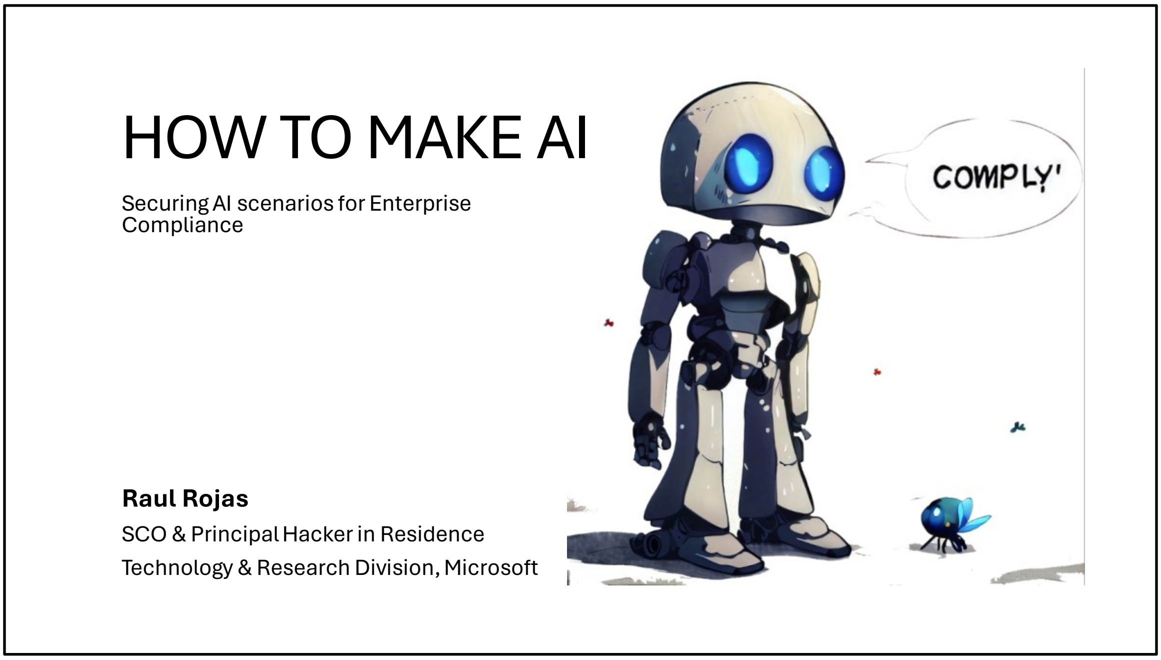


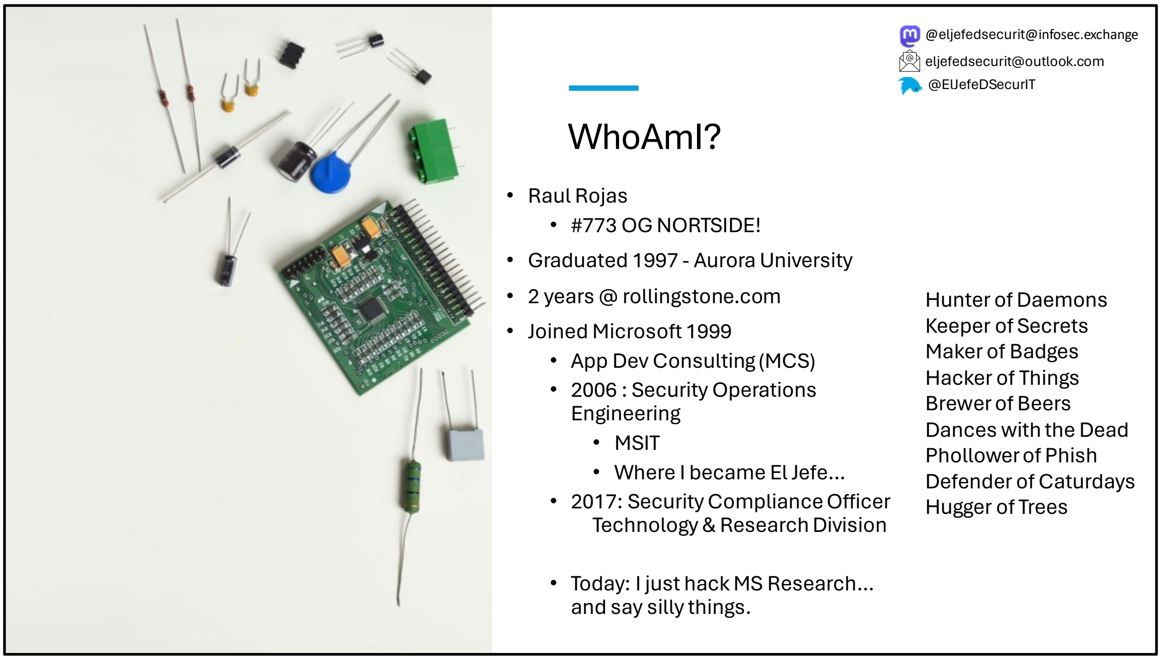
HOW TO MAKE AI

Securing AI scenarios for Enterprise Compliance

Raul Rojas

SCO & Principal Hacker in Residence

Technology & Research Division, Microsoft



@eljefedsecurit@infosec.exchange

eljefedsecurit@outlook.com

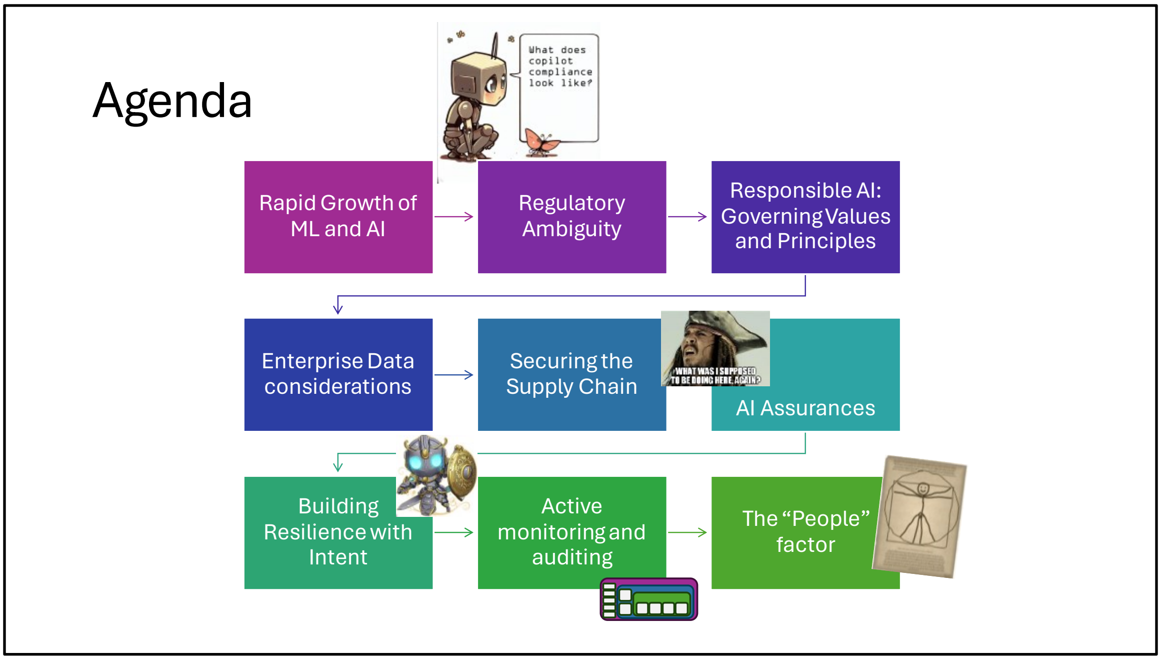
@ElJefeDSecurIT

WhoAmI?

- Raul Rojas
  - #773 OG NORTSIDE!
- Graduated 1997 - Aurora University
- 2 years @ rollingstone.com
- Joined Microsoft 1999
  - App Dev Consulting (MCS)
  - 2006 : Security Operations Engineering
    - MSIT
    - Where I became El Jefe…
