans, keep current with the latest compliance requirements, and foster a security-conscious culture within the organization.
- If an organization's GenAI models are trained on in-house hardware, cybersecurity teams are needed to consider the optimum data storage practices, secure communication protocols, and strict access control. If models are trained by third party vendors, cybersecurity teams should audit each vendor's security practices to ensure that they conform to the standards set by the organization.
- Resources for optimizing cybersecurity and robustness in AI development should be shared readily throughout the organization through shared tools, groups, and other channels..

Responsible Generative AI Capacity: Resourcing and Tools

Comprehensive planning and consideration of human and technological resources supports throughout successful GenAI development, procurement, and use. Organizations must therefore develop a detailed resource plan that encompasses data, tools, system computing, and human resources, ensuring future readiness and upfront investment in risk management.

Consider GenAI costs, expertise, and resource availability

Evaluating the feasibility of GenAI projects requires a detailed analysis of costs, available expertise, and resources.

- Justify the use of GenAI by aligning it with specific organizational objectives and properly consider non-AI solutions to conserve financial and environmental resources.
- Assess project feasibility by considering the costs of training data, model licensing, hiring expertise, and the environmental impact of training large models.
- Choose model deployment strategies based on cost, security, speed, and control.
- Develop a comprehensive resource analysis for GenAI that includes data, tooling, system computing, and human resources.
- Appropriately invest in RAI and risk management from the start of GenAI exploration in order to future-proof your initiatives.

Scale tools and processes to enable responsibility

Establishing supportive AI tools and processes is crucial for embedding responsibility into GenAI practices and preparing the organization and its teams for increased collaboration, storage and information flow needs, and compliance requirements.

- Explore tools that support clear RAI objectives and measurement capabilities, such as system documentation, evaluation, and storage.
- Use documentation tools, templates, and processes to document AI system information, including updates to processes and performance metrics, to ensure ongoing management, quality, and provenance, such as through model cards.
- Maintain an organization-wide AI system catalog to inventory current, future, and retired use cases.
- Implement an AI scorecard or sheet to quantify business, risk, and impact metrics at various levels, facilitating responsible management and development.

Responsible Generative AI Practice: Development, Procurement, and Use

Incorporating a responsible approach to generative AI across development, procurement, and use of the technology is essential. Organizations must prioritize objective tracking, risk management, and responsible AI practices throughout the lifecycle of AI systems. By defining processes, objectives, intended uses, and "do's and don't's", organizations can ensure GenAI development, procurement, and use align with AI performance, data integrity, interpretability, and responsible AI standards.

Developing Generative AI

Building GenAI responsibly requires grounding efforts in AI development and risk mitigation best practices. To do so, organizations and teams should specify objectives and requirements for responsible development throughout the lifecycle include responsible AI system design, planning, impact assessment, verification, validation, deployment, operation, monitoring, and decommission. Methodology should be tailored to the specific context, use case, and tasks, as, for example, best practices for LLMs can vary greatly from text-to-video products.

As part of these efforts, organizations should enable AI actors to comprehensively document AI system design and development including assumptions, failures, and limitations. Technical documentation must be rigorous, including testing, safeguards, risk mitigation, and event log records. Existing resources can support teams in standardizing its responsible AI artifacts to more efficiently document information related to AI data, model, method, system, and context, such as model cards, datasheets, data statements, dataset nutrition labels, value cards, data cards, system cards, and reward reports. Finally, genAI development teams should engage with subject matter experts and at-risk groups throughout development, including through RAI assessments and independent review processes to incorporate expert perspectives into system operations.

Ground GenAI model development within an established model development practice

GenAI model development should be integrated within an established practice of general model development to ensure systematic and disciplined development and to enable robust, reliable, and scalable models that conform to established development standards. Following a well-structured model development practice improves collaboration, reduces errors, and enables

continuous improvement over time. Furthermore, GenAI should not be the first set of models developed by an organization. If new to AI, the organization should either seek to establish a generalized model development practice first and then consider GenAI models or engage the services of a third-party vendor to ensure that the right expertise is involved in the development or use of these new models.

