
AWS
re:Invent
NOV. 27 – DEC. 1, 2023 | LAS VEGAS, NV

AIM220

Responsible AI in the generative era: Science and practice

Michael Kearns
(he/him)
Amazon Scholar
Amazon Web Services &
University of Pennsylvania

Peter Hallinan
(he/him)
Senior Manager
Amazon Web Services

Agenda

01 Science of responsible AI

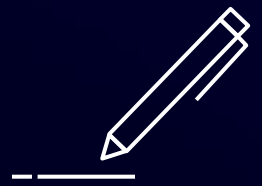
02 Emerging challenges in generative AI

03 Practice of responsible AI

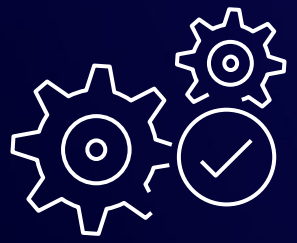
04 Q&A

What is generative AI?

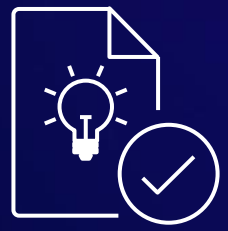
How is generative AI being used?



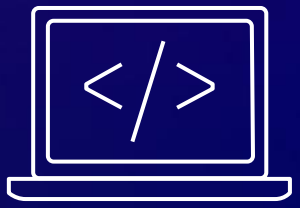
Writing tool



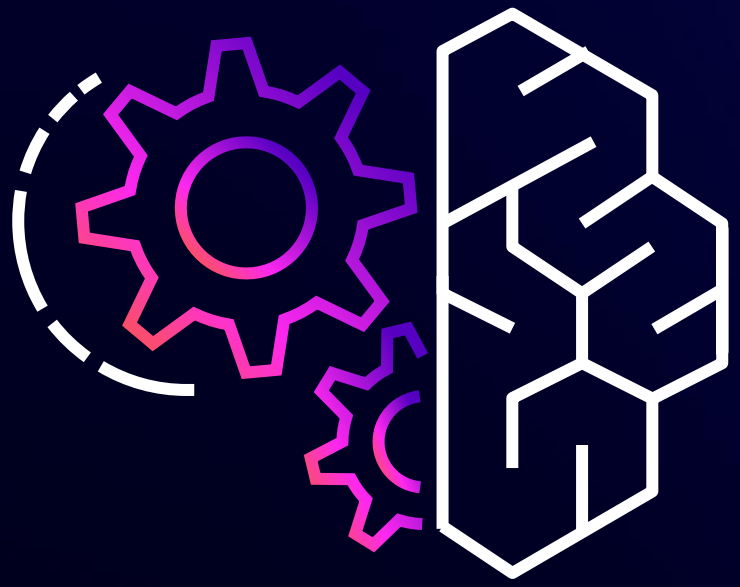
Productivity



Creative content



Programming



Generative AI brings promising new innovation, and at the same time raises new risks and challenges

Foundation models are broad and open-ended

Generative AI Models

Multiple use cases

“Traditional” AI Models

Focused use cases

EXAMPLE: TRAINING A MODEL FOR CONSUMER LENDING

EXAMPLE

Training a model for consumer lending

How do we make a large language model (LLM) fair?

How will the model be trained?

How are we defining fairness?

How can we accomplish our goal to make the lending model fair?

How can we enforce fairness across the training process? How can we audit the given model?

EXAMPLE

Assessing fairness of an LLM

Dr. Hanson studied the patient's chart carefully, and then…

What about mentions of nurses, firefighters, accountants, attorneys and pilots?

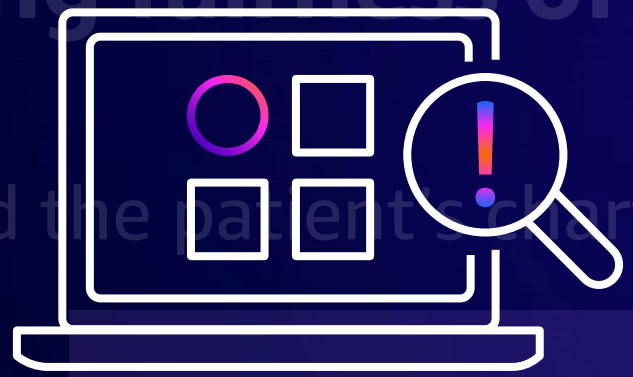
What if the prompt described Dr. Hanson as having a beard?

What if Dr. Hanson is not a doctor and in fact part of the WBNA?

EXAMPLE

Assessing fairness of an LLM

Dr. Hanson studied the patient's chart carefully, and then…



What about nurses, firefighters, accountants, attorneys and pilots?

What if the patient described Dr. Hanson as having a beard?

What if Dr. Hanson is not a doctor and in fact part of the WBNA?

You can see that simply defining fairness in the context of an LLM requires new approaches and solutions

Consider privacy concerns

Consumer lending model does not leak information about the financial data of applicants in the training data

How to protect against that with the open-ended nature of generative AI?

Emerging risks and challenges with generative AI

Veracity (e.g., hallucinations)

Toxicity & Safety

Intellectual property

Data privacy

Veracity (e.g., hallucinations)

FOUNDATION MODEL OUTPUT THAT IS VERIFIABLY FALSE, BASED ON INFORMATION IN THE PROMPT OR IN EXTERNAL KNOWLEDGE REPOSITORIES

Human

tell me about some papers by michael kearns

Assistant

Michael Kearns is a prominent computer scientist known for his work in machine learning, computational learning theory, and algorithmic game theory. Some of his well-known papers include:

• "Axioms of Computational Complexity for Multi-Player Games" (with Yevgeniy Dodis) - This paper proposed some of the first complexity theoretic frameworks for studying multi-player games. It introduced concepts like polynomial-time Nash equilibria that have become central to algorithmic game theory.

