Responsible Use of AI Guide

DECEMBER 2024

Contents

Responsible AI at AWS

At Amazon Web Services (AWS), we see the transformational nature of artificial intelligence (AI) across industries every day. AI is used to help improve healthcare, advance brain research, enable sustainable aquaculture practices, and accelerate the building and deployment of climate solutions—among many more use cases that help address some of society's greatest challenges. Given the breadth and depth of AI tools and technologies, many customers are asking for perspectives on how to responsibly design, develop, deploy, and operate AI systems. **At AWS, we are committed to developing AI responsibly and take a people-centric approach that prioritizes education, science, and our customers, to integrate responsible AI across the end-to-end AI lifecycle.** We believe the use of AI must respect the rule of law and human rights, and we encourage the safe and responsible development of AI as a force for good.

This document shares some recommendations that can be used across four major phases of the AI lifecycle: design, develop, deploy, and operate. **The field of responsible AI is a rapidly developing area, so these recommendations should be viewed as a starting point and not the final answer.** We encourage readers to consider the spirit and intent behind the recommendations. Responsible AI requires a shared commitment between developers, deployers, and end users of AI systems.

At AWS, we are committed to working alongside others to develop AI technology responsibly and build trust. We're collaborating with the U.S. Artificial Intelligence Safety Institute Consortium, established by the National Institute of Standards and Technology (NIST), and are engaging with multistakeholder organizations, such as the OECD.AI working groups, the Partnership on AI, the Responsible AI Institute, Frontier Model Forum, and EqualAI, alongside strategic partnerships with universities on a global scale, to promote responsible AI practices. Read more about our commitment to the responsible use of AI.

We are also eager to receive feedback and appreciate the opportunity to contribute to this important topic while continuing to learn from the broader community.

How to use this guide

This guide offers considerations for designing, developing, deploying, and operating AI systems responsibly, based on our extensive learnings and experience in AI. It was written with a set of diverse AI stakeholders and perspectives in mind—including, but not limited to, builders, decision makers, and end users.

Recommendations in this guide should also be considered along with other third-party and AWS resources for responsible development and operation of AI systems, such as AWS AI Service Cards. To learn more about how to build and operate AI systems responsibly using AWS services, refer to the AWS Responsible AI page.

Excelling at responsible AI

Organizations build responsible AI capabilities through a programmatic approach with specific objectives, dedicated leaders, metrics, mechanisms, and resourcing. The journey typically proceeds in four phases: awareness, foundations, exploration, and scaling.