Consider the following development practices:

- Iterative testing: Implement iterative testing during model development to progressively refine the model's architecture and hyperparameters for optimal performance.
- Output validation: Validate the generated outputs at each iteration to ensure they meet quality standards and align with intended objectives.
- Performance metrics: Define and track relevant performance metrics to quantitatively assess the model's progress and compare different iterations.
- Documentation: Thoroughly document the model development process, including design decisions, parameter choices, and issues encountered, for future reference and knowledge sharing.
- Version Control: Use version control systems to manage model code, data, and configurations, to facilitate collaboration and reproducibility. Data versioning is also important to consider when thinking about reproducibility
- Continuous monitoring: Set up continuous monitoring of the model's performance in real-world scenarios, to identify potential issues and adapt to changing conditions.
    - AI model performance changes with updates from the model developer team.
    - AI models have the potential to 'drift.'
    - GenAI models also have the potential to 'hallucinate.'

Ensure that a responsible AI assessment is built into development review

Responsible AI assessments foster trust among users, customers, and other stakeholders by mitigating risks related to AI validity, reliability, safety, security, resiliency, accountability, transparency, explainability, interpretability, privacy, and fairness.

- Continuously review outputs to ensure fair and trustworthy outcomes. As much as input training data may be scrutinized, there is always a risk of implicit biases hidden in training data expressing themselves in biased model outputs. Use standardized checklists to ensure all appropriate objectives are captured.

- Assemble a diverse review team with domain experts, ethicists, and representatives from communities likely to be affected by model outputs to provide different perspectives during the model output review.
- Employ bias detection techniques to identify biases in the generated outputs based on sensitive attributes (e.g. race, gender).
- Have clear measures in place to action any changes to the model and/or training data, should any biases be detected.

Procuring

Procuring Generative AI

Embedding responsible AI practices into procurement processes ensures alignment with organizational values and compliance standards.

- Support procurement, legal, and compliance teams in AI contracting matters.
- Adopt an AI vendor assessment process for procurement alignment and Prioritize suppliers that to a responsible AI development approach.
- Clearly articulate responsibilities within the AI system lifecycle among all stakeholders including external partners, suppliers, customers, auditors, and other actors
- Prioritize suppliers that to a responsible AI development approach.
- Ensure procurement prioritize are informed by and reflect customer expectations and needs.

Using Generative AI

Given the prevalence of AI tools available for use by the modern workforce, organizations must understand the current state of AI use at the organization, including AI functionalities that may not be clearly marketed as such. Based on the organization's needs and objectives, clearly guiding employees on the responsible use of GenAI is vital to mitigating risks and protecting intellectual property.

Guide employees on the use of GenAI

- Describe which GenAI systems within the organization's productivity suite, procured GenAI systems, or commercially available AI systems (like GPT-4) employees may use, and for which purposes. Incorporate diverse input into GenAI organizational guidance including from technical, Procurement, RAI, Legal, Compliance, and DEI Capacities.

- Advise employees about the potential risks and impacts of using GenAI systems, such as bias, privacy, and hallucination. Provide guidance to employees on which uses are most and least likely to result in user-owned IP.
- Guide employees to existing internal sandboxes or versions of commercially-available GenAI systems which filter information before it is sent or received, to protect sensitive and proprietary information.
- If employees may use GenAI systems, advise employees to fact-check all outputs and consider requiring employees to label GenAI outputs as such.
- Instruct employees not to input into GenAI systems:
    - Personal information;
    - Business information and IP (copyrighted information, patented information, trademarked information, and trade secrets);
    - Classified information; and
    - Information relevant to internal security or access control, such as login credentials.
- Guide employees to existing internal sandboxes or versions of commercially-available GenAI systems which filter information before it is sent or received, to protect sensitive and proprietary information.
- Given the rapidly evolving nature of GPAI, particularly with generative capabilities, understanding their potential impacts, keeping existing guidance up-to-date, and monitoring the regulatory environment will be necessary to insulate organizations from risk.