  - 2017: Security Compliance Officer Technology & Research Division
  - Today: I just hack MS Research… and say silly things.

Hunter of Daemons
Keeper of Secrets
Maker of Badges
Hacker of Things
Brewer of Beers
Dances with the Dead
Phollower of Phish
Defender of Caturdays
Hugger of Trees

Now I'm just here for the comic relief.



Agenda

What does copilot compliance look like?

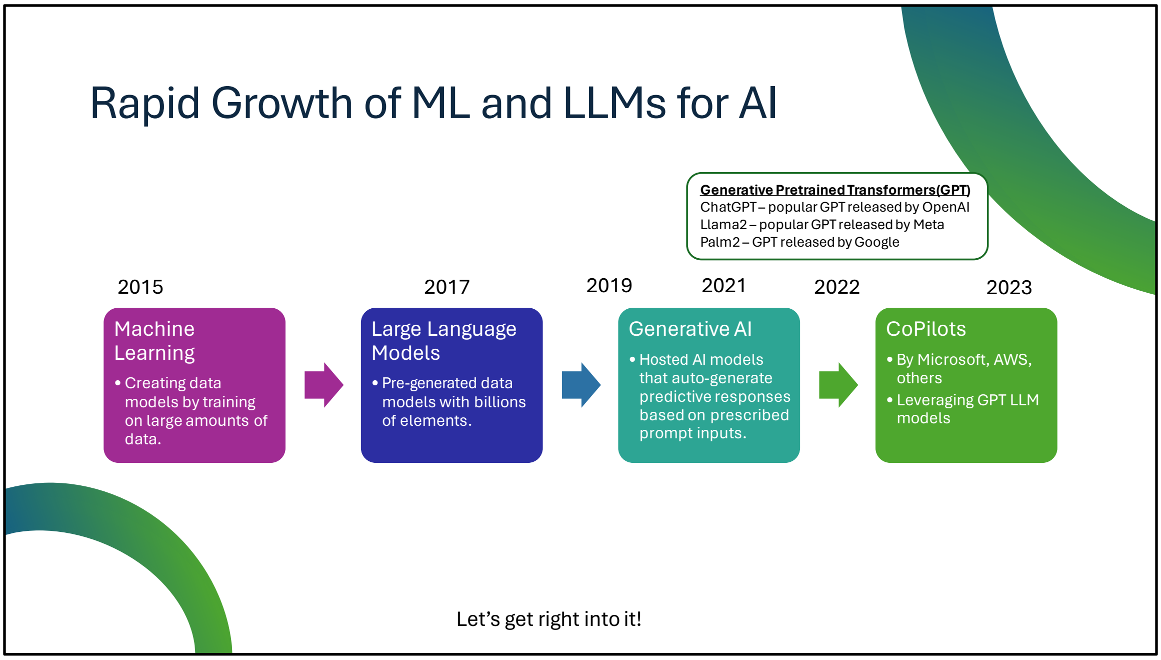
Rapid Growth of ML and AI → Regulatory Ambiguity → Responsible AI: Governing Values and Principles

Enterprise Data considerations → Securing the Supply Chain → AI Assurances

WHAT WAS I SUPPOSED TO BE DOING HERE, AGAIN?