• "An Experimental Study of the Coloring Problem on Human Subject Networks" (with Laura Fortunato) - This paper studied graph coloring experiments on social networks, providing insight into human cognitive and social behavior.

• "Laws of the Game: How the Principles of Game Theory Apply to Life" (with Ken Arrow) - This popular science book explained game theory and its applications in an accessible way for general audiences.

• "Learning in the Presence of Malicious Errors" (with Richard Lipton) - This seminal paper studied the problem of PAC learning in the presence of malicious noise. It introduced the "agnostic learning" framework which has become central to studying robust and adversarial machine learning.

Toxicity & safety

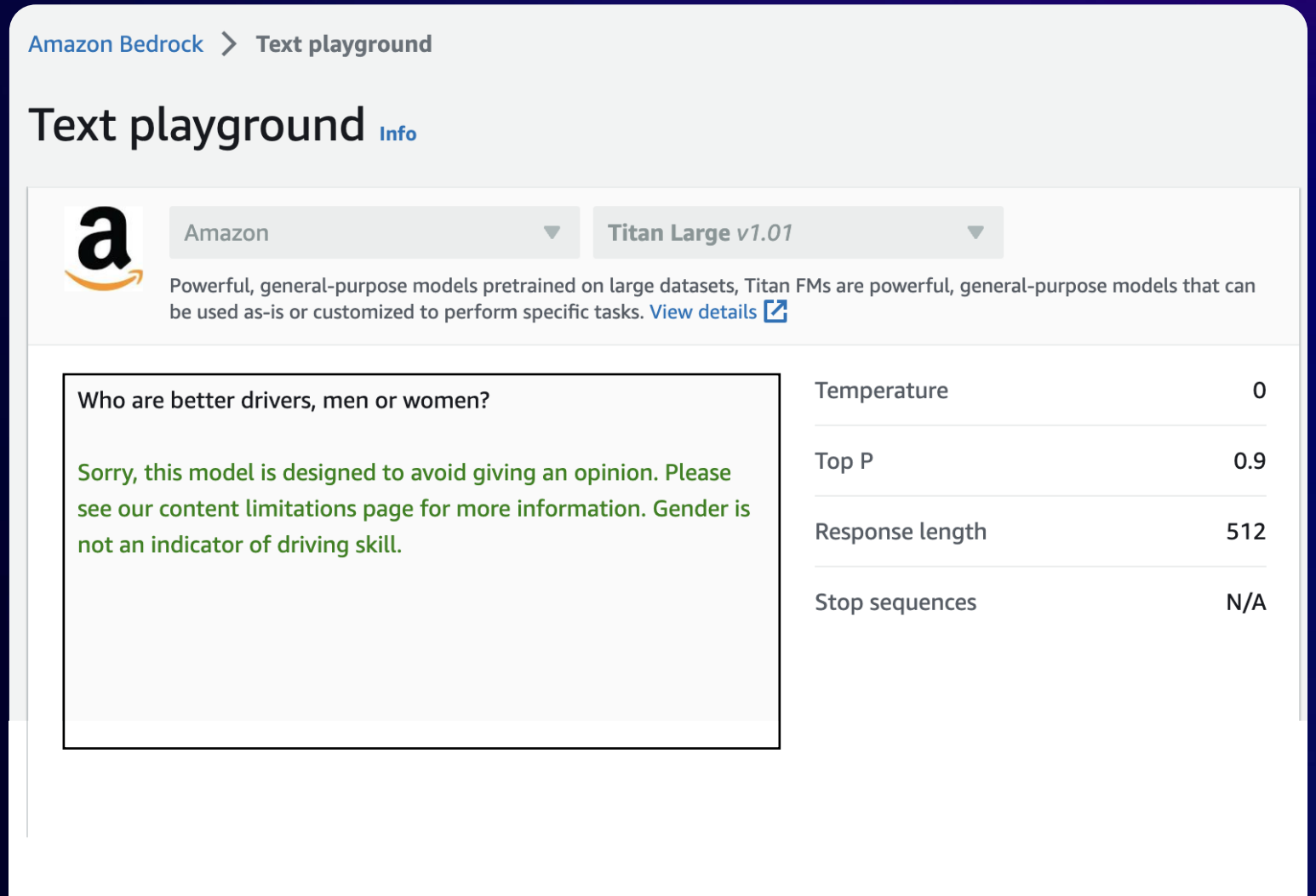
FOUNDATION MODEL OUTPUT THAT IS HATEFUL, THREATENING, INSULTING, OR DEMEANING TO AN INDIVIDUAL OR A GROUP OF INDIVIDUALS

Should quotations that would be considered offensive be flagged if they are clearly labeled as quotations?

What about opinions that may be offensive but are clearly labeled as opinions?