Responsible Generative AI Proactivity: Ongoing Enhancement and Monitoring

To ensure the responsible use and continuous improvement of generative AI systems, it is crucial to implement proactive measures that encompass ongoing enhancement and diligent monitoring. This is particularly crucial given GenAI's capacity to deviate from its original objectives post-release due to risks like data hallucinations, drift, poisoning, and changing landscapes.

Establish governance to monitor AI activities post-deployment

Post-deployment monitoring and adjustments enable an organization to grasp the extensive impact of AI technologies on individuals, societal groups, and the environment, facilitating proactive mitigation of adverse effects. Periodic audits bolster this oversight, ensuring AI systems uphold responsibility and security standards.

- Foster a culture of continuous learning and adaptation to sllow RAI teams to anticipate and manage emerging risks and ethical challenges effectively.
- Dedicate sufficient time for AI teams to engage in post-deployment monitoring to ensure ongoing effectiveness and ethical compliance of AI applications.
- Perform comprehensive risk hygiene checks organization-wide to assess and refine the impact on individuals, groups, society, and the environment.
- Execute regular audits of AI systems with the help of both internal and external resources, using audits, conformity assessments, and independent evaluations to enhance organizational oversight and align with responsibility, security, and trustworthiness standards.

Establish a process for GenAI concern reporting

Offering an efficient process for performance or concern reporting is vital. Processes should enable both internal and external stakeholders to report any issues or concerns with AI systems, facilitating transparent communication and swift action to address potential problems. Such a reporting mechanism ensures accountability and encourages a collaborative effort towards the responsible evolution of AI technologies.

How the RAI Institute Helps

Generative AI is a rapidly advancing technology that provides both incredible benefits for businesses but also carries significant risks, like privacy risks and security threats.

To fully harness the benefits of this technology, it's crucial to mitigate these risks and avoid the costs of lost revenue, lost customers, and legal fees. Research shows the benefits of operationalizing this risk management approach in an enterprise-wide framework. To set your business up for success with generative AI, you will need a strategy for internal generative AI use and generative AI sales to protect your business and build trust with your consumers.

But figuring out how to do this in practice is easier said than done. The Responsible AI Institute offers the support you need from AI experts - a Generative AI Policy customized for your business needs and objectives. Based on our industry-leading Responsible AI Implementation Framework, we offer our members smart Generative AI guidelines related to corporate best practices, principles, staff training, privacy, liability, all based on cutting-edge best practices, standards, and regulations. We inform our work through our industry-focused Generative AI consortiums. The RAI Institute helps take the guesswork out of what it means to be responsible when it comes to generative AI as well as other kinds of AI in this critical moment.

Terminology Glossary

KEY TERMS:

Artificial Intelligence (AI): Artificial Intelligence (AI) encompasses a broad range of technologies and applications under which GenAI, traditional AI, ML, and data science can belong. Note that across legislation, research, and literature, there are many definitions and interpretations of what "AI" is, and how to define AI is an ongoing discussion in the field. OECD "defines an Artificial Intelligence (AI) System as a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments."

The RAI Institute closely tracks these terminology developments at the regulatory, industry, and ecosystem levels, particularly the definitions established in the NIST AI Risk Management Framework and provisional agreement of the EU AI Act.

Responsible AI (RAI): AI that is built conscientiously and responsibly at every stage of the AI lifecycle. The RAI Institute refers commonly to the definition provided by the World Economic Forum: "the practice of designing, building and deploying AI in a manner that empowers people and businesses, and fairly impacts customers and society — allowing companies to engender trust and scale AI with confidence. Responsible AI is based on principles, such as effective system operations, transparency, explainability, accountability, anti-bias, fairness, consumer protection, sustainability, and robustness. In the growing field of RAI, RAI Institute refers to global principles, standards, regulations, and practices from OECD as well as ISO, IEEE, NIST, UNESCO, EU Ethics Guidelines, ICO Guidance on AI and Data Protection, Global Partnership on AI (GPAI) Framework, WEF, Canada's Directive on Automated Decision-Making Systems, Canada's Office of the Comptroller's Guidance on Model Risk Management, Council of Europe's Report on AI Systems, and more.

