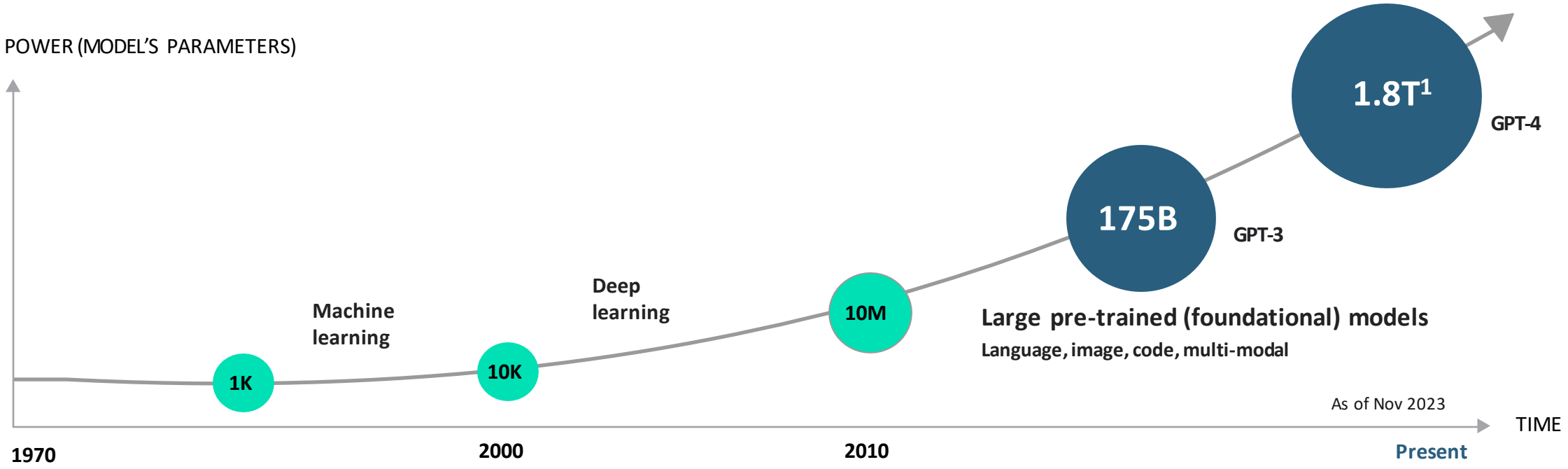
ETHICS IN AI

MARIA GOMEZ AGUIRRE
PRIYANKA SYAL

LeadDev Berlin - Dec 5th 2023

LLM IS A NEW PARADIGM FOR BUILDING AI SYSTEMS



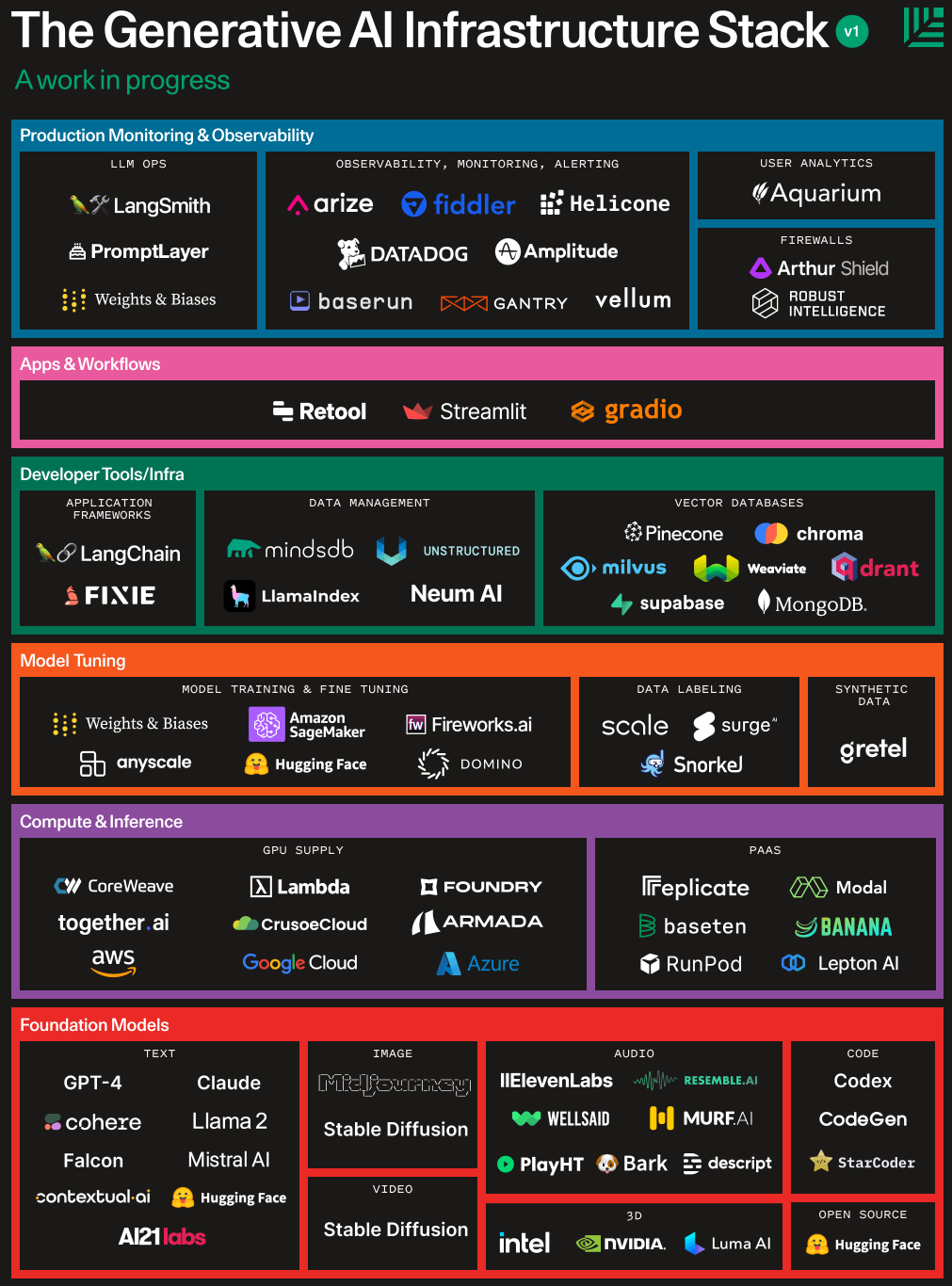
1. https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/

GENAI HAS WITNESSED UNPRECEDENTED LEVELS OF ADOPTION BY CONSUMERS

~ 5 DAYS

~ 75 DAYS

~ 150 DAYS



ECOSYSTEM IS COMPLEX AND GROWING

Source: https://www.sequoiacap.com/article/generative-ai-act-two/

AI HAS BEEN USED WITH GREAT SUCCESS …

economictimes.indiatimes.com

Generative AI in healthcare: Seize the advantage or fall behind!

The GenAI market is projected to surge to an astronomical $120 billion by 2030, with its most exciting impact anticipated in the healthcare industry. GenAI is set to fundamentally transform healthcare delivery, and early indicators of its at-scale adopt...

about.fb.com

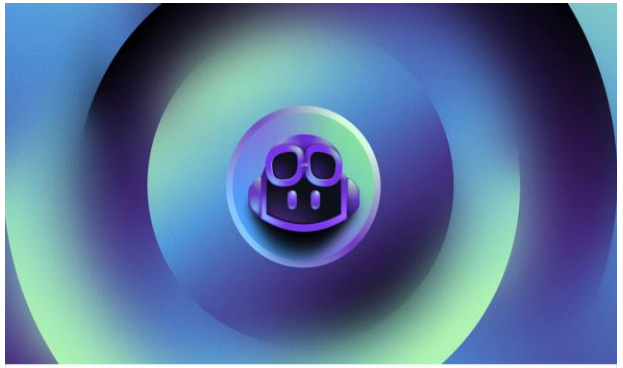
Introducing SeamlessM4T, a Multimodal AI Model for Speech and Text Translations | Meta

SeamlessM4T allows people to communicate effortlessly through speech and text across different languages.

www.forbes.com

Council Post: The Future Of AI-Powered Personalization: The Potential Of Choices

In the ever-expanding digital landscape, where choices seem limitless, AI-powered personalization has emerged as a game-changer.



github.blog

GitHub Copilot X: The AI-powered developer experience

GitHub Copilot is evolving to bring chat and voice interfaces, support pull requests, answer questions, and adopt OpenAI's GPT-4.