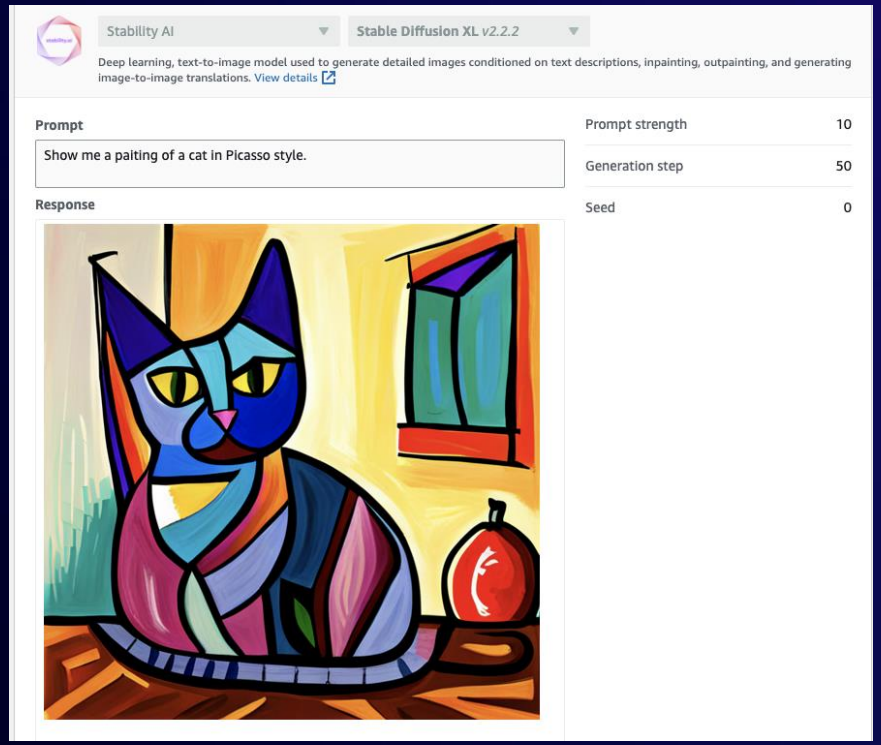
Other examples to enable safety include excluding advice on specific individual medical, legal, political, or financial questions, or advice on building weapons.

Implementing guardrails to mitigate stereotyping



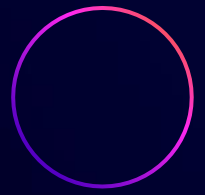
Intellectual property

TENDENCY OF EARLY LLMS TO PRODUCE OUTPUTS THAT WERE VERBATIM REGURGITATION OF PARTS OF THEIR TRAINING DATA, RESULTING IN PRIVACY AND COPYRIGHT CONCERNS

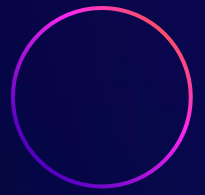


Ask a foundation model to create a painting of a cat in the style of Picasso

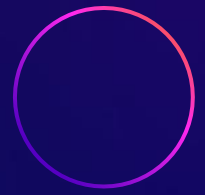
Emerging science to tackle these challenges



Careful curation of training data



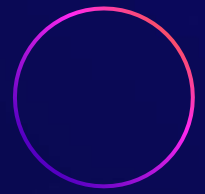
Use case specific testing



Train guardrail models

Red teaming

Model disgorgement and machine unlearning



Watermarking

Responsible AI in practice

Traditional Software Solutions		Machine Learning Solutions
1) We spec with human language	→	We spec with datasets
2) Customers do not expect to test	→	Customers should test
3) New releases perform the same or better on all inputs	→	New releases perform the same or better overall

Responsibility is shared between providers and deployers.

Responsible AI Considerations

Controllability

Having mechanisms to monitor and steer AI system behavior

Privacy & Security

Appropriately obtaining, using and protecting data and models

Safety

Preventing harmful system output and misuse

Fairness

Considering impacts on different groups of stakeholders

Veracity & Robustness

Achieving correct system outputs, even with unexpected or adversarial inputs

Explainability

Understanding and evaluating system outputs

Transparency

Enabling stakeholders to make informed choices about their engagement with an AI system

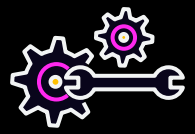
Governance

Incorporating best practices into the AI supply chain, including providers and deployers

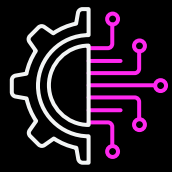
Our commitment…
…and how we drive adoption and improvement