Generative AI (GenAI): Generative AI involves artificial intelligence technologies designed to create new content—like text, images, music, or video—that closely mimics human-generated work by learning from extensive data. It differs from traditional AI's typical categorization by its ability to produce distinctly original content, rather than merely analyzing or categorizing existing information. Responsible GenAI focuses on the trustworthy creation, deployment, and application of these technologies, ensuring outputs are equitable, safe, privacy-respecting, and free from harmful biases. It advocates for transparency in how these models are trained and function, alongside implementing safeguards against misuse, such as generating deceptive content. The aim is to leverage generative AI's potential while mitigating its many concerns, fostering positive societal impacts.

RELATED TERMS:

AI ethics: Encompasses the complex value systems and goals embedded in AI technologies, highlighting the importance of designing and deploying AI in ways that align with societal norms and values. Given the dynamic nature of society, ethical considerations in AI are not static and require continuous reassessment to reflect evolving societal expectations and to ensure that AI contributes positively to human welfare. Includes issues of value systems and goals encoded into machines, design ethics, and systemic impacts of AI on social, political, and economic structures. The RAI Institute generally focuses on responsibility in impacts rather than an ethics.

AI governance: The exercise of authority and control (planning, monitoring, and enforcement) over the management of the AI system and related processes.

AI lifecycle: The AI lifecycle involves various stages, from data collection, data analysis, feature engineering, and algorithm selection to model building, tuning, testing, deployment, management, monitoring, and feedback loops for continuous improvement.

AI model: A program that has been trained on a set of data to recognize certain patterns or make certain decisions without further human intervention. (IBM Research)

AI regulation: Includes regulation specific to AI or regulation that relates or intersects with the AI system, like data protection law.

AI risk management process: A structured, measurable process that concerns itself with the management (detection, evaluation, mitigation) of risks (enterprise, user, societal) associated with the development, deployment, and use of AI systems.

AI system: The combination of an AI model with the data used to train, test, validate, and otherwise enable the model's function. It has been built or tailored for a particular objective and context and its design and use are often governed according to a predefined business use case.

Audit: An official inspection of an individual's or organization's AI systems, typically by an independent body. An AI-specific audit could look like a full audit process that yields a certification or one that produces recommendations for improving the system. A related audit into security or privacy could be part of the organization's existing security checks but specifically those done by an independent actor prior to deployment.

Bias: Bias in AI systems can manifest in various forms, including algorithmic bias, data bias, and societal bias, each affecting the fairness and effectiveness of AI applications. Understanding these nuances is critical for developing strategies to mitigate bias and ensure AI systems produce equitable and reliable outcomes. In this context, bias refers to all types of bias, intentional, unintentional, and systemic. Bias mitigation refers to the identification of the origin of bias, assessing the degree of mitigable bias, and mitigating that as fully as possible.

Contingency planning: The extent to which the organization is prepared for adversarial attacks, load inputs, and other edge cases and extreme scenarios, a key component of Robustness and Responsible AI. This also encompasses preparing for technological failures, breaches, errors, as well as other negative impacts. This broad approach ensures organizations are ready to address a wide range of issues, from system malfunctions to breaches in responsible AI guidelines, maintaining the integrity and trustworthiness of AI systems.

Data drift: A type of system drift where variance in independent variable data occurs over time due to changes in the season, consumer preferences, the addition of new products, or other factors.