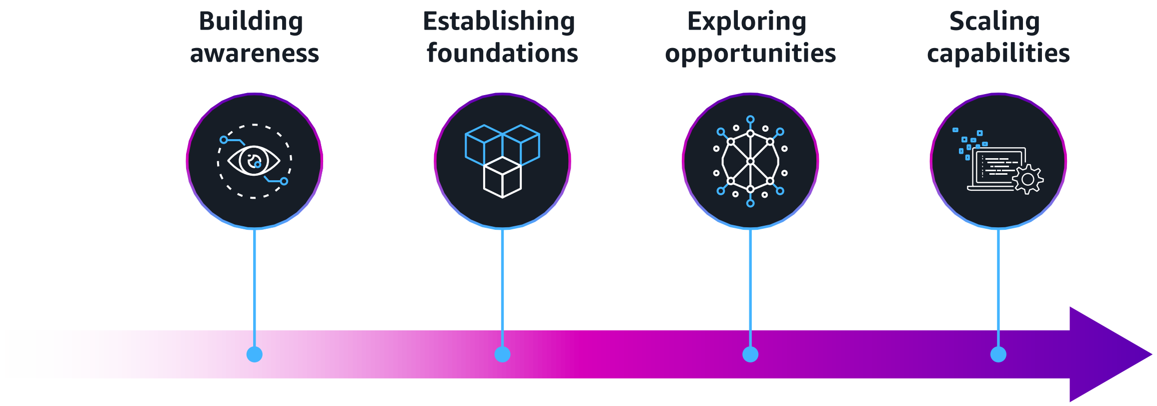


FIGURE 1: OVERVIEW OF THE RESPONSIBLE AI JOURNEY FROM BUILDING AWARENESS TO SCALING CAPABILITIES

To start, your organization should build its awareness of the general opportunities that AI offers and the general responsible AI challenges, such as regulatory and technical, that AI poses. What business use cases, either internally or for your customers, might benefit the most from AI support? For each of the use cases, which of the three roles—developer, deployer, end user—would your organization be playing? With example use cases ready, investigate technical feasibility and specific responsible AI risks, zoom in on a target opportunity set, and identify the specific gaps in organizational capabilities needed to turn the opportunities into actual gains.

In the second phase, lay the foundations for responsible AI. You should consider creating a multidisciplinary organizational focal point, such as a core responsible AI team or dedicated responsible AI group, and run lightweight trial projects. A focal point does not replace organizational processes but rather explores solutions for responsible AI use case assessment, risk triage, and other governance issues. It will drive thoughtful questioning about which responsible AI practices will be best practices in your organization's business contexts. Your trial projects will test the relevance of your existing governance practices and can help you converge on the right AI stack to simultaneously support technical development and regulatory compliance.

In the exploration phase, run a carefully chosen set of pilot projects that build end-to-end solutions for your internal or external customers. These projects will have sharply defined use case targets and be expected to deliver business benefits. Teams working on these projects will likely encounter issues of fairness, privacy, and transparency, among others, and should work with the organization's responsible AI focal point to set internal standards, and with external stakeholders to secure feedback.

Finally, with pilot results in hand, scale up investment to pursue your full range of AI opportunities. This will require integrating responsible AI practices into the core operations and decision-making processes of your organization. Consider whether it makes sense to pursue international standard certifications, such as ISO 42001, and engage with key stakeholders in jurisdictions relevant to your business.

Even within your organization, you will find there is no one-size-fits-all approach to responsible AI. For some use cases, your organization may build solutions from the ground up; for others, especially generative AI use cases, your organization may prefer to use existing fully managed solutions and focus solely on operating an application.

AI roles: Developers, deployers, and end users

When it comes to designing, developing, deploying, and operating AI systems, we distinguish between developers, deployers, and end users of AI systems and emphasize the shared responsibility between everyone involved in the AI lifecycle:

AI developers
Those who create and develop AI models or systems

AI deployers
Those who deploy an AI system to end users

AI end users
Those providing inputs or receiving outputs from an AI system or model

Each of these roles plays a part in the AI lifecycle. AI developers define intended use cases and assess potential risks of an AI system based on the ways that deployers might integrate it. AI deployers compare their unique operating context to the intended use of the system, carefully assessing the suitability and performance. AI end users are encouraged to share feedback with developers or deployers, as their insights can contribute to improvements.

In some cases, AI deployers may choose to adapt or create bespoke AI systems themselves and thereby become AI developers. This underscores the fluidity between the roles and the need for collaboration across the entire AI supply chain.

Responsible AI considerations throughout the AI lifecycle

When designing, developing, deploying, or operating an AI application, try to systematically consider potential limitations and risks that may arise. One way to do this is by establishing a set of guiding principles, or dimensions, that can be applied at various stages of the AI lifecycle. These dimensions of responsible AI can vary depending on the organization and its specific needs, but responsible AI dimensions should all serve to promote and integrate responsible design, development, deployment, and operation of AI. Dimensions can be used to create an organizational structure and mechanisms that allow builders, decision makers, and users to systematically ask questions and make decisions.

At AWS, we consider the following dimensions: fairness, transparency, privacy and security, explainability, safety, controllability, veracity and robustness, and governance. By implementing dimensions and practices for responsible AI within your organization, or by adopting established frameworks, you can create effective AI applications while also mitigating potential harms and risks. For example, at AWS, we frame veracity, or truthfulness, and robustness as "achieving correct system outputs, even with unexpected or adversarial inputs."