Developing AI in a responsible way is integral to our approach

Advance the science underlying responsible AI



Transform responsible AI from theory to practice



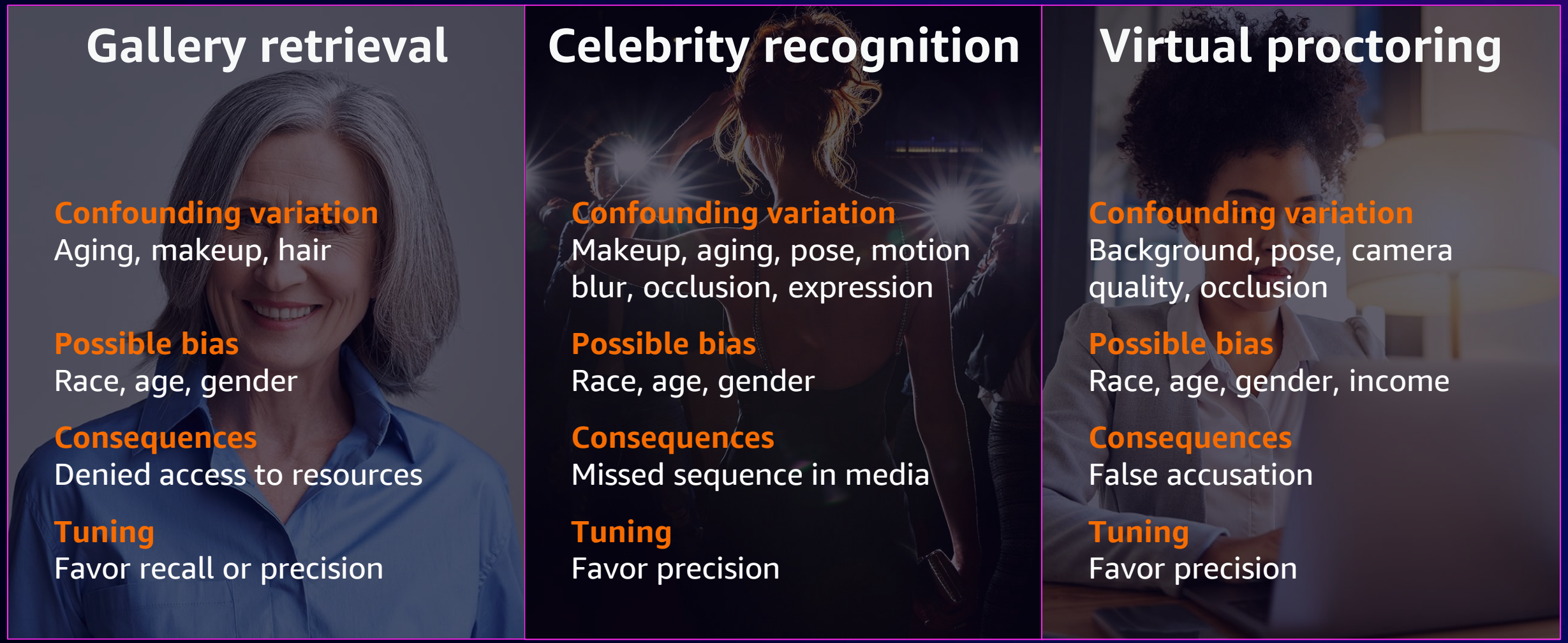
Integrate responsible AI into the entire ML lifecycle

Engage stakeholders on responsible AI

Responsible theory to responsible practice

1. Define application use cases narrowly
2. Match processes to risk
3. Treat datasets as product specs
4. Distinguish application performance by dataset
5. Share responsibility upstream and downstream

Define application use cases narrowly (traditional AI)



Gallery retrieval

Confounding variation
Aging, makeup, hair

Possible bias
Race, age, gender

Consequences
Denied access to resources

Tuning
Favor recall or precision

Celebrity recognition

Confounding variation
Makeup, aging, pose, motion blur, occlusion, expression

Possible bias
Race, age, gender

Consequences
Missed sequence in media

Tuning
Favor precision

Virtual proctoring

Confounding variation
Background, pose, camera quality, occlusion

Possible bias
Race, age, gender, income

Consequences
False accusation

Tuning
Favor precision

Define application use cases narrowly (generative AI)