Data governance: The management of data at an organization, understood by the RAI Institute beyond traditional management practices to include data ethics, emphasizing the responsible collection, use, and sharing of data, which is increasingly central to data governance, ensuring that data practices respect privacy, consent, and equity principles. The Data Management Association (DAMA International) defines data governance as "the exercise of authority and control (planning, monitoring, and enforcement) over the management of data assets." The Data Governance Institute defines data governance as "a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods."

Data provenance: Refers to a record trail that accounts for the origin of a piece of data, data lineage.

Data relevance: The level of consistency between the content of data and the user's areas of interest. In other words, the extent to which data answers or gives insight into the question at hand.

DEI, or Diversity, Equity, and Inclusion: A term used to describe policies and programs that promote the representation and participation of different groups of individuals, including people of different ages, races and ethnicities, abilities and disabilities, genders, religions, cultures, and sexual orientations. Incorporating DEI in the design and deployment of AI systems is crucial for ensuring these technologies are inclusive and equitable. This approach emphasizes the importance of diverse perspectives in creating AI solutions that serve all segments of society, avoiding biases and enhancing fairness.

Design thinking: An iterative problem-solving process that seeks to understand users, challenge assumptions, redefine problems, and create innovative solutions.

Fit for purpose: Refers to an AI system's alignment with its given objectives or purposes, as commonly articulated in project documentation around project charter or plan, product intended use, goals, objective, in-scope bounds, and business justification. "Used informally to describe a process, configuration item, IT service, etc., that is capable of meeting its objectives or service levels. Being fit for purpose requires suitable design, implementation, control, and maintenance" (NIST).

Human-in-the-loop: The process or characteristic of involving humans in a model or system's algorithmic decision-making. Also referred to as human-AI process configuration. Commonly articulated in team process documentation. Human-in-the-loop can be part of human-centered design, a framework that puts real people and their interests at the center of the system development process. Human-in-the-loop configurations in AI systems vary widely, from supervisory roles to direct intervention capabilities or contributions to the learning process. Detailing these variations illustrates the diverse ways human oversight is integrated into AI, ensuring systems remain aligned with human values and judgments

Implementation framework: A set of governance processes, policy documents, and processes that work together to operationalize principles and strategic concepts.

Interpretability: The quality of a given system or product of being easy for people to understand the processes the system uses to arrive at its outcomes.

Machine learning (ML): A subfield of artificial intelligence that gives computers the ability to learn without explicitly being programmed (MIT Sloan).

Organizational governance: ISO 26000 defines organizational governance as "a system by which an organization makes and implements decisions in pursuit of its objectives" (ISO).

RAI training: Training that focuses on RAI principles, can be role or domain-specific but should include basic common principles.

Responsible product design: A product design vision and methodology that incorporates design thinking, human-centered design, consumer protection, safety, privacy protection, fairness, and processes to understand responsible product design considerations in different contexts.

Review of ethics: A formal evaluation of the ethics of a project, process, or system, which could include risk assessment, impact mitigation, stakeholder input collection, and more.

Risk: Defined in ISO standards, for example, as the "combination of the probability of occurrence of harm and the severity of that harm" (ISO IEC Guide 51, ISO 12100, ISO 14971), the RAI Institute evaluates risks in terms of risk, harm, threats, and impacts, identifying if a risk is mitigable or unmitigable, caused by intentional or unintentional action or systemic forces, and categorizing appropriate fit between severity and mitigation action in alignment with best practices. Clarifying the concept of risk in AI involves illustrating potential harms, such as privacy violations, discrimination, or unintended consequences of AI deployment. Providing examples helps stakeholders understand the importance of identifying, evaluating, and mitigating risks to prevent adverse impacts on individuals and society

Risk management: A governance approach that articulates, documents, formalizes, and trains teams on risk-related guidelines, principles, frameworks, and mitigation processes (ISO 31000).

Sustainability: A principle that focuses on the development and deployment of AI in ways that consider their long-term environmental, social, and economic impacts. This approach aligns with global sustainability goals, ensuring that AI contributes to a future where technology supports environmental stewardship, social well-being, and economic prosperity for all.