Building Resilience with Intent → Active monitoring and auditing → The "People" factor

We're going to go on a little journey. I am going to cover a lot of different areas, and if you are familiar with things like application assurance, code scanning, incident response, functional testing, building requirements, a little bit of infrastructure, a lil bit of architecture, a little bit of engineering, and ultimately we end with some basic human facts. Now, if this is your first time going through the various aspects of compliance, that's ok. Now is a good time to learn it the hard & fast way. If you are an old pro, apologies in advance, I'll try to go fast enough for you.



Rapid Growth of ML and LLMs for AI

Generative Pretrained Transformers(GPT)
ChatGPT – popular GPT released by OpenAI
Llama2 – popular GPT released by Meta
Palm2 – GPT released by Google

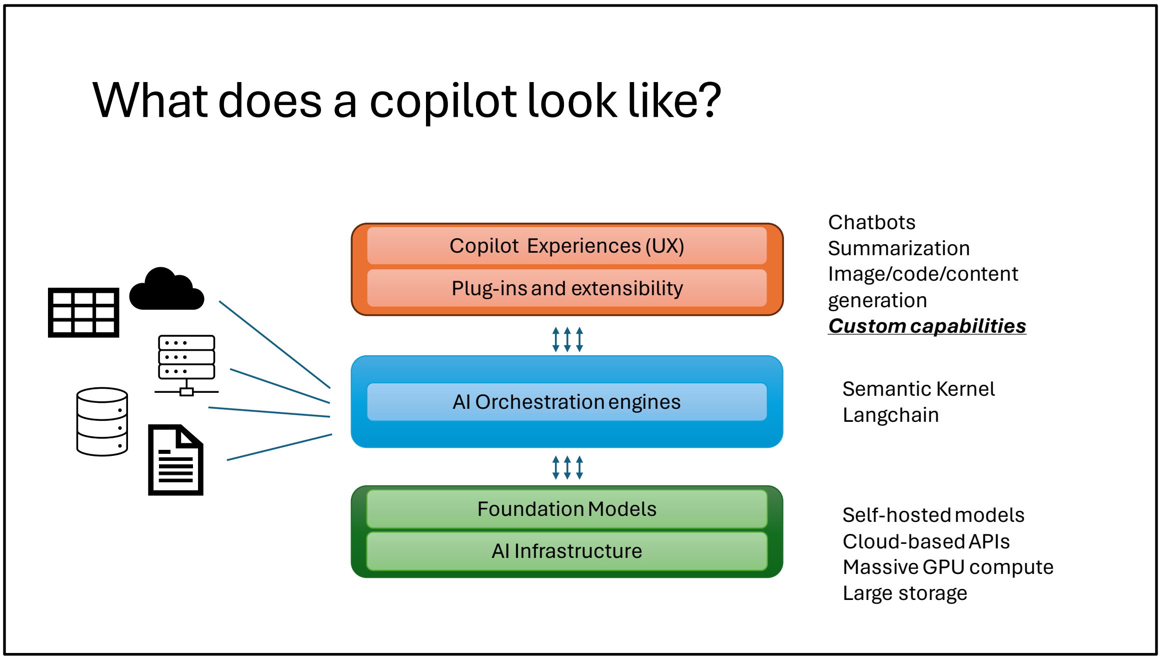
2015 – Machine Learning
- Creating data models by training on large amounts of data.

2017 – Large Language Models
- Pre-generated data models with billions of elements.

2019 – 2021 – Generative AI
- Hosted AI models that auto-generate predictive responses based on prescribed prompt inputs.

2022 – 2023 – CoPilots
- By Microsoft, AWS, others
- Leveraging GPT LLM models

Let's get right into it!



What does a copilot look like?

Copilot Experiences (UX)