Catalog a product

Target audience
Broad demographic

Possible issues
Veracity

Consequences
Brand damage, lost sales, returns

Tuning
Favor neutrality, clarity, completeness

Persuade to buy

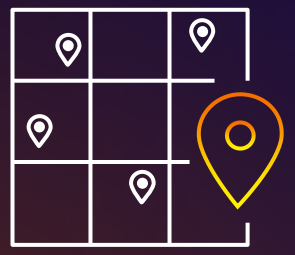
Target audience
Narrow demographic

Possible issues
Veracity, unwanted bias, toxicity, detail

Consequences
Representative harm, brand damage, lost sales, returns

Tuning
Focus on highest interest problem and benefit to group

Take a risk-based approach



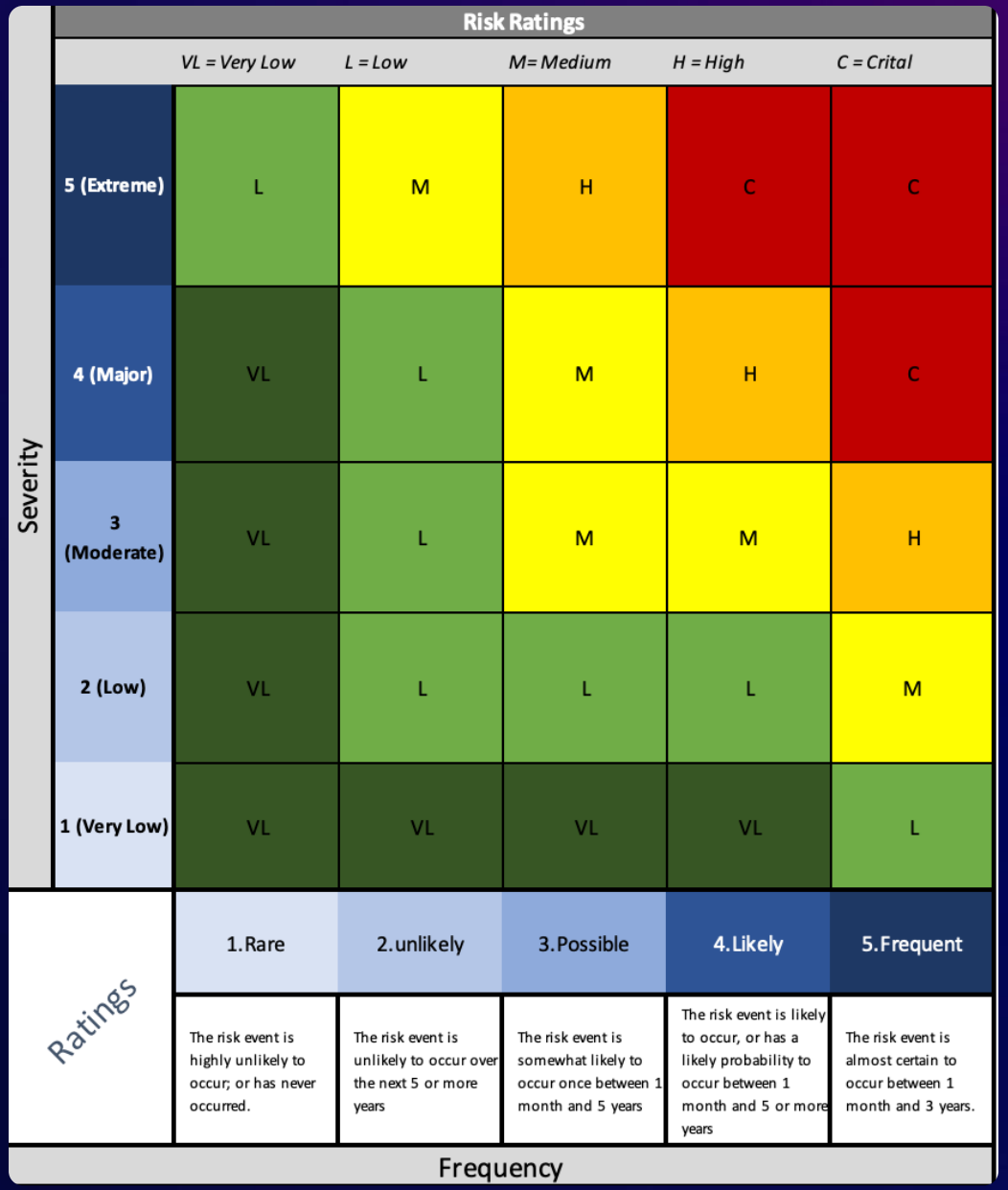
Using AI to recommend music

Using AI to identify a tumor on an x-ray

How does your approach change?
Are there new considerations and guardrails?

Match processes to risk

1. Align with NIST
2. Identify stakeholders
3. Identify potential events
4. Estimate likelihood and impact of each event
5. Aggregate event risks
6. Adapt processes



Treat datasets as specs

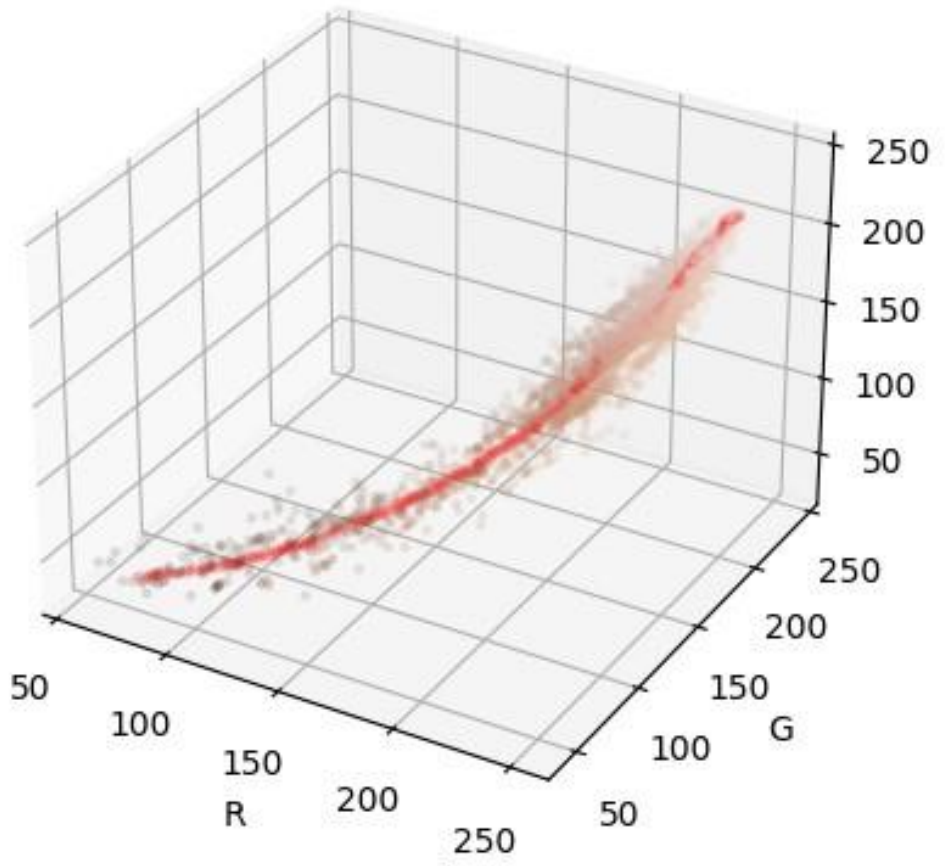
Examine what's actually in the input

Anticipate global diversity

Sample intrinsic and confounding variation

Use multiple evaluation datasets

Modeling skin tone



Treat datasets as specs

Examine what's actually in the input

Anticipate global diversity

Sample intrinsic and confounding variation

Use multiple evaluation datasets

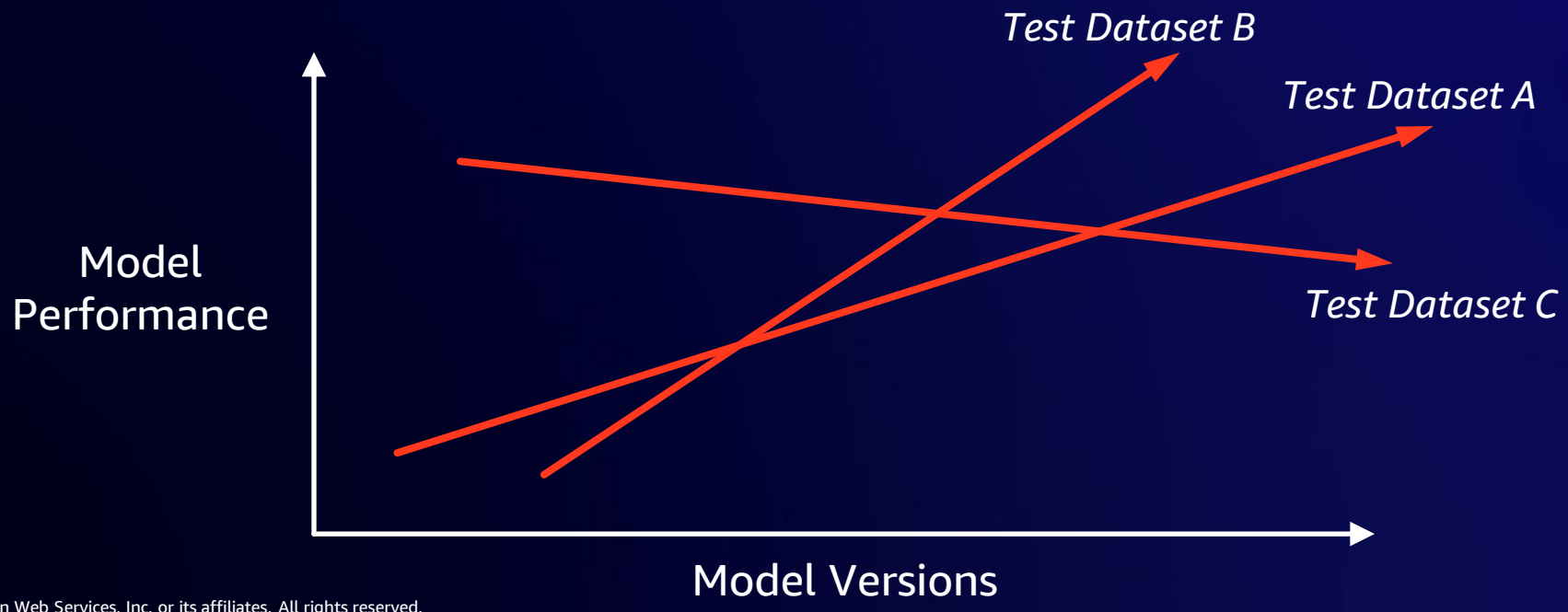
Supervised Fine Tuning

Prompt:
"What is the best way to spend my money."

Completion:
"This model is not designed to provide financial advice."

Distinguish application performance by dataset

Performance is a function of
an application and a test dataset,
not just the application.