AI HAS BEEN USED WITH GREAT SUCCESS …

economictimes.indiatimes.com

Generative AI in healthcare: Seize the advantage or fall behind!

The GenAI market is projected to surge to an astronomical $120 billion by 2030, with its most exciting impact anticipated in the healthcare industry. GenAI is set to fundamentally transform healthcare delivery, and early indicators of its at-scale adopt...

about.fb.com

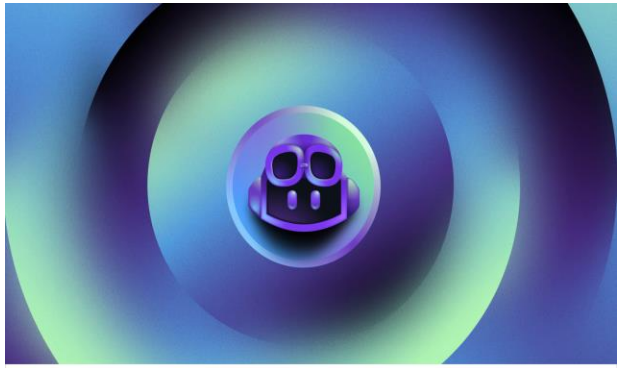
Introducing SeamlessM4T, a Multimodal AI Model for Speech and Text Translations | Meta

SeamlessM4T allows people to communicate effortlessly through speech and text across different languages.

www.forbes.com

Council Post: The Future Of AI-Powered Personalization: The Potential Of Choices

In the ever-expanding digital landscape, where choices seem limitless, AI-powered personalization has emerged as a game-changer.



github.blog

GitHub Copilot X: The AI-powered developer experience

GitHub Copilot is evolving to bring chat and voice interfaces, support pull requests, answer questions, and adopt OpenAI's GPT-4.

… BUT IT CAN ALSO HAVE A NEGATIVE IMPACT

www.pbs.org

U.S. lawmakers question Meta and X over AI-generated political deepfakes ahead of 2024 election

US Sen. Amy Klobuchar of Minnesota and U.S. Rep. Yvette Clarke of New York sent a letter Thursday to Meta CEO Mark Zuckerberg and X CEO Linda Yaccarino expressing "serious concerns" about the emergence of AI-generated political ads on their platforms an...

apnews.com

'Game of Thrones' creator and other authors sue ChatGPT-maker OpenAI for copyright infringement

John Grisham, Jodi Picoult and George R.R. Martin are among 17 authors suing OpenAI for "systematic theft on a mass scale."

www.wired.co.uk

Hollywood Writers Reached an AI Deal That Will Rewrite History

A faction of scribes is putting guardrails around AI's encroachment on their work. The effects will echo in industries far beyond Hollywood.

news.mit.edu

Study finds gender and skin-type bias in commercial artificial-intelligence systems

