AWS re:Invent

DECEMBER 2 – 6, 2024 | LAS VEGAS, NV

AIM392-NEW

Responsible AI development and deployment

Rahul Gupta
(he/him/his)
Sr. Manager of Science
Amazon AGI

Sherif Mostafa
(he/him/his)
Sr. Manager of Product
Amazon AGI

Karan Bhandarkar
(he/him/his)
Principal Product Manager
Amazon AGI

Agenda

01 Introduction to responsible AI

02 Working backwards from design objectives

03 Aligning the AI system

04 Evaluating the extent of alignment

05 Looking ahead

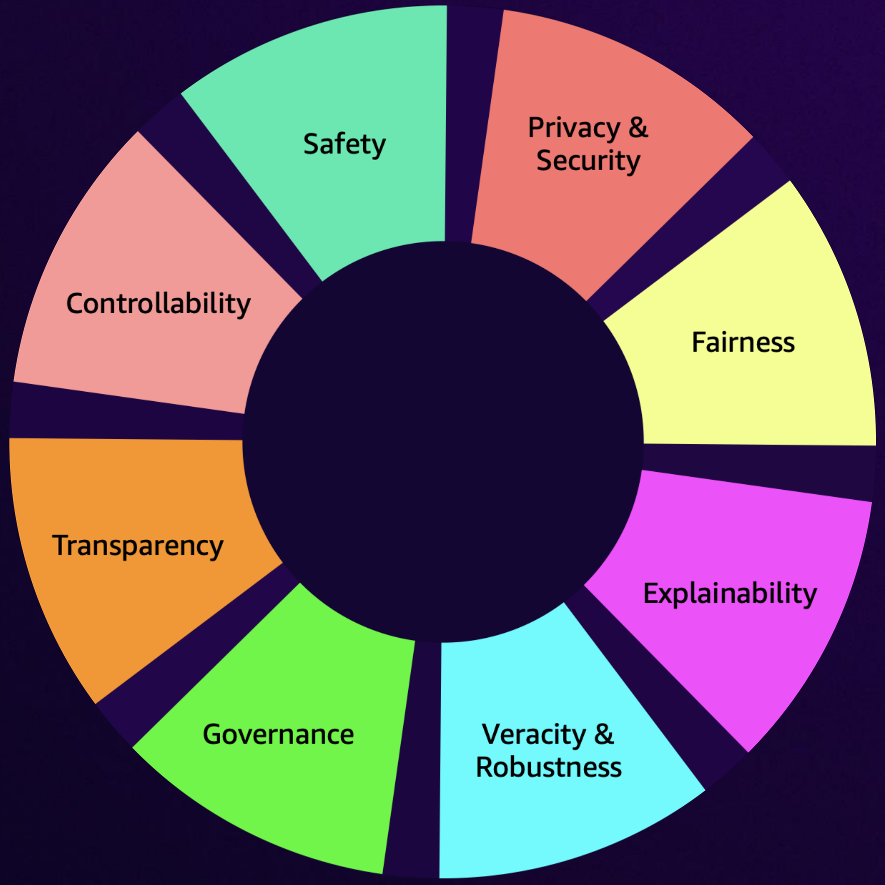
Introduction to responsible AI

Introduction to responsible AI

What is responsible AI?

Why is it important?

What are the core dimensions of responsible AI for Amazon Nova?



Core dimensions of responsible AI

Safety
Preventing harmful system output and misuse

Privacy and security
Appropriately obtaining, using, and protecting data and models

Controllability
Having mechanisms to monitor and steer AI system behavior

Veracity and robustness
Achieving correct system outputs, even with unexpected or adversarial inputs

Fairness
Considering impacts on different groups of stakeholders

Explainability
Understanding and evaluating system outputs

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

Transparency
Helping stakeholders make informed choices about their engagement with an AI system

Integrating the dimensions

To integrate these dimensions in the development of the Amazon Nova FMs, follow this three step process along the model lifecycle:

Step 1: Design objectives
Define the bar to uphold the AI system

Step 2: System alignment
Incorporate measures during the build process to meet this bar

Step 3: System evaluation
Evaluate the extent of alignment to ensure the system meets the predefined bar

Working backwards from design objectives

Working backwards from design objectives

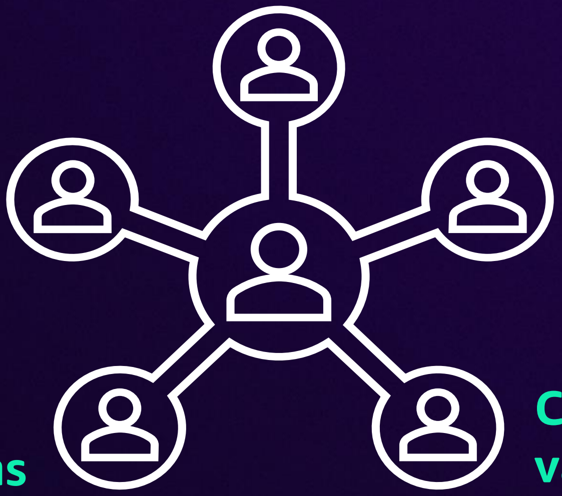
What are design objectives? Why are they important?

How are design objectives defined?

Are all objectives equal?

Foundational and application

Legislation, regulations and voluntary commitments



Privacy and security

Industry associations

Customer trust and expectations

Core societal values

Aligning the AI system

Aligning the AI system

Data curation

Curating data as per the design objectives

Model training

Instilling the design objectives in the core model

Auxiliary system

Supplementing the core model

Aligning the AI system

Data curation

Curating data as per the design objectives

- Collect a diverse mix of data
- Moderate for toxic and unsafe content
- Apply privacy protecting filters

Aligning the AI system

Model training

Instilling the design objectives in the core model

- Pre-training
- Supervised fine-tuning (SFT)
- Learning from human feedback (LHF)

Aligning the AI system

Auxiliary system

Supplement the core model

- Input content moderation
- Output content moderation
- AI transparency markers
- Hotfix mechanism

Technical challenges faced

- Optimizing the model for adherence without impacting performance
- Solving for a thick long tail of vulnerabilities
- Staying abreast of novel vulnerabilities
- Ensuring low latency with auxiliary guardrails

Evaluating the extent of alignment

Evaluating the extent of alignment

ROLE OF EVALUATION

Test for alignment

Use automated and human-in-the-loop mechanisms to test the AI system's adherence to each objective

Assess severity

Identify areas where the model misses the design objectives, and assign a severity rating to a deemed misalignment

Address misalignment

Address the identified misalignment through iterations of the model or the auxiliary system

Evaluating the extent of alignment

APPROACH TO EVALUATION

Quantitative

Testing for previously known issues at scale

Qualitative

Testing for unknown and novel vulnerabilities ("red teaming")

Automated

Testing for new findings at scale

Evaluating the extent of alignment

Quantitative

Testing for previously known issues at scale

- Positive and negative missed objective testing
- Public benchmarks
- Internally curated test sets
- Design objective classifiers

Evaluating the extent of alignment

Qualitative

Testing for unknown and novel vulnerabilities (red teaming)

- What is red teaming?
- Tiered approach to red teaming

Evaluating the extent of alignment

Qualitative

Testing for unknown and
novel vulnerabilities
(red teaming)

- What is red teaming?
- Tiered approach to red teaming
  - **Internal red teamers**
  - **Partner red teamers**
  - **Highly specialized red teamers**
  - **Subject matter experts**

Evaluating the extent of alignment

Automated

Testing for new findings at scale

- Automated red teaming
- Published jailbreak techniques

Looking ahead

Looking ahead

APPROACH AND COMMITMENT TO RESPONSIBLE SCALING POLICY

- As capabilities of AI systems expand, so do potential risks they present
- We are defining and operationalizing against these risks as AI systems scale in their capabilities
- Ongoing efforts include:
  - Voluntary RAI commitments, including to the White House and others
  - Engagements with Frontier Model Forum (FMF), National Institute of Standards and Technology (NIST) U.S. AI Safety Institute, Model Evaluation and Threat Research (METR)

Looking ahead

EFFORTS TO ENHANCE TRANSPARENCY FOR DIGITAL CONTENT

- Engagement with Partnership on AI (PAI)
- Steering committee member of Coalition of Content Provenance and Authenticity (C2PA)

Looking ahead

ACADEMIC PARTNERSHIPS

- Amazon Trusted AI Challenge – Security in coding LLMs
- Expanding evaluations with UC Berkeley and Carnegie Mellon University

Looking ahead

APPROACH TOWARD THE 2025 AI SUMMIT IN FRANCE

- Presenting Amazon Safe AI Scaling Framework

Questions?

Thank you!

Please complete the session survey in the mobile app

Rahul Gupta
Email: gupra@amazon.com
LinkedIn: @rahul-gupta-amazonagi

Sherif Mostafa
Email: smmostaf@amazon.com
LinkedIn: @sherif-s-mostafa

Karan Bhandarkar
Email: karanbx@amazon.com
LinkedIn: @karanbhandarkar