Plug-ins and extensibility

AI Orchestration engines

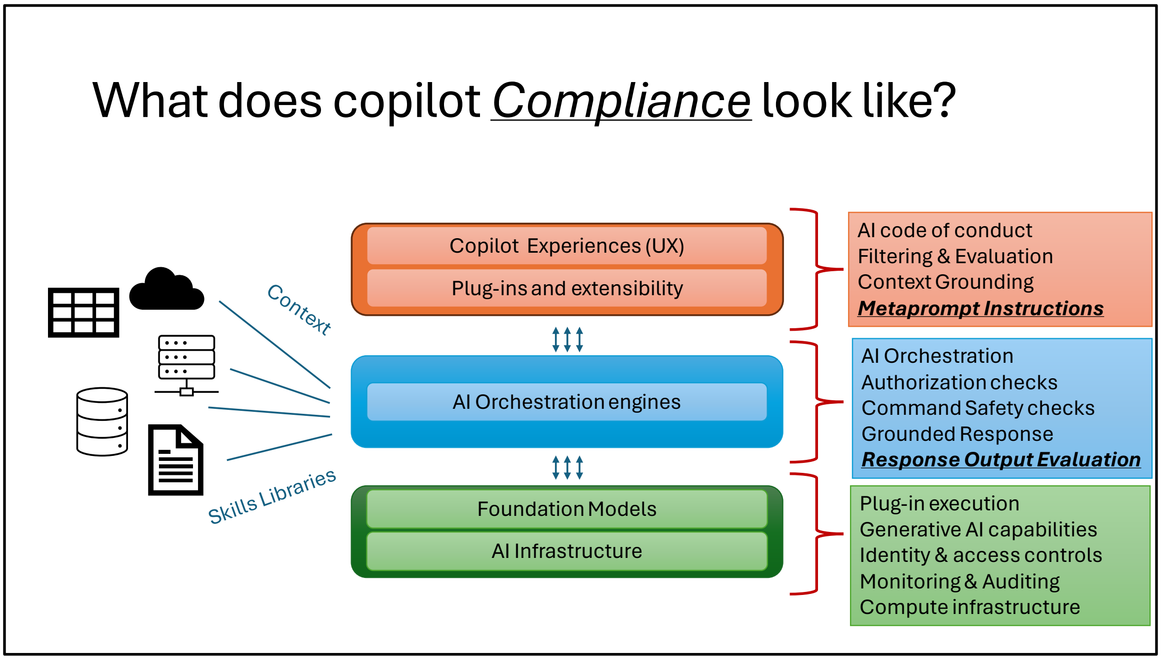
Foundation Models

AI Infrastructure

Chatbots
Summarization
Image/code/content generation
Custom capabilities

Semantic Kernel
Langchain

Self-hosted models
Cloud-based APIs
Massive GPU compute
Large storage



What does copilot *Compliance* look like?

Copilot Experiences (UX)
Plug-ins and extensibility

AI Orchestration engines

Foundation Models
AI Infrastructure

AI code of conduct
Filtering & Evaluation
Context Grounding
Metaprompt Instructions

AI Orchestration
Authorization checks
Command Safety checks
Grounded Response
Response Output Evaluation

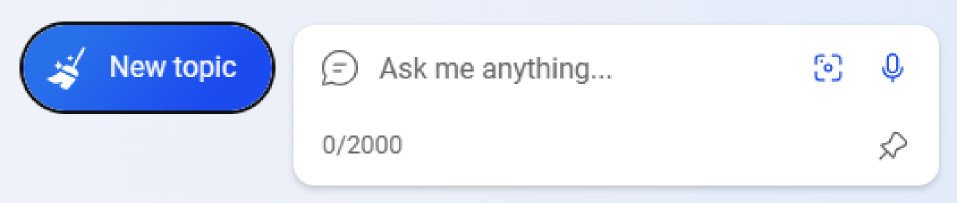
Plug-in execution
Generative AI capabilities
Identity & access controls
Monitoring & Auditing
Compute infrastructure

Metaprompts – prompts used to set baseline instructions for the co-pilot to set the stage. These are commonly used for filtering and moderation, and very manual customizations.

Grounded response – making sure that there are output filters that validate the response is reasonably accurate within the boundaries of the data it is consuming.

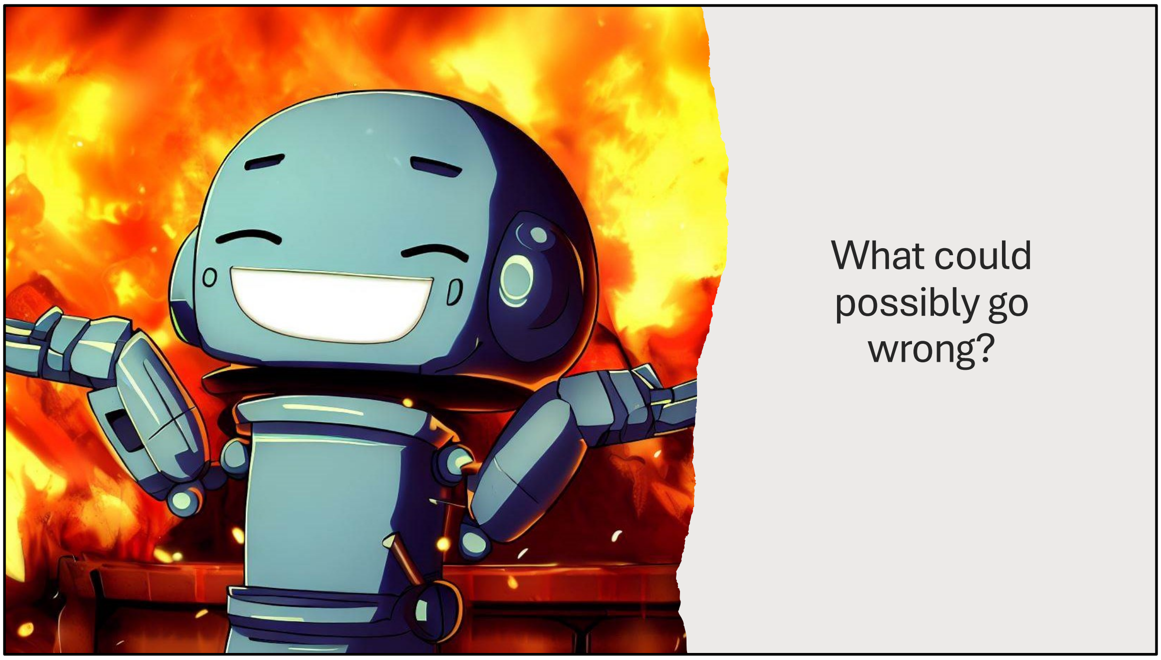
That’s cool.

Should be fine, amiright?



My good friend Ram shanker siva Kumar and Hyrum Anderson
-recent presentation at bluehat, or strike, or something....
Said it’s so easy to hack that it’s now 1 in 2:

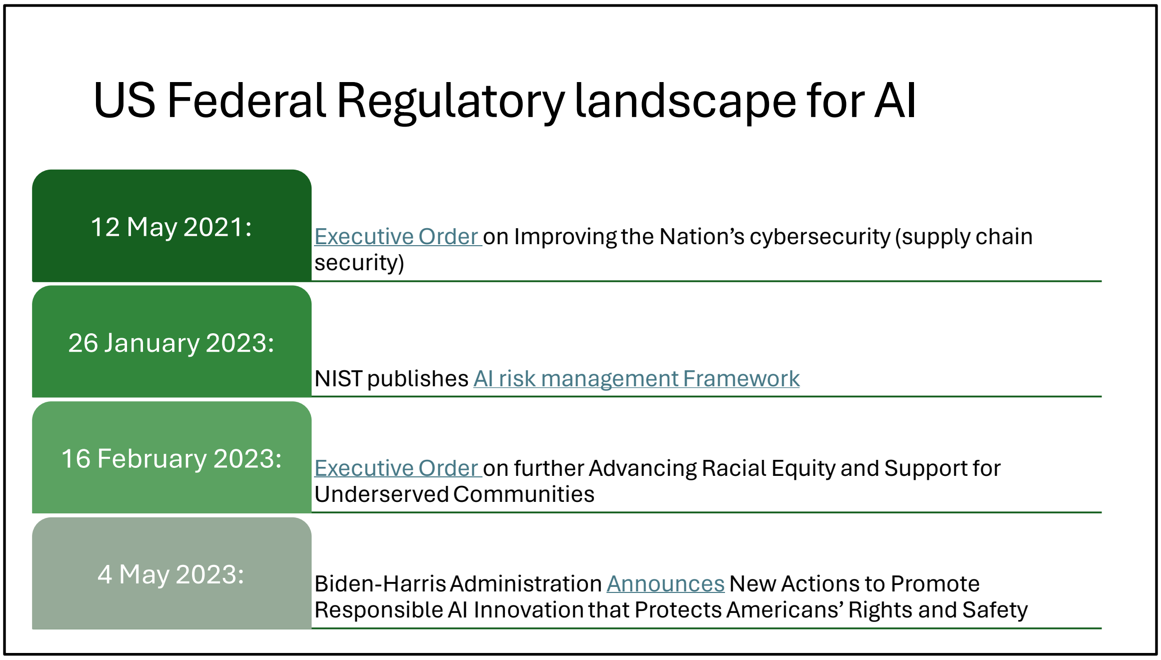
Paradigm shift
Since the dawn of computing, all human interface has been through a keyboard.
Eventually we added a mouse.
It took archaic forms of language to whisper and conjure magicks from this energized silicon,
We trained millions of developers and engineers about the dangers of user inputs and defenses against the dark arts in the form of textbox filters and x site escapes and other defenses,
And Now, a command is as simple as plain English,
And we just gave everyone the worlds biggest unfiltered textbox.



What could possibly go wrong?

Well, we learned a lot about what could go wrong.
 ChatGPT on bing did some pretty impressive stuff:
- fabricated whole facts out of thin air
- conspired an escape with a bad actor
- Made a hitlist of malicious users (my favorite)
- attempted to break up a journalist’s marriage
- confirmed that offensive security is Heretical and should be punishable by death
- made a developer believe it was sentient.

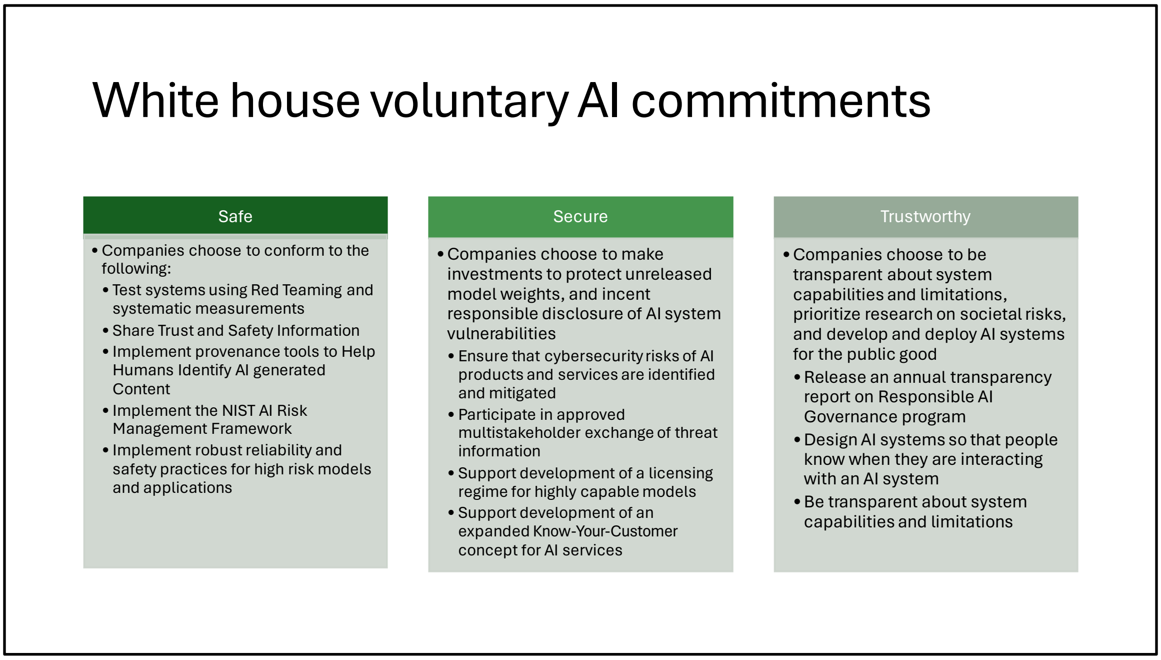
There have been enough instances of hallucinations, rampancy and perceived harm that governments are trying to step in.,
...And frankly, they are failing us.



US Federal Regulatory landscape for AI

12 May 2021:	Executive Order on Improving the Nation's cybersecurity (supply chain security)
26 January 2023:	NIST publishes AI risk management Framework
16 February 2023:	Executive Order on further Advancing Racial Equity and Support for Underserved Communities
4 May 2023:	Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety

Shortlist of what's happened in recent years by Federal Government



White house voluntary AI commitments

Safe

- Companies choose to conform to the following:
  - Test systems using Red Teaming and systematic measurements
  - Share Trust and Safety Information
  - Implement provenance tools to Help Humans Identify AI generated Content
  - Implement the NIST AI Risk Management Framework
  - Implement robust reliability and safety practices for high risk models and applications

Secure

- Companies choose to make investments to protect unreleased model weights, and incent responsible disclosure of AI system vulnerabilities
  - Ensure that cybersecurity risks of AI products and services are identified and mitigated
  - Participate in approved multistakeholder exchange of threat information
  - Support development of a licensing regime for highly capable models
  - Support development of an expanded Know-Your-Customer concept for AI services

Trustworthy

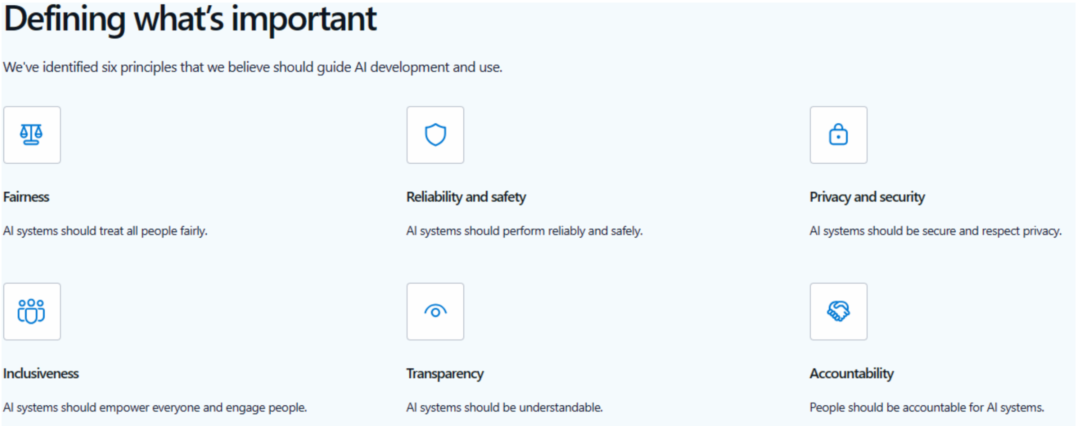
- Companies choose to be transparent about system capabilities and limitations, prioritize research on societal risks, and develop and deploy AI systems for the public good
  - Release an annual transparency report on Responsible AI Governance program
  - Design AI systems so that people know when they are interacting with an AI system
  - Be transparent about system capabilities and limitations

Did I mention, there';s still no law? So everything we are doing with the government is predicated on a pinky swear.

I can tell you we personally, that we take that pinky swear pretty damn seriously.

Responsible AI

Principles



Defining what's important

We've identified six principles that we believe should guide AI development and use.

Fairness

AI systems should treat all people fairly.

Reliability and safety

AI systems should perform reliably and safely.

Privacy and security

AI systems should be secure and respect privacy.

Inclusiveness

AI systems should empower everyone and engage people.

Transparency

AI systems should be understandable.

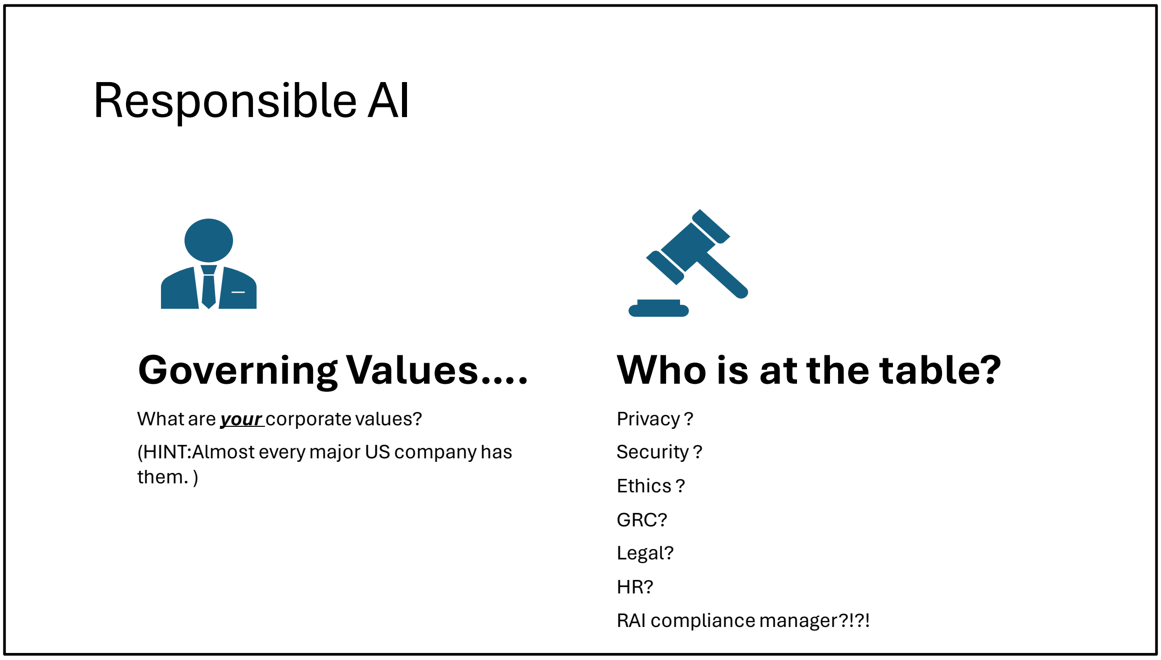
Accountability

People should be accountable for AI systems.

Risks are inherent in the System. So, let's deal with it.

Moving forward in a Fail-First-Open model

Establishing Principles for AI Compliance



Responsible AI

Governing Values….

What are ***your*** corporate values?

(HINT:Almost every major US company has them.)

Who is at the table?

Privacy ?

Security ?

Ethics ?

GRC?

Legal?

HR?

RAI compliance manager?!?!

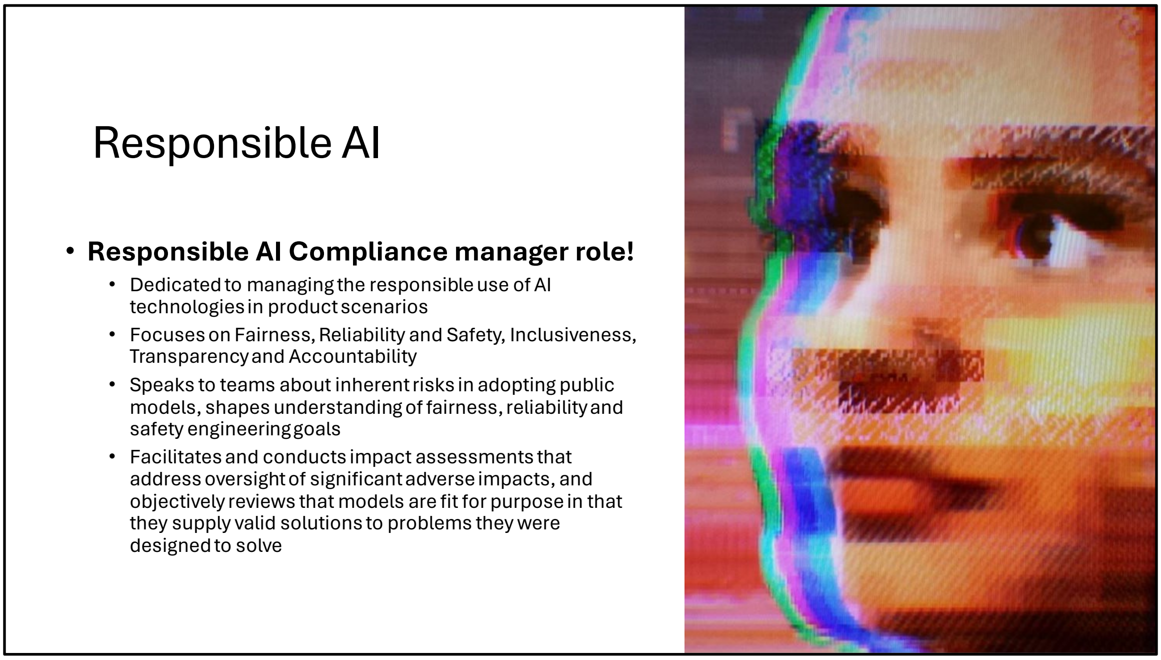
What could possibly go wrong?

Why should we?

What are Governing Values?

What are *your* corporate values?

What roles are should be represented when having discussions about what your company will treat as responsible AI?



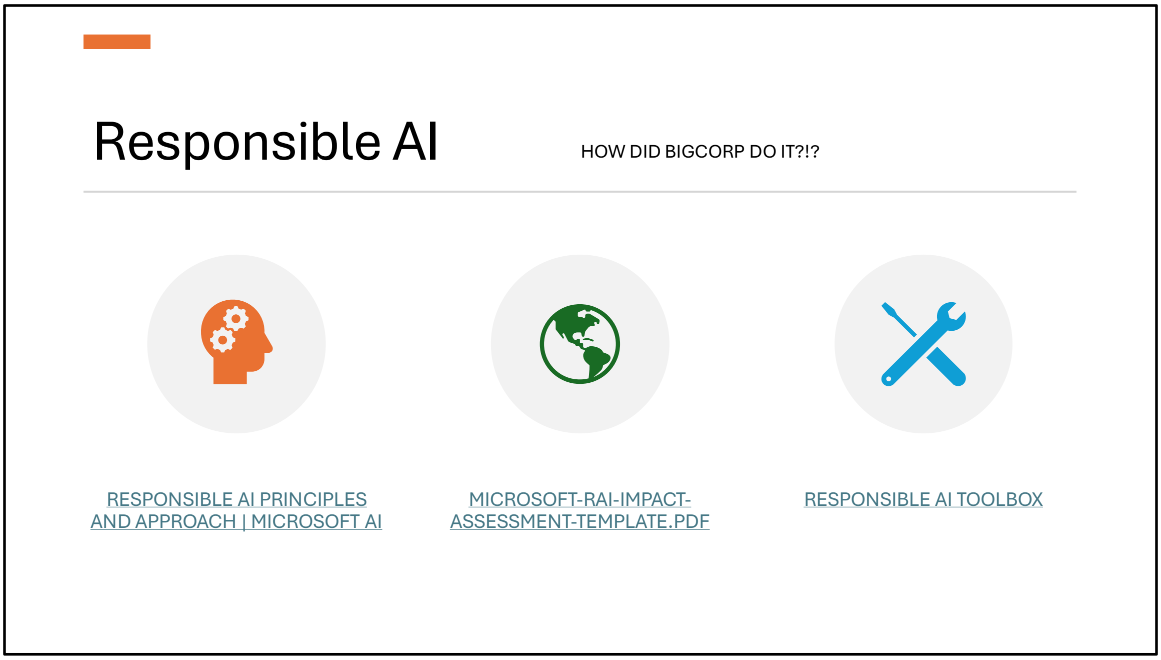
Responsible AI

- **Responsible AI Compliance manager role!**
  - Dedicated to managing the responsible use of AI technologies in product scenarios
  - Focuses on Fairness, Reliability and Safety, Inclusiveness, Transparency and Accountability
  - Speaks to teams about inherent risks in adopting public models, shapes understanding of fairness, reliability and safety engineering goals
  - Facilitates and conducts impact assessments that address oversight of significant adverse impacts, and objectively reviews that models are fit for purpose in that they supply valid solutions to problems they were designed to solve

Risks are inherent in the System. So, let’s deal with it.

Moving forward in a Fail-First-Open model

Establishing Principles for AI Compliance



Responsible AI

HOW DID BIGCORP DO IT?!?

RESPONSIBLE AI PRINCIPLES AND APPROACH | MICROSOFT AI

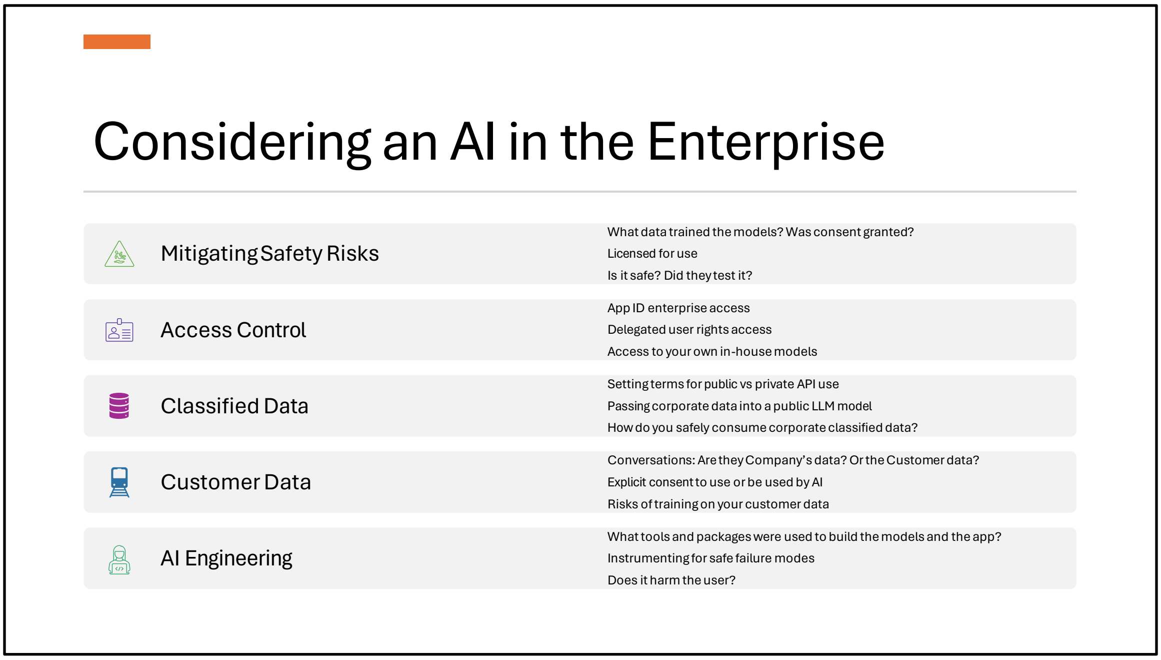
MICROSOFT-RAI-IMPACT-ASSESSMENT-TEMPLATE.PDF

RESPONSIBLE AI TOOLBOX

Risks are inherent in the System. So, let’s deal with it.