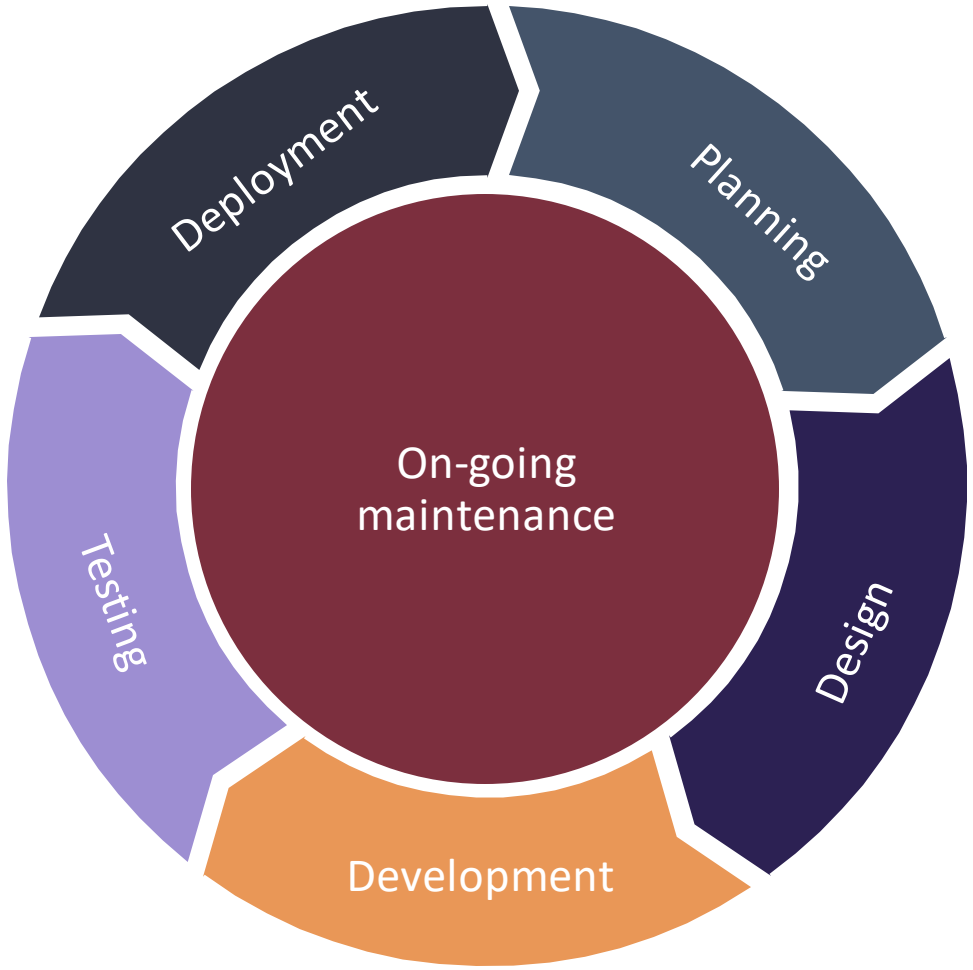
A new paper from the MIT Media Lab's Joy Buolamwini shows that three commercial facial-analysis programs demonstrate gender and skin-type biases, and suggests a new, more accurate method for evaluating the performance of such machine-learning syste...

Image source: DALL·E 3

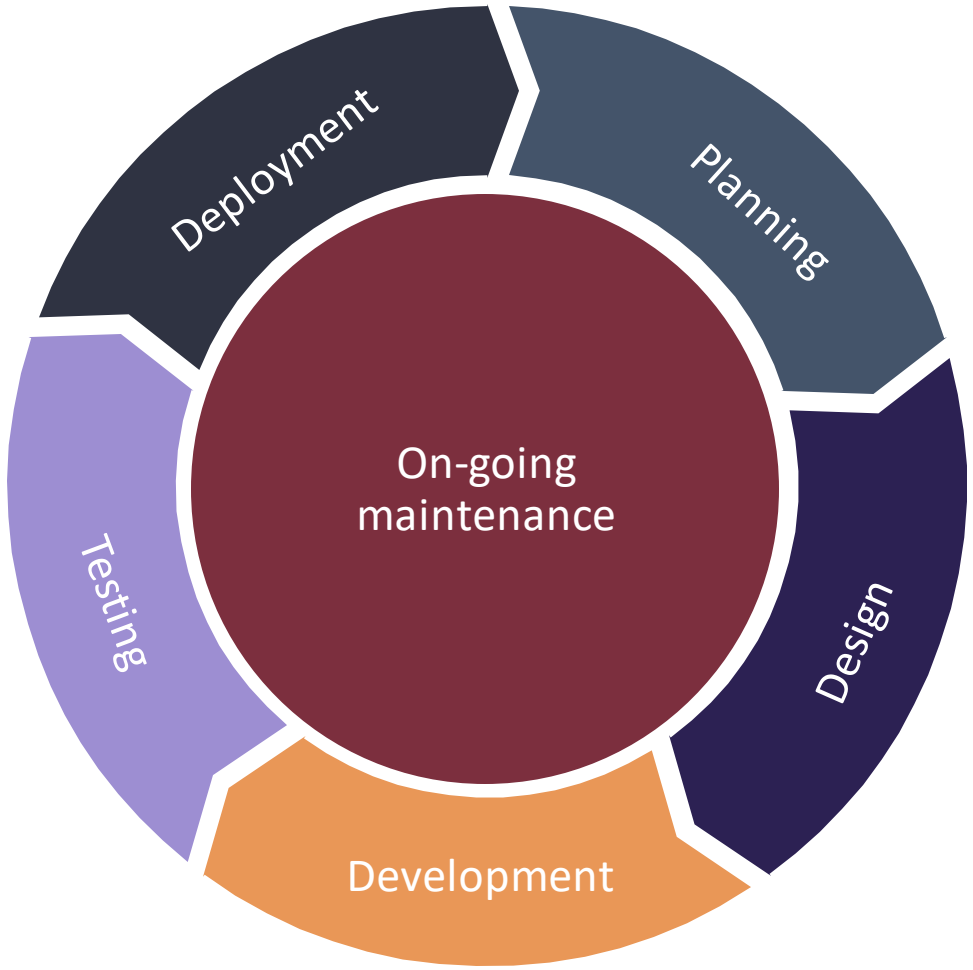
WHY SHOULD WE CARE?

- We are building more and more AI products
- Regulators are still trying to play catch-up
- We need to think about the human impact of our work as engineering community

THERE ARE TOOLS AND PRACTICES WE CAN APPLY IN EVERY STEP OF THE SDLC



THERE ARE TOOLS AND PRACTICES WE CAN APPLY IN EVERY STEP OF THE SDLC



PLANNING & REQUIREMENTS ANALYSIS

Ethical requirements

PLANNING & REQUIREMENTS ANALYSIS

Ethical requirements

New capabilities and skills

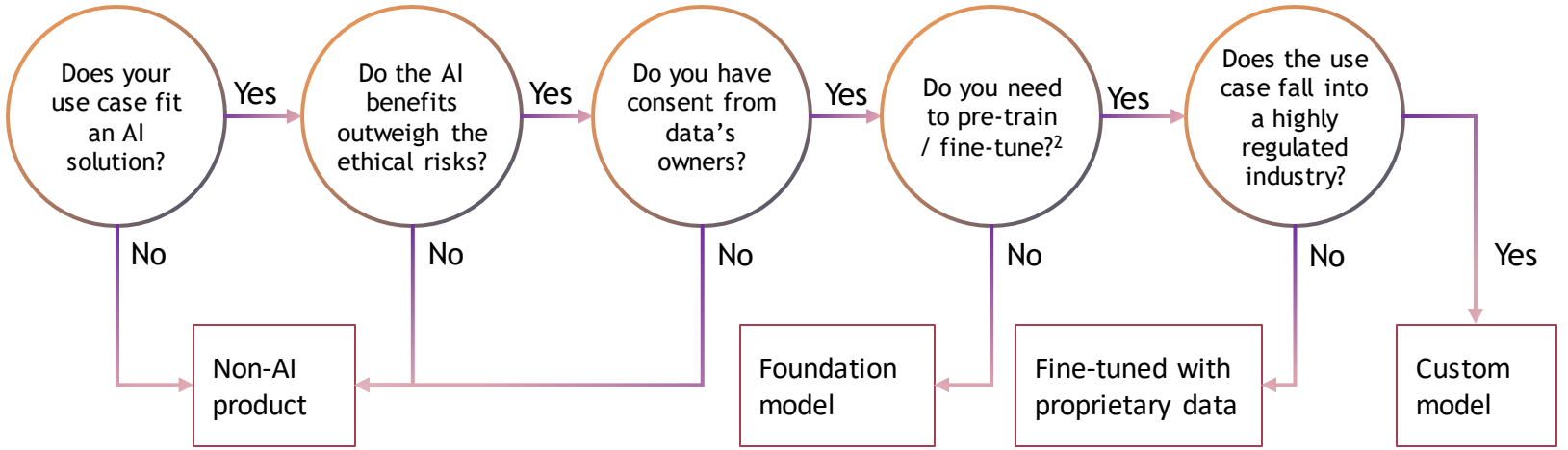
- Technical (ie MLOps, AgentOps)
- Non-technical (ie Social science, regulatory knowledge)

Image source: DALL·E 3

Is this FOMO?

PLANNING & REQUIREMENTS ANALYSIS

Is this FOMO?



DESIGN

Consider the disrupted

DESIGN

Consider the disrupted

Engage end users from the start

DESIGN

Consider the disrupted

Engage end users from the start

Design for explainability & transparency

DEVELOPMENT

Transparency

What's the minimum amount of data needed?
What is the new use of the user data?
How can users opt out?

Source: https://www.honeycomb.io/blog/llms-demand-observability-driven-development; https://github.com/NVIDIA/NeMo-Guardrails/blob/main/docs/README.md

DEVELOPMENT

Transparency

What's the minimum amount of data needed?
What is the new use of the user data?
How can users opt out?

Authenticity

Guardrails tools

- Azure AI Content Safety
- ChatGPT Question Filter
- NeMo Guardrails

Source: https://www.honeycomb.io/blog/llms-demand-observability-driven-development; https://github.com/NVIDIA/NeMo-Guardrails/blob/main/docs/README.md

DEVELOPMENT

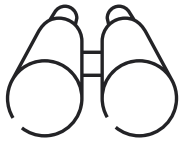
Transparency

What's the minimum amount of data needed?
What is the new use of the user data?
How can users opt out?

Authenticity

Guardrails tools

- Azure AI Content Safety
- ChatGPT Question Filter
- NeMo Guardrails



Observability

OpenTelemetry

LangChain
LangSmith

honeycomb.io

PromptLayer

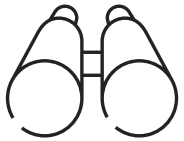
Weights & Biases

Source: https://www.honeycomb.io/blog/llms-demand-observability-driven-development; https://github.com/NVIDIA/NeMo-Guardrails/blob/main/docs/README.md

DEVELOPMENT

Transparency

Authenticity



Observability

Identify and mitigate limitations of models & datasets

Source: https://www.honeycomb.io/blog/llms-demand-observability-driven-development; https://github.com/NVIDIA/NeMo-Guardrails/blob/main/docs/README.md

TESTING

Bias and authenticity

Aequitas
Bias & Fairness Audit

PAIR

AIF360

Source: https://blog.google/technology/ai/google-responsible-generative-ai-best-practices/; https://owasp.org/www-project-top-10-for-large-language-model-applications/

TESTING

Bias and authenticity

Correctness

Aequitas
Bias & Fairness Audit

PAIR

AIF360

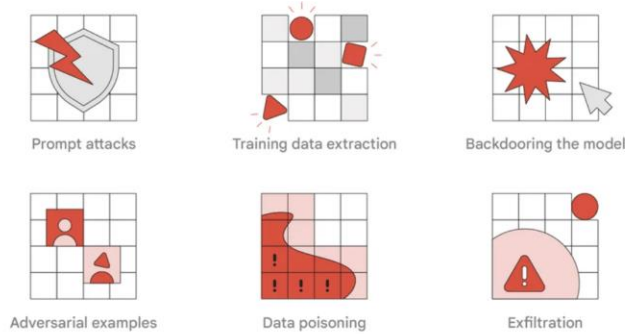
DeepEval.

Source: https://blog.google/technology/ai/google-responsible-generative-ai-best-practices/; https://owasp.org/www-project-top-10-for-large-language-model-applications/

TESTING

Bias and authenticity

Correctness



Robustness

Aequitas
Bias & Fairness Audit

PAIR

AIF360

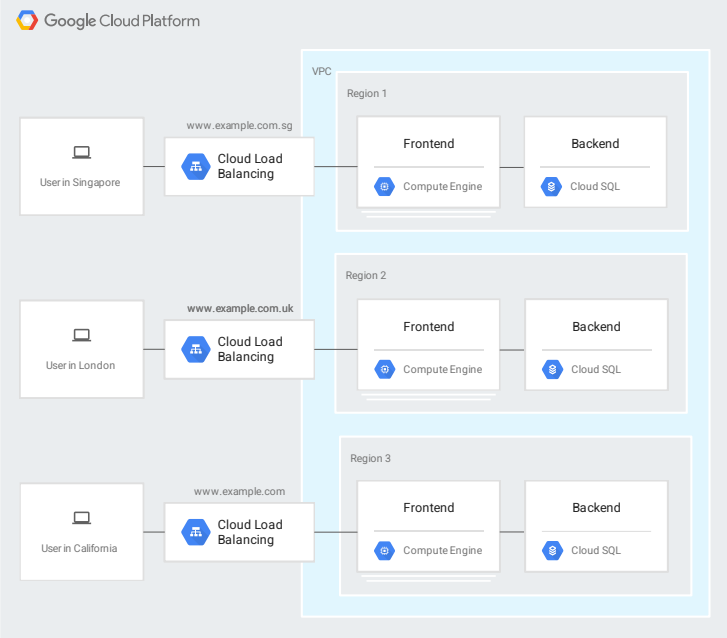
DeepEval.

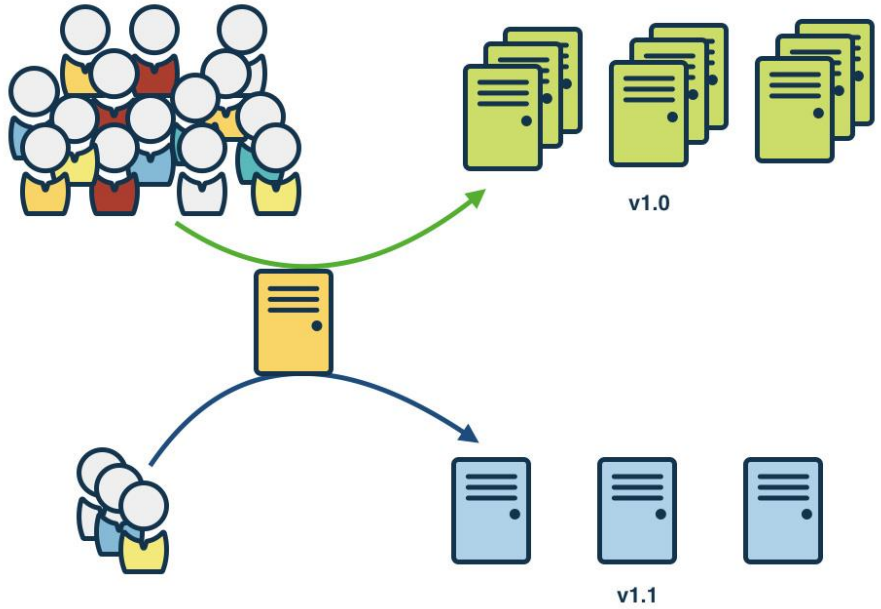
OWASP®

Adversarial Robustness Toolbox

Source: https://blog.google/technology/ai/google-responsible-generative-ai-best-practices/; https://owasp.org/www-project-top-10-for-large-language-model-applications/

DEPLOYMENT & ROLL OUT





Security and data privacy best practices

Staged rollouts

Source: https://www.gocd.org/2017/08/15/canary-releases/, https://cloud.google.com/solutions/best-practices-compute-engine-region-selection

MAINTENANCE

Align

Measure

Iterate

MAINTENANCE

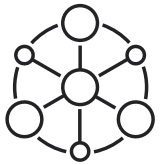
Measure

Model quality

Authenticity / Correctness Rate

Safety Score

Biased responses

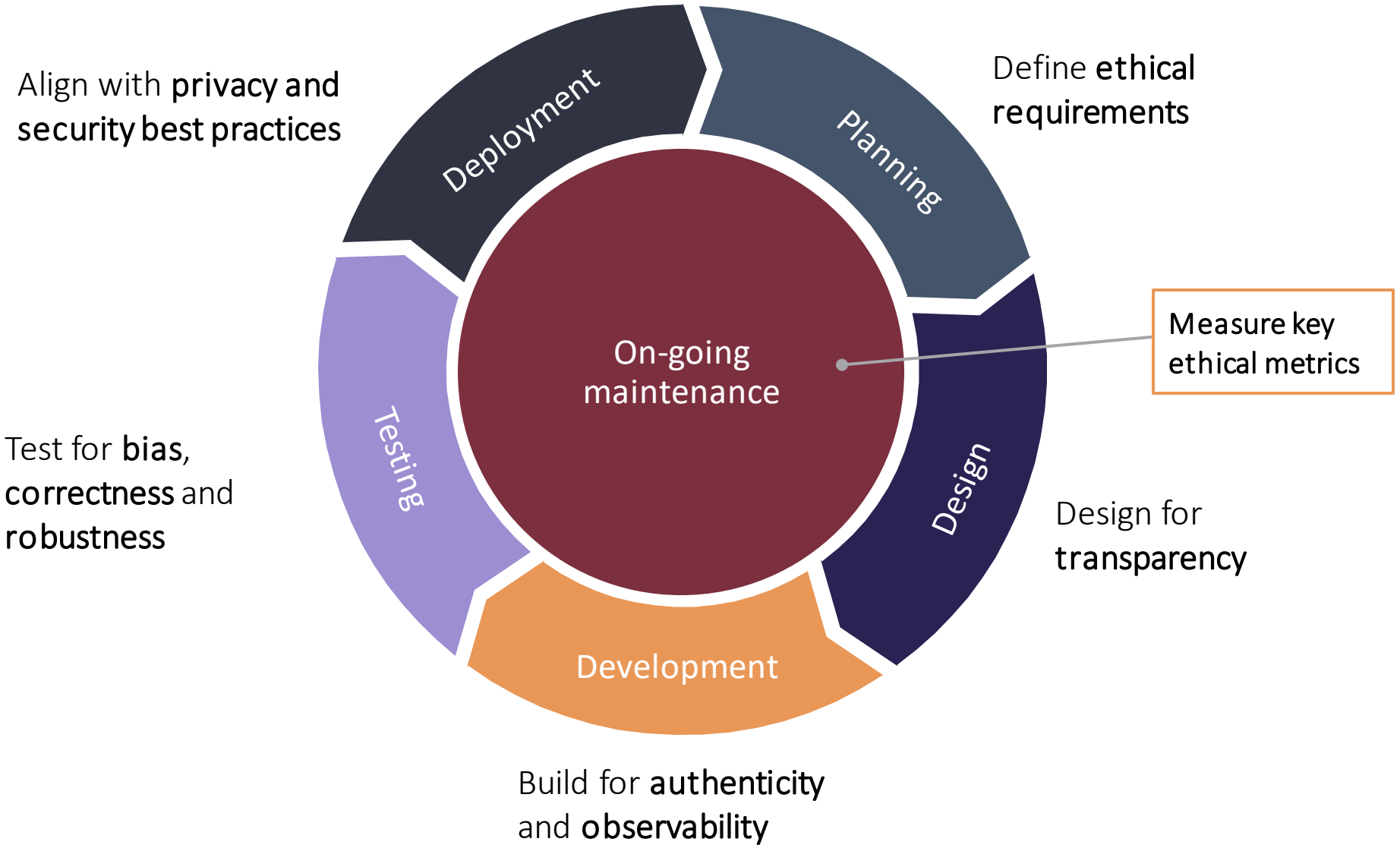


System quality

Data relevance

Business impact

User satisfaction


Align with privacy and security best practices
Deployment
Define ethical requirements
Planning
On-going maintenance
Measure key ethical metrics
Test for bias, correctness and robustness
Testing
Design
Design for transparency
Development
Build for authenticity and observability

- Regulations are behind technology. Go beyond them
- It is our responsibility to adapt and act
- You won't make it 100% right but <u>you can make better</u>

THANK YOU!

MARIA GOMEZ AGUIRRE @mariascandella
PRIYANKA SYAL

LeadDev Berlin - Dec 5th 2023