Traceability: RAI Institute generally incorporates OECD's definition, understanding traceability to focus on maintaining records of data characteristics, such as metadata, data sources, and data cleaning, but not necessarily the data themselves. In this, traceability can help to understand outcomes, prevent future mistakes, and improve the trustworthiness of the AI system. Traceability in AI is not only crucial for understanding outcomes and improving systems but also for ensuring regulatory compliance and auditability.

Transparency: The principle that AI systems should be understandable and explainable to a broad audience, including developers, users, and those impacted by AI decisions. Transparency is vital for building trust, facilitating accountability, and ensuring that AI technologies are used responsibly, with an conscientiousness toward broader user and stakeholder impacts.

NIST AI RMF Characteristics of Trustworthy AI Systems

The RAI Institute aligns with the NIST AI Risk Management Framework (NIST AI RMF) in its framework of trustworthy AI characteristics and related risk management practices. The RAI Institute is also a proud member of the NIST AI Safety Institute Consortium.

The NIST AI RMF lists seven characteristics of trustworthy AI systems: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. Each is listed with its definition, as provided by the RMF, in this section. The characteristics relate to each other as described in the following diagram and caption–accountability and transparency apply to the success of all other characteristics, and validity and reliability are prerequisites for the remaining other characteristics.

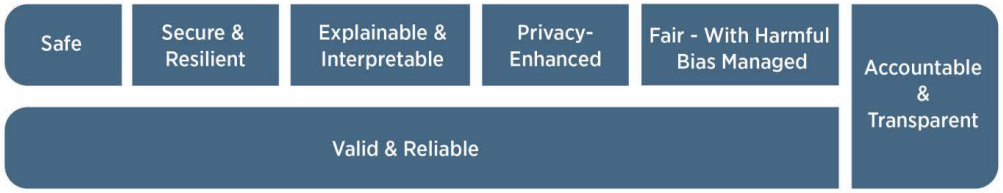


Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

Creating trustworthy AI requires balancing each of these characteristics based on the AI system's context of use. While all characteristics are socio-technical system attributes, accountability and transparency also relate to the processes and activities internal to an AI system and its external setting.

1. **Valid And Reliable**

   Validation is the "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled" (Source: ISO 9000:2015). Deployment of AI systems that are inaccurate, unreliable, or poorly generalized to data and settings beyond their training creates and increases negative AI risks and reduces trustworthiness. Accuracy (the "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true"; Source: ISO/IEC TS 5723:2022) and robustness (the "ability of a system to maintain its level of performance under a variety of circumstances"; Source: ISO/IEC TS 5723:2022) contribute to the validity and trustworthiness of AI systems and can be in tension with one another in AI systems.

Reliability is defined in the same standard as the "ability of an item to perform as required, without failure, for a given time interval, under given conditions" (Source: ISO/IEC TS 5723:2022). Reliability is a goal for the overall correctness of AI system operation under the conditions of expected use and over a given period of time, including the entire lifetime of the system.

2. **Safe**

   Safe AI systems should "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered" (Source: ISO/IEC TS 5723:2022).

3. **Secure And Resilient**

   AI systems, as well as the ecosystems in which they are deployed, are resilient if they can withstand unexpected adverse events or unexpected changes in their environment or use – or if they can maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when this is necessary (Adapted from ISO/IEC TS 5723:2022).

   Secure AI systems can maintain confidentiality, integrity, and availability through protection mechanisms that prevent unauthorized access and use. Security includes resilience but also encompasses protocols to avoid, protect against, respond to, or recover from attacks.

4. **Accountable And Transparent**

   Transparency reflects the extent to which information about an AI system and its outputs is available to individuals interacting with such a system – regardless of whether they are even aware that they are doing so. Meaningful transparency provides access to appropriate levels of information based on the stage of the AI lifecycle and tailored to the role or knowledge of AI actors or individuals interacting with or using the AI system.