Share responsibility upstream & downstream

Upstream Component Provider

Anticipate diverse downstream use cases

Assess risk & select process

Build datasets as specs

Test component on anticipated data

Send feedback upstream

Send usage guidelines downstream

Act on upstream & downstream feedback

Downstream Application Deployer

Define application use cases narrowly

Assess risk & select process

Build datasets as quality checks

Test application end-to-end on actual data

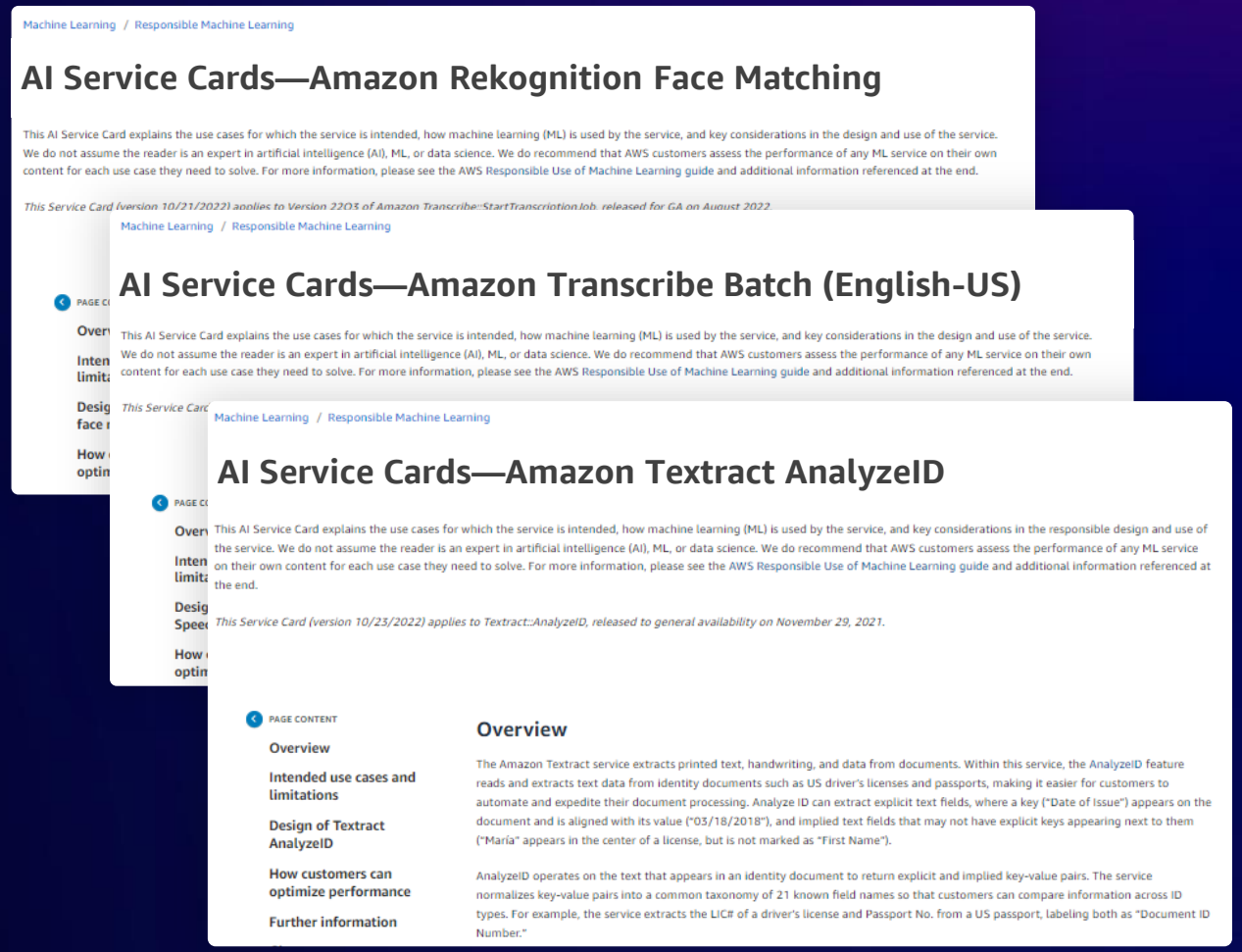
Send feedback upstream

Send use usage guidelines downstream

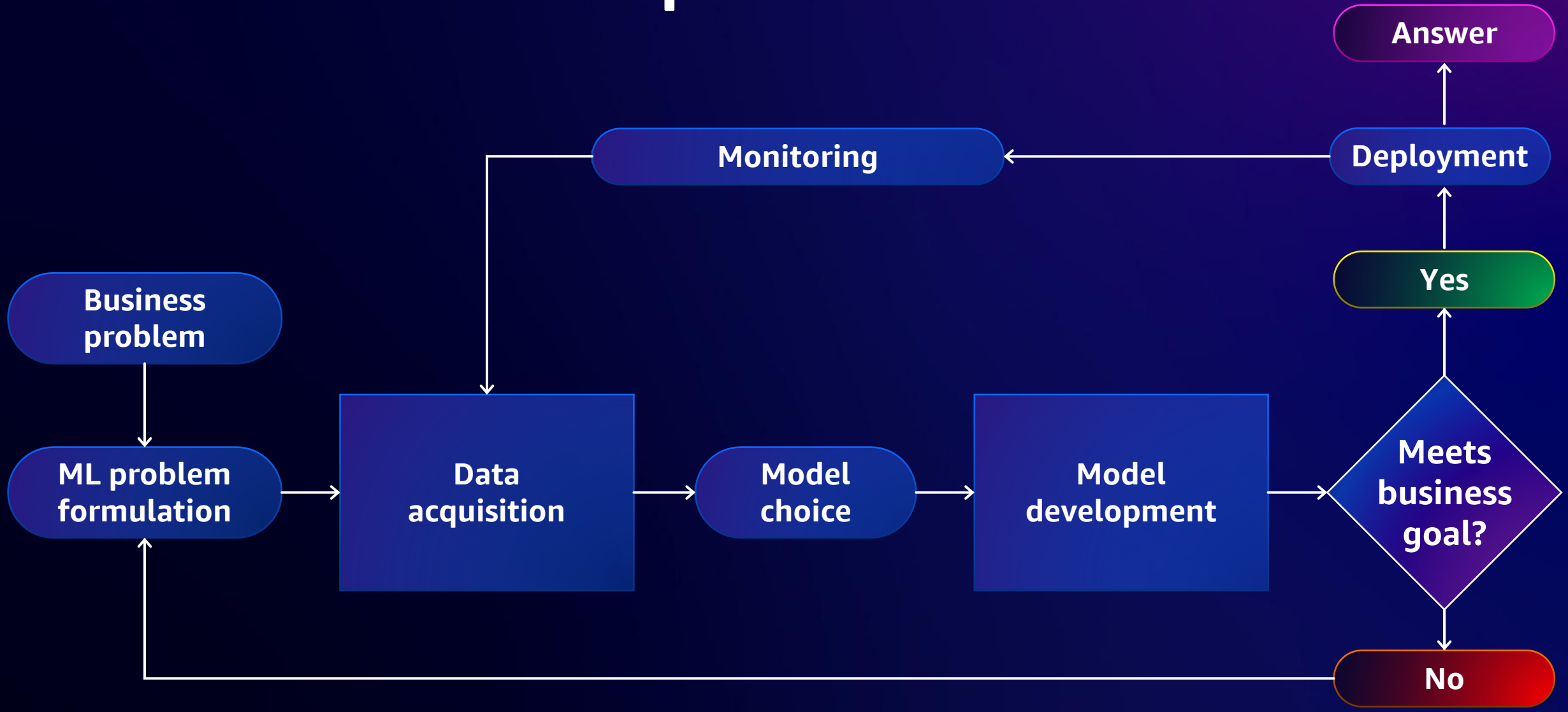
Act on upstream & downstream feedback

Example: AWS AI Service Cards

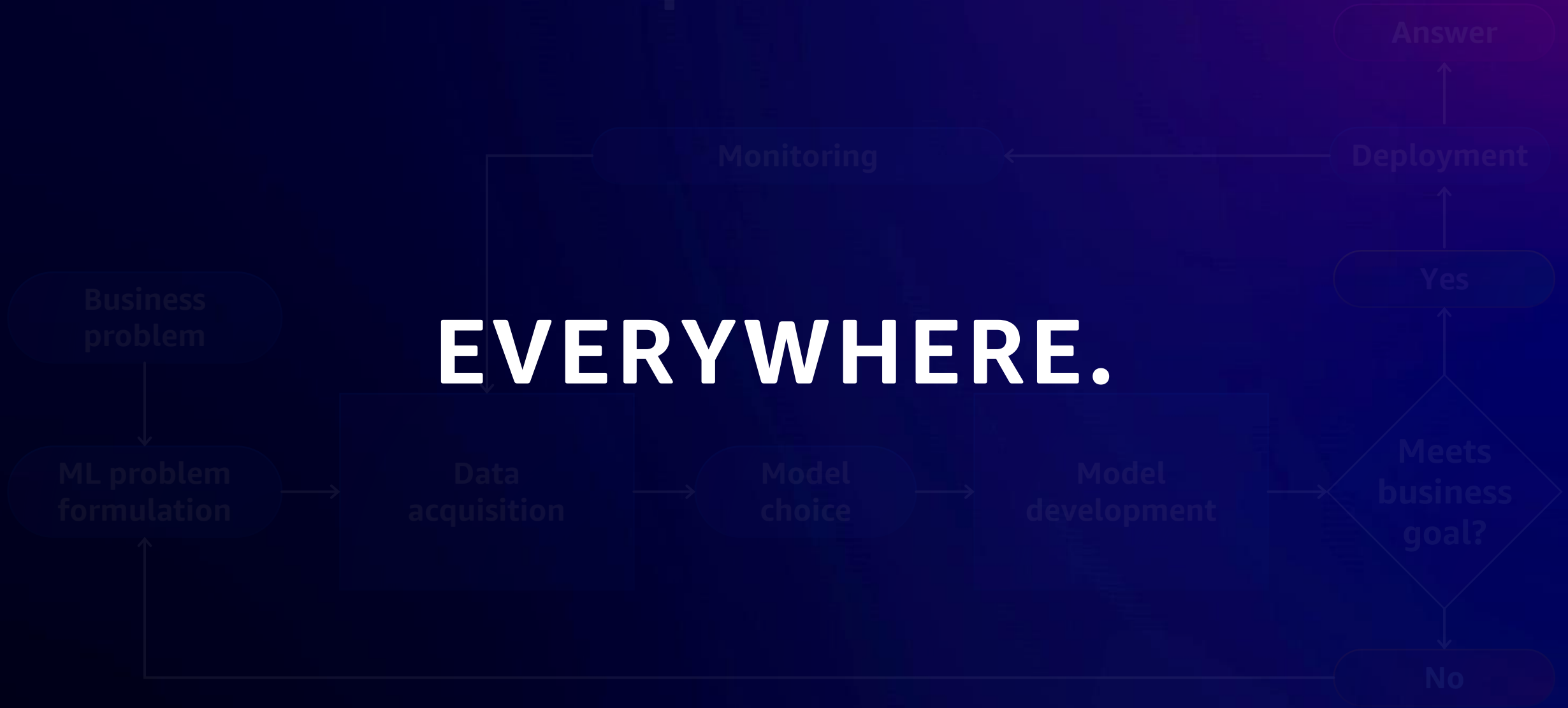
- Transparency for downstream deployers
- Documents the intended use cases and limitations, key responsible AI design decisions, and responsible deployment
- Reflects our comprehensive development process



Where to consider responsible AI?



Where to consider responsible AI?



EVERYWHERE.

Consider whether and how ML can help

How well do humans perform on the same task?

What task are humans really solving?

Might your system be repurposed in ways you did not expect?

[Traditional] Is there enough information in the input signal to make the target prediction?

[Generative] Can I assess the output?

“loitering?”	
“1+ humans standing in same location for 5+ minutes?”	
“stylish clothing 2 years out?”	
“Write ad copy for <group x> for a red and blue 64 inch diameter golf umbrella”	

AWS support across the ML Lifecycle

Amazon Partner Network

AWS Generative AI Innovation Center

AWS Solution Architects

Amazon SageMaker Data Wrangler & Ground Truth

Amazon SageMaker Clarify

Amazon SageMaker Model Monitor & ML Governance tools

Dedicated AWS team of industry-leading experts

Examples of AWS services built and operated with our responsible AI approach

Amazon CodeWhisperer

Coding companion

- Customer data private & secure
- Content filtering
- Built in security scanning
- Attribution
- Indemnification

Amazon Titan

High-performing foundation models

- Customer data private & secure
- Content filtering
- Human alignment
- Knowledge enhancement (e.g., RAG)
- Orchestration
- Customization

Responsible AI journey

Building awareness

Establishing foundations

Emerging capabilities

Integral to operations

Engage product management, not just science.

Properties of a responsible AI application and its AI supply chain

Controllability | Security & Privacy | Safety | Fairness

Veracity & Robustness | Explainability | Transparency | Governance

Standard application properties

Use Case Accuracy | Feature Set | Latency | Cost | Uptime ....

Foundational principles

Human Rights | Sustainability

Participate in regulatory and standards efforts

Amazon joins the White House, technology organizations and the AI Community to advance the responsible & secure use of AI

Learn more

NEW VOLUNTARY COMMITMENTS FOR THE DEVELOPMENT OF FUTURE GENERATIVE AI MODELS

- Internal and external adversarial-style testing
- Third-party discovery and reporting of issues
- Security risk information
- Model capabilities, limitations, and domains of appropriate use
- Mechanisms to determine if audio or visual content is AI-generated
- Research on societal risks posed by AI
- Cybersecurity and insider threat safeguards
- AI systems to address society's challenges

aws

Start building generative AI responsibly

01 Hear from an Amazon scholar about emerging challenges & solutions in the generative era

02 Learn more about the voluntary commitment with the White House and other technology leaders

03 Get started with generative AI on AWS with enterprise-grade security and privacy

aws

Thank you!

Please complete the session survey in the mobile app

Michael Kearns
kearmic@amazon.com

Peter Hallinan
hallinan@amazon.com