Due to the inherent need to balance tradeoffs between dimensions depending on context and use case, it can also be helpful to view responsible AI dimensions as considerations rather than requirements. Note that these dimensions may vary from project to project and might also change in the future as new scientific progress makes updates necessary, as was the case for generative AI systems in recent years. For example, the open-ended nature of generative systems and the rapid adoption of this new technology have given rise to new risks and challenges, such as veracity from the example above. Also keep in mind that responsible AI dimensions should be accompanied by a set of mechanisms that help weave responsible practices into every stage of the AI lifecycle. The mechanisms will differ depending on the responsible AI dimension you consider. For instance, while a mechanism supporting security might involve implementing encryption protocols and access controls to protect data, a mechanism for fairness could include the curation of diverse training datasets.

In addition to formulating responsible AI dimensions, it can also be helpful to look at how these dimensions are applied in different ways across the AI lifecycle, from design to operation. Note that there is no one-size-fits-all approach to the different AI lifecycle stages. For certain use cases, your organization might consider working through all three stages, from design of the application to operating the final product. For other use cases—especially generative AI use cases—your organization may prefer to use existing fully managed solutions and focus solely on deploying and operating an application.

Considerations that apply to all phases

Certain practices and considerations apply to all phases of the AI lifecycle, as they help form the foundation for responsible AI. These practices include promoting governance practices and cultivating a culture of responsible innovation.

Include diverse backgrounds: Consider diverse perspectives, backgrounds, skills, and experiences on teams that are developing AI systems. Assess whether teams include a wide array of genders, races, ethnicities, abilities, ages, religions, sexual orientations, military statuses, backgrounds, and political views. Further, monitor whether teams may have gaps and consider adding underrepresented perspectives to fill them. Successful teams will likely have cross-functional expertise, such as technologists, academics, industry experts, lawyers, social scientists, and other stakeholders, and diverse characteristics to help ensure important perspectives are taken into account. Consider resources, such as user testing, focus groups, or third-party organizations, to obtain additional perspectives from outside parties.

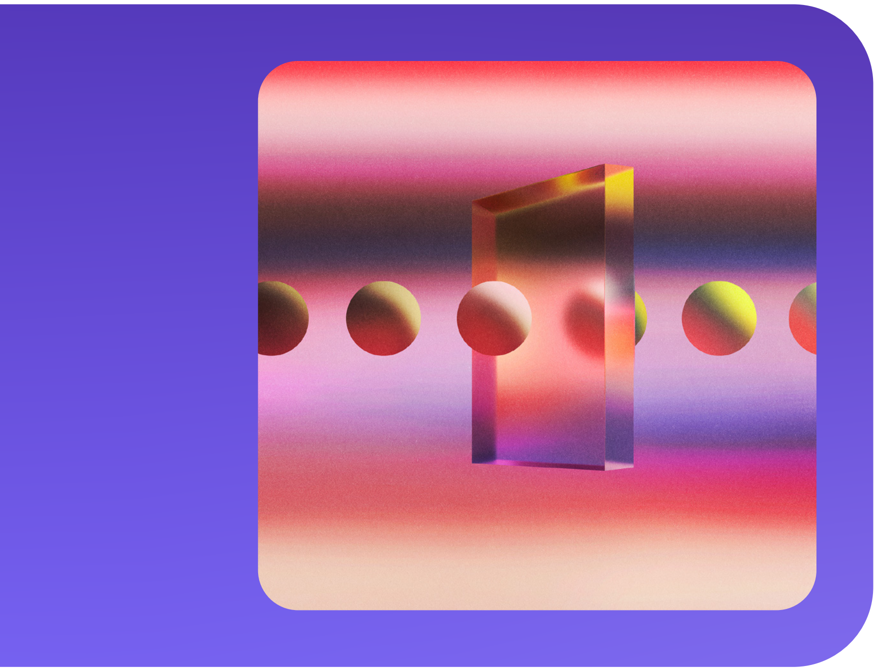
Engage with independent assessors: Consider establishing independent, diverse teams to help assess and test for potential harms or issues throughout the AI lifecycle. In some cases, it may be appropriate to have external parties conduct these types of evaluations rather than rely solely on internal teams. Having independent, multidisciplinary groups assess responsible AI systems can provide an important check and balance. These teams can bring diverse perspectives, expertise, and a degree of distance from the immediate build process.