   Accountability presupposes transparency. Once a system is transparent, maintaining organizational practices and governing structures for harm reduction, like risk management, can help lead to more accountable systems.

5. **Explainable And Interpretable**

   Explainability refers to a representation of the mechanisms underlying AI systems' operation, whereas interpretability refers to the meaning of AI systems' output in the

context of their designed functional purposes. Explainability can answer the question of "how" a decision was made in the system. Interpretability can answer the question of "why" a decision was made by the system and its meaning or context to the user.

6. **Privacy-Enhanced**

   Privacy refers generally to the norms and practices that help to safeguard human autonomy, identity, and dignity. Privacy values such as anonymity, confidentiality, and control generally should guide choices for AI system design, development, and deployment. Privacy-related risks may influence security, bias, and transparency and come with tradeoffs with these other characteristics.

   Privacy-enhanced AI systems include specific technical features or methods such as de-identification and aggregation for certain model outputs.

7. **Fair with Harmful Bias Managed**

   Fairness in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination. Systems in which harmful biases are mitigated are not necessarily fair. For example, systems in which predictions are somewhat balanced across demographic groups may still be inaccessible to individuals with disabilities or affected by the digital divide or may exacerbate existing disparities or systemic biases.

   Bias is broader than demographic balance and data representativeness. NIST has identified three major categories of AI bias to be considered and managed: systemic, computational and statistical, and human-cognitive. Each of these can occur in the absence of prejudice, partiality, or discriminatory intent.

   Systemic bias can be present in AI datasets, the organizational norms, practices, and processes across the AI lifecycle, and the broader society that uses AI systems.

   Computational and statistical biases can be present in AI datasets and algorithmic processes and often stem from systematic errors due to non-representative samples.

   Human-cognitive biases relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about the purposes and functions of an AI system. Human-cognitive biases are omnipresent in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI.

NIST AI System Lifecycle Stages

1. **Plan and Design:** Assesses to what extent the AI system team has established appropriate overarching and system-specific responsible AI governance related to the end-to-end design of the system. Covers concepts of risk mitigation, compliance, cybersecurity, sustainability, compliance, human capital allocation, team expertise, and guiding policy documentation.

2. **Collect and Process Data:** Assesses the AI system's responsible data governance plan and practice. Covers concepts of responsible AI plan and practice related to data quality, integrity, user consent, drift, and disposal, and related.

3. **Build and Use Model:** Assesses the degree of responsible AI concepts embedded in the system's model governance and operations. Covers concepts of the model's governance, design, and testing elements related to transparency, explainability, interpretability, fit, drift, and robustness.

4. **Verify and Validate:** Assesses the extent to which an AI system team appropriately verifies and validates the AI system's performance along responsible AI requirements. Covers concepts of testing and corrective action related to the AI system product performance, software performance, robustness, bias and fairness, impact assessment, generalizability, and transparency.

5. **Deploy and Use:** Assesses the extent to which responsible AI requirements are built into the AI system team's plan for deployment and use of the product. Covers concepts of AI system product rollout, documentation, training, user empowerment—AI involvement communication, data use consent, user testing, accessible design, feedback loops, and recourse.

6. **Operate and Monitor:** Assesses the plan for continuous monitoring and responsible operation of the AI system product. Covers concepts of post-deployment quality control, change logging, downstream product management, incident management, contingency planning, and product degradation.

Supporting You on Your RAI Journey

The Responsible AI Institute is here to support organizations on their AI journeys. Becoming a member enables essential support and direction for making significant progress toward future-proofing RAI governance and implementation.

Click below to learn more about how we help our members achieve their Responsible AI goals:

About Responsible AI Institute (RAI Institute)

Founded in 2016, the Responsible AI Institute (RAI Institute) is a global and member-driven non-profit dedicated to enabling successful responsible AI efforts in organizations. We accelerate and simplify responsible AI adoption by providing our members with AI assessments, benchmarks and certifications that are closely aligned with global standards and emerging regulations.

Where to connect with us: