Responsible AI for mission-based organizations

Machine learning (ML) and artificial intelligence (AI) are transformative technologies, enabling organizations of all sizes to further their mission in ways not previously possible. From understanding member sentiment better to assisting individuals displaced during a humanitarian disaster to improved intelligent document processing (IDP), ML is helping mission-based organizations more fully meet their mission objectives.

It is critical to think responsibly about these technologies so that all users of the ML system are treated fairly, data is appropriately protected, and individuals can make informed choices about consent. In this post, I discuss responsible AI and how you should think about your workloads. This approach will help ensure your AI systems are fair, transparent, and secure.

What is responsible AI?

As technology evolves, the definition of responsible AI will also evolve. At its heart, responsible AI is about respecting the rule of law, human rights, and values of equity, privacy, and fairness. As you read this post, consider these responsible AI recommendations and the intent behind them, along with third-party and other guidance from Amazon Web Services (AWS).

AWS views responsible AI across eight areas:

- Fairness
- Explainability
- Privacy and security
- Safety
- Controllability
- Veracity and Robustness
- Governance
- Transparency
- Considering ML workloads across these dimensions is critical for mission-based organizations, as a lapse in any area could dramatically affect an organization's ability to deliver on its mission and negatively impact its stakeholders.

Fairness

It's important to consider how an ML system might impact a subpopulation of users. Gender and economic bias can easily creep into an ML system without diligent effort. Mission-based organizations must be particularly cautious here, as unintentional bias in an ML system could result in an inaccurate medical diagnosis or being unfairly denied a financial product or important service.

Organizations can minimize the effect of bias in their ML systems by ensuring they staff their teams with individuals from diverse backgrounds, skills, perspectives, and demographics. They can create fairness goals across various subpopulations and periodically test to ensure they're meeting their

PUBLIC SECTOR

Responsible AI for mission-based organizations

Machine learning (ML) and artificial intelligence (AI) are transformative technologies, enabling organizations of all sizes to further their mission in ways not previously possible. From understanding member sentiment better to assisting individuals displaced during a humanitarian disaster to improved intelligent document processing (IDP), ML is helping mission-based organizations more fully meet their mission objectives.

It is critical to think responsibly about these technologies so that all users of the ML system are treated fairly, data is appropriately protected, and individuals can make informed choices about consent. In this post, I discuss responsible AI and how you should think about your workloads. This approach will help ensure your AI systems are fair, transparent, and secure.

What is responsible AI?

As technology evolves, the definition of responsible AI will also evolve. At its heart, responsible AI is about respecting the rule of law, human rights, and values of equity, privacy, and fairness. As you read this post, consider these responsible AI recommendations and the intent behind them, along with third-party and other guidance from Amazon Web Services (AWS).

AWS views responsible AI across eight areas:

- Fairness
- Explainability
- Privacy and security
- Safety
- Controllability
- Veracity and Robustness
- Governance
- Transparency
- Considering ML workloads across these dimensions is critical for mission-based organizations, as a lapse in any area could dramatically affect an organization's ability to deliver on its mission and negatively impact its stakeholders.

Fairness

It's important to consider how an ML system might impact a subpopulation of users. Gender and economic bias can easily creep into an ML system without diligent effort. Mission-based organizations must be particularly cautious here, as unintentional bias in an ML system could result in an inaccurate medical diagnosis or being unfairly denied a financial product or important service.

Organizations can minimize the effect of bias in their ML systems by ensuring they staff their teams with individuals from diverse backgrounds, skills, perspectives, and demographics. They can create fairness goals across various subpopulations and periodically test to ensure they're meeting their

objectives. They can use independent teams to test the system for bias, and when discrepancies are found, develop a plan to identify the root cause, acquire more data to help mitigate the bias, and retrain the model.

Additionally, organizations using Amazon SageMaker should consider using SageMaker Clarify. Using SageMaker Clarify during the build phase of the ML lifecycle can help you identify imbalances and biases in your input features. It offers the ability to balance your data through various techniques. Once your model is trained, SageMaker Clarify can check your model for potential bias by identifying predictions that occur more frequently for one group over another.

Explainability

Historically, ML systems have not provided very good justification for their predictions. While statistical systems (like linear or logistic regression, or analysis of variance) typically provide good explainability because of their straightforward calculations, predictions from more complicated ML systems—especially deep learning systems—are difficult to explain.

Some ML models are easier to explain than others, especially regression or classification-type problems based on tabular data. However, it's important to understand that there is no one-size-fits-all approach for understanding why a model made a specific prediction.

As with fairness, organizations that use SageMaker should also consider the use of SageMaker Clarify. SageMaker Clarify can provide metrics to help identify which features of your dataset contributed the most to certain predictions. SageMaker Clarify works with tabular, natural language processing (NLP), and computer vision models.

Consider the intended use of your ML system to determine if explainability is required. A system that extracts document contents may not require explainability. But a system that potentially impacts the safety or security of an individual or subpopulation very likely requires some sort of prediction explainability.

Privacy and Security

Security is job zero at AWS. This means that security is the most important thing we do and it's fundamental to every job we carry out. ML workloads introduce new security threats that organizations must take care to mitigate. Organizations must protect their ML models running in production from fraud, waste, and abuse, in addition to the data used to train those models.

In some cases, the ML workloads running in mission-based organizations contain extremely sensitive data that could harm the organization or individuals if it's breached. It's also important to be diligent about protecting the ingest process for critical data. By targeting the data ingest process, a threat actor may be able to poison training data, which could result in a weaker ML model, or even render the ML model unable to make accurate predictions.

Read about how to improve the security of your ML workloads. As a best practice, you should review your workload using the security pillar of the Well-Architected Framework Machine Learning Lens.

Safety

Safe ML systems prevent harmful system output and misuse of the ML system. The principal of safety is applicable to every ML system, but it's most common to see this property in action with generative AI workloads. The advent of generative AI has the potential to radically transform virtually every area of a customer or member experience, and AWS is providing a number of solutions to ensure generative AI is used in a way that cannot generate harmful output. Amazon is committed to promoting AI safety and recently joined the Frontier Model Forum to help advance AI safety.

The Amazon Titan Image Generator is a foundation model that allows users to create realistic, studio-quality images in large volumes and at low cost, using natural language prompts. All images generated by the Amazon Titan Image Generator include an invisible watermark by default, which helps content creators, news organizations, and fraud detection teams identify AI-generated images that were created using the Amazon Titan Image Generator model. This feature helps organizations mitigate harmful content and reduces the spread of misinformation.

Amazon Bedrock also implements automated abuse detection mechanisms to identify and prevent harmful content from being generated. Amazon Bedrock uses classifiers to automatically detect harmful content, recurring behavior, and detects and blocks child sexual abuse material (CSAM).

Controllability

Controllability involves having mechanisms to monitor and steer AI system behavior. Like safety, the principle of controllability applies to all ML systems, but is commonly seen in generative AI workloads. Guardrails for Amazon Bedrock is one set of tools organizations can leverage to improve the controllability of their ML workloads. Guardrails can be applied across all models used within Amazon Bedrock, including custom fine-tuned models. Organizations can leverage Guardrails with configurable thresholds to filter hate speech, insults, misconduct, and prompt injection attacks. Guardrails can also redact sensitive PII it detects and can block interactions with your generative AI workloads when specific words or phrases are detected.

Guardrails are also available in Amazon Q Business, which can be used to control how Amazon Q, a generative-AI powered assistant, responds to specific topics in the chat. Amazon Q can be configured with how it should respond when relevant answers aren't found in your enterprise data. You can also apply topic level controls to specific users or groups.

Veracity and robustness

If an ML system is making important decisions, an organization will want confidence that the system is not easily fooled. If the ML system can produce potential inaccuracies, consider how this data could be used. An organization may want to limit the scope of the ML system and reduce access to it. It's critical to develop a plan for how to deal with inaccuracies.

An ML model might be designed to operate in very specific conditions and should not be used outside of those parameters. In many organizations, ML practitioners struggle with democratizing information about ML models across the enterprise. Amazon SageMaker Model Cards solves this problem by providing a single source of truth about a model, its intended uses, training details, metrics, observations, and additional call-outs. Amazon SageMaker users will find that their model

training details will transfer automatically to their model card. Organizations using third-party models should consider reviewing the ML model's card to understand any important details and use cases before running the model in production. Organizations using AWS' AI Services (like Amazon Rekognition and Amazon Transcribe) should consider reviewing the AWS AI Service Cards for the applicable service.

Mission-based organizations should consider the adoption of model cards into their ML lifecycle. Through the use of model cards, an organization can consistently document model behavior and assumptions. Model cards are also useful documentation in regulated environments where audits are common and provide an easy way to share model information with external sponsors.

Governance

Governance is the process to ensure that responsible AI is happening consistently across the organization. Amazon's founder, Jeff Bezos, has said that good intentions don't work. Telling someone to "try harder" or "do better next time" rarely results in the outcome you want. Instead, you must build a mechanism that replaces a human best effort with a scalable, repeatable process so that you achieve the desired outcome.

With ML governance, you must build a mechanism within the organization to ensure the steps required during the development and operation of the ML workload are being completed appropriately. Having a written record can help both internal teams and external auditors validate the presence of a well-functioning governance process. Organizations should consider adding ML governance as an activity to their Cloud Center of Excellence (CCoE).

Transparency

Transparency involves the communication from the organization to the users of the ML system. An organization may wish to consider letting all users of the system know to what extent an ML model is being used. Consider accessibility factors to ensure that all intended users of a system have access to the same level of service.

In cases where the ML model may provide low confidence in the prediction, organizations should consider using tools like Amazon Augmented AI (Amazon A2I). Amazon A2I allows organizations to create a human-in-the-loop workflow to validate low-confidence predictions or periodically spot-check ML predictions for accuracy.

As an example, a membership application may allow visitors to upload photos. In the terms of use of the site, the organization may state what constitutes an appropriate image and that images will be processed using ML to detect inappropriate content. Inappropriate content that is detected with low confidence could be reviewed by a human workforce to validate predictions and allow for images that are false positives to be processed

Conclusion

In this post, I outlined the current areas in which AWS thinks about responsible AI. This is a continually evolving field, but it's important to consider these factors in your workloads so you can maintain the security, privacy, and trust of your users. Responsible AI is especially important for mission-based organizations to consider, as a lapse in any of the areas discussed could result in an impact on mission delivery or on the organization's users.

I encourage you to consider using SageMaker Clarify for bias detection and explainability and SageMaker Model Cards to help democratize ML information across your organization. Also, be sure to review your ML workloads using the Machine Learning Lens of the Well-Architected Framework, and consider using Amazon A2I for situations where you require a human-in-the-loop as part of your ML workflow.

For generative AI workloads, consider using Amazon Bedrock to take advantage of the built-in safety features. Consider using Guardrails for Amazon Bedrock and guardrails in Amazon Q Business to prevent generative AI models from discussing topics you want to avoid.

For further reading, check out the AWS guidance on the responsible use of AI and ML.

Thank you for downloading this AWS Blog! Carahsoft is the distributor for AWS public sector solutions available via GSA, NASPO, The Quilt and other contract vehicles.

To learn how to take the next step toward acquiring AWS's solutions, please check out the following resources and information:

For additional resources:
carah.io/AWS-Resources

For upcoming events:
carah.io/AWS-Events

For additional AWS solutions:
carah.io/AWS-Solutions

For additional public sector solutions:
carah.io/AWS.Solutions

To set up a meeting:
AWS@carahsoft.com
888-662-2724

To purchase, check out the contract vehicles available for procurement:
carah.io/AWS-Contracts