Consider relevant laws and regulations: Engage with legal advisors to assess your use case for compliance with applicable laws and regulations throughout all phases of design, development, deployment, and operation of AI systems. This may include vetting legal rights to use data and models and determining the applicability of laws around privacy, biometrics, antidiscrimination, and other regulations on specific use cases. Be mindful of differing legal requirements across states, provinces, and countries, as well as new AI regulations being considered and proposed around the world. Given the rapid evolution of AI technology and capabilities, anticipate procedural and regulatory changes, and assess whether to update internal policies and processes accordingly. Also consider implementing restrictions and human oversight for AI systems with significant potential for misuse or unintended consequences when deployed. AI is a constantly evolving landscape, and new techniques, technologies, laws, and social norms will continue to be developed and evolve over time. Therefore, it is critical that everyone working with AI systems stays educated on these issues.

Establish governance mechanisms: Consider establishing best practice guides and policies within your organization that outline the acceptable use and management of AI within your organization. Involve diverse perspectives in the policy-setting process to align with organizational goals and values, and consider implementing training programs or materials to educate members about roles, responsibilities, and responsible AI fundamentals. Periodically review and update processes to maintain the relevance and effectiveness of these governance mechanisms over time. Beyond internal governance, consider whether external policies are appropriate to govern AI technology use. For example, at AWS, we created the AWS Responsible AI Policy, which applies for the use of AI services, features, and functionality that AWS provides.

Create transparency artifacts: Documentation facilitates communication of information about the AI system between stakeholders. It captures relevant design decisions and inputs, making it useful for tracking potential problems and assisting both internal and external teams in evaluating the AI system. To support transparency, consider creating documentation artifacts, such as datasheets, model cards, or system cards. These documents provide valuable information that can be used to assess AI applications and give guidance for builders, decision makers, and users, enabling them to make more informed decisions about adopting or using an AI system.

At AWS, we provide various transparency artifacts, such as AWS AI Service Cards, that list the intended use cases and limitations, responsible AI design choices, and deployment and performance optimization best practices for some of our AI services. AWS AI Service Cards are part of our comprehensive process to responsibly build AWS AI services and provide increased transparency to help customers better understand our AI services. Another mechanism to support transparency are model cards. Model cards are structured documents that provide information about a model's intended use cases, training details, evaluation metrics, results, observations, and recommendations, along with other relevant information. Model cards can guide users to operate the model in a way that is appropriate and safe.



The design phase

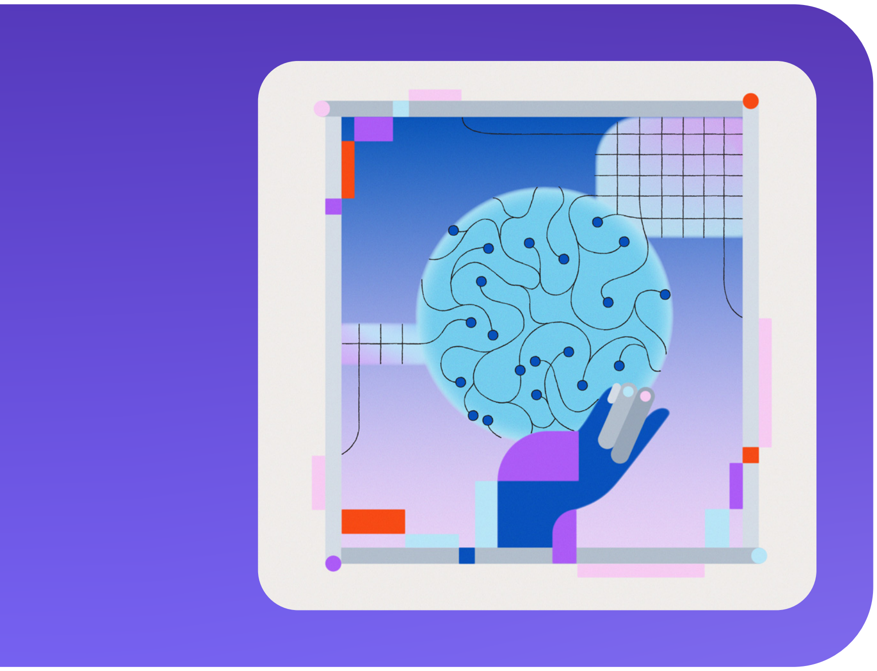
The design phase includes defining use cases and requirements for an AI system, establishing performance criteria, and exploring the potential impact of the system on users and other parties.

Define use case: There are a wide variety of use cases that can incorporate AI, with different goals, characteristics, user bases, and potential impacts. Defining a use case consists of creating a description of the business problem and the workflow that solves the problem, including key inputs and outputs. In some cases, the workflow may contain multiple components—in those instances, it can be helpful to describe how different components are interacting with each other. In addition to the business problem itself, the use case description should also contain a list of stakeholders that are involved in the business problem and their objectives. Finally, consider confounding or intrinsic variation in inputs and types of errors and their impact of the proposed solution. **Consider the potential impact of an AI system on stakeholders that are not customers or direct users of the system but may still be affected.** For example, if an autonomous vehicle is not operating as expected, it could have an impact on passengers, other drivers, pedestrians, or property. AI developers should consider the intended purpose of the system and, as appropriate, anticipate other likely uses and foreseeable misuses.

Assess risks: Risk management is a process that helps minimize the effect of potential negative impacts while also providing opportunities to create better products or applications for end users. **Consider the benefits and potential risks of your specific use case.** Given the broad nature and applicability of AI, many applications may pose limited or no risk, such as movie recommendation systems, while others could involve significant risk, especially if used in a way that impacts privacy or safety. Examples of risks worth carefully evaluating include **technical limitations of an AI system, over-reliance on limited data or inaccurate output, the potential for bias in training data or the model itself, and intentional or unintentional misuse.** When performing a risk assessment, try to consider the severity of the risk and its likelihood in order to help prioritize risk treatment.

There are many ways to conduct a risk assessment. You can use existing risk management frameworks, such as NIST AI Risk Management Framework, or create a framework that suits your team or organization. Generally, risk reports will include a detailed description of the use case, an overview of stakeholders, a list of potential risks, and a suggested rating and mitigation action that can lower the risk. Going through the risk assessment exercise can allow you to gain deeper insights into the potential impact and risks for a wide range of stakeholders and help create a more trustworthy, robust, and safer AI application. For a walk-through example of a risk assessment, you can review the AWS blog "Learn how to assess the risk of AI systems".

Identify limitations: Builders and decision makers should understand the nature, capabilities, and limitations of AI systems, including important concepts like the probabilistic nature of algorithms, confidence levels, and human review. Many AI systems predict a possible or likely answer, not the answer itself. **The probabilistic nature of AI means that use cases that require definitive answers, as opposed to possible or likely answers, may benefit from additional guardrails.** This holds especially true for generative AI systems, a technology that creates content that varies with repeated tries. This is in contrast to more traditional applications of machine learning, which typically solve focused and narrow prediction problems. Additionally, generative AI introduces challenges that are new or different than those with predictive models—such as the potential for AI systems to generate factually incorrect statements. When considering the capabilities of AI systems, it is also valid to ask whether AI is the right approach to solve the problem altogether as some problems are not suited to be solved using AI. Also keep in mind that AI applications should be monitored and updated continuously, which may also require additional effort to uphold performance.



The development phase

The development phase is a dynamic and iterative process. For AI developers, this phase includes collecting and curating training and testing data, building, and testing system components. For AI deployers, the goal of the development phase is to adapt an AI system into a functional application.

Define requirements: In addition to traditional performance metrics, consider the need to explain the methodology and important factors that influence the AI system's output. Note that there is currently no one-size-fits-all solution for explaining outputs, something that holds especially true for highly complex and large models, such as foundation models (FMs). While some areas, such as the explainability of models using structured tabular data, have seen significant progress and can aid in clarifying certain predictions, the field of explainability continues to evolve. The importance of explainability can also vary per use case. Systems that have low or no risk may not require explainability while AI systems with output that may be used in a manner that could impact human rights or safety will likely need a method for explaining or providing insights into how the system performed its analysis or the factors influencing outputs. If explainability is not technically feasible, consider whether other mechanisms, such as human review, auditability, and refocusing or limiting the scope of the use case, might serve as an appropriate alternative.

Anticipate use cases: AI developers should develop metrics and a test plan to measure performance of the system against anticipated production uses. They should consider running ongoing tests against a frequently updated, high-quality dataset. For deployers, testing should include not only the AI system itself but also the overall process it's a part of, including decisions or actions that might be taken based on system output. In some situations, it may not be appropriate to use the system if testing does not reach a specified accuracy level.

Train and test data: Consider how you will acquire data to train and test your AI system or application. For example, data may be available through a variety of sources, including publicly available and licensed data and proprietary data. Involve your legal and procurement teams as appropriate to assess the impact of any privacy considerations or other relevant laws, licenses, or contractual requirements that may impact your collection or use of the data. **Consider any necessary processes for handling data securely and safely and ways to mitigate risk.** For example, if certain portions of a dataset are sensitive but aren't necessary for development of the model, consider whether you can discard that content.

It's important to appropriately secure data used for building and testing AI systems and applications. Consider using encryption and secure storage measures to protect data, particularly when dealing with sensitive information. Encryption can be used to protect data both at rest and in transit so that it remains confidential and secure. Make sure to comply with applicable data protection and privacy laws, and use data minimization techniques where appropriate. A thoughtful data preparation process is fundamental for constructing effective, responsible AI systems, including data provenance and modifications made to raw data.

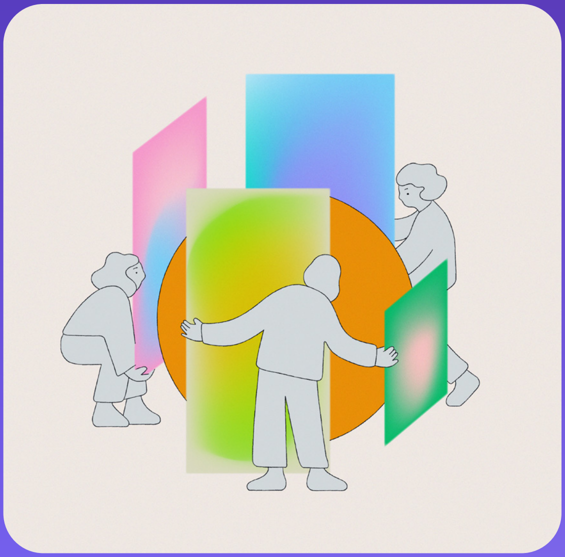
Assess representativeness of training and testing data: When collecting data to develop and test AI systems or applications, consider the data's completeness and representativeness. Diversity of data—in terms of source, type, demographic, geographic, temporal, or other aspects—can help the AI system or application operate as intended. This is particularly important in circumstances where a system is being used as part of a decision-making process that may have a material impact on a person's fundamental rights, health, or safety. **Develop mechanisms to evaluate whether the data appropriately represents real-world use, and collect and test additional data to address underrepresented attributes.** For example, for audio transcription, you may need data with different accents, speech speeds, vernacular, and background environments. Autonomous transport systems, however, may need data from different terrains and obstacles, such as cobblestones, dirt, and cracked sidewalks. Also review data for freshness because it may be outdated and in need of replacement and examine potential sources of error, which may be inherent to the data itself, in its structure and organization, or introduced during annotation. When working with data, be mindful of potential cognitive biases that could inadvertently impact the training process, such as confirming preconceived notions or neglecting contradictory evidence. Also make sure to create separate sets of data for training and testing of systems or applications, both of which should be complete and representative.

Consider adopting datasheets or similar governance mechanisms to document the composition, licensing, and provenance of a dataset. Datasheets can also help tackle responsible AI concerns by stating the purpose of the data collection, providing an overview of representation concerns if applicable, and outlining the intended use of the dataset. When curating a dataset for your use case, the creation of a datasheet allows other members in your team or organization to better understand limitations.

Use adversarial-style testing: Adversarial testing, or red teaming, is an adversarial attack simulation of the AI system usually conducted by AI developers with the goal to identify vulnerabilities which might be exploited by an attacker. Red teaming is mostly appropriate for complex, large models and may not be necessary for more traditional AI models and use cases. Also keep in mind that red teaming alone is not a comprehensive solution to validate all real-world harms associated with AI systems. It should be included with other forms of testing, evaluation, and verification, such as assessments by independent third-party teams. While it is possible to achieve zero errors against a fixed test set, the goal of red teaming is to iteratively explore more use cases and prompt variations, so it’s important to continue to red team with every model iteration.

Assess performance: Whenever possible, try to use multiple datasets and human workforces to evaluate the performance of your system or application, as it is unlikely that a single evaluation dataset can provide an absolute picture of performance. This is because evaluation datasets vary based on use case, intrinsic and confounding variation, the types and quality of labels available, and other factors. While automated testing provides useful feedback, it does not always correlate well with human assessment. Using human judgement is critical for assessing the effectiveness of generative AI systems or applications, because people are better equipped to understand the context, intent, and nuances of complex creative tasks. Keep in mind that the particular medium in which an AI system is trained on, such as images, tabular data, and spoken or written language, matters greatly in how we analyze and understand it.

Implement risk mitigation techniques: Consider mitigations depending on the type of AI application and the risks that were identified as part of the risk assessment process. For example, for generative AI applications, mitigation techniques may include value alignment through reinforcement learning from human feedback (RLHF), creation of prompt templates, or augmentation of the system to include additional data sources that can help improve the truthfulness of the output. Different risks will require different approaches for mitigation to be implemented. For example, when the training data for a model contains personal information, you might want to consider privacy preservation techniques; for data that includes sensitive attributes, additional fairness measures can be added to the model training process to reduce disparity in model performance. Check out how to tune models and the full tutorial.



The deployment phase

In the deployment phase, the AI system is actively moved into production. For a deployer, this includes preparing the AI system for use. For a developer, this includes understanding and accounting for risks and limitations associated with the specific use case and deployment in an application that's facing end users.

Oversee AI systems: As noted earlier, AI systems generate predictions of a possible or likely answer, not the answer itself. If confidence indicators are available, take them into account (or instruct your users to take them into account) when reviewing and acting on system outputs. Be mindful of overreliance on confidence indicators or situations where confidence scores may be used as shortcuts to make decisions. In highly dynamic settings or scenarios where context is relevant, confidence scores may lead to misguided decisions. Therefore, regardless of confidence levels, consider whether human review or oversight over the system operation may be appropriate or necessary, such as in situations where AI systems may be used in a manner that impacts human rights or safety. If it is, consider how to best incorporate such human input into the overall operation of the system. Human reviewers should be appropriately trained on real-world scenarios, including examples where the system fails to properly process inputs or cannot handle edge cases, and have ways to exercise meaningful oversight.

Test for specific use cases: Deployers should consider whether a particular AI system is appropriate for their use case, including any benefits, limitations, and risks. This should be reassessed if the system is used for new or different use cases or beyond original scope, and it should also be cross-referenced with any relevant intended-use information provided by the AI developer. Before live deployment, it is important to test AI systems in the operational environments and on the data on which they will be deployed. Deployers should also factor in localization considerations when deploying an AI system into a new region or geography—for example, real estate pricing models in different geographic areas or voice recognition systems deployed in areas with different dialects or accents.

Validate and improve: AI systems can be subject to concept drift, where system behavior changes as a result of changes in users, environments, or data over time. Develop and run ongoing performance tests and use these test results and feedback to identify areas where additional data or development may improve your system's performance. Continue to assess accuracy and monitor for potential bias, including that your models perform as expected across different segments. Consider appropriate adjustments to both the system and overall processes that involve the system, such as updated training, new notices or restrictions, or optimizing the ways system output is evaluated and used.

Consider versioning and rollback: Maintain version control of model and data updates, and put a well-documented process in place to enable rolling back to a previous version. If needed, create a backup version that provides basic functionality. This can be especially important in situations where a model or system update unexpectedly introduces unintended behavior. Regularly review the versioning and rollback processes and test them to ensure they work as expected. When making updates to the AI system, plan and communicate necessary changes, monitor the transition, and be prepared to revert if issues arise.

The operate phase

The operate phase deals with the ongoing operation of the system after it's developed and deployed. Note that many considerations and questions from earlier phases are still relevant.

Notify users and consider accessibility: Consider whether to inform end users about the use of AI in the system they are interacting with, such as notifying them that they are interacting with a chatbot and not a live human. Also consider whether it's appropriate or feasible to allow end users to opt out or bypass interacting with the AI system and offer an alternate method to accomplish the use case. For example, some users may prefer not to use a facial recognition authentication system and request a different method of authentication. Consult accessibility resources to assess whether the system is usable by the target audience and provides appropriate access options to all intended users.

Provide and use feedback mechanisms: Since AI systems can continue to learn and improve throughout their lifecycle, an important aspect of improvement involves receiving and incorporating feedback from users and stakeholders. Consider soliciting feedback through programmatic and manual methods, including in-system mechanisms or third-party outreach through surveys and focus groups. If appropriate for the use case, consider mechanisms for users or stakeholders to request more information about how system output is used.

Use content authentication and tracking: Content authentication can help increase transparency around AI-generated content. Watermarks are one type of content authentication mechanism that can be used to verify whether digital content, such as images and videos, was AI-generated. Additionally, consider including Content Credentials, another provenance/authentication technology. Content Credentials are based on an open technical specification developed and maintained by the Coalition for Content Provenance and Authenticity (C2PA), a cross-industry standards development organization. They can enable users to identify AI-generated content, provide transparency about the source and creation process (origin and history of content), allow verification of content provenance, and empower users to make informed decisions about the use of AI-generated content.

Implement safeguards: As a deployer of AI systems, consider using safeguarding mechanisms, such as guardrails, to enhance the safety and reliability of your AI system. Safeguarding mechanisms can act as protective barriers and help limit undesirable or harmful outputs. For instance, you should consider using guardrails to constrain the inputs or outputs for a deployed AI system, which can help ensure that they operate within predefined boundaries. Safeguarding mechanisms can range from simple lists of words to filter to regular expressions or fully automated metric-based or model-based guardrails that can identify the intent of the user or the response. Depending on the application, you might want to either completely prevent certain outputs, such as hate speech, or it might be sufficient to obfuscate specific parts of the output, such as personally identifiable information.

At AWS, we developed Amazon Bedrock Guardrails, an API which allows you to implement safeguards for your generative AI applications based on your use cases and responsible AI policies. For some use cases, it may also be helpful to refine system outputs by creating a human review workflow.

Conclusion

By embracing responsible AI, your organization can harness the transformative power of this technology while proactively mitigating risks and building trust with customers and stakeholders. This involves establishing awareness, building foundations, and scaling up capabilities, allowing you to incorporate responsible AI principles and practices throughout the AI lifecycle. The recommendations provided in this guide serve as a starting point, with the understanding that responsible AI practices will need to adapt and advance alongside the progress of AI capabilities. For more details about responsible AI at AWS, including tools and services, refer to the AWS Responsible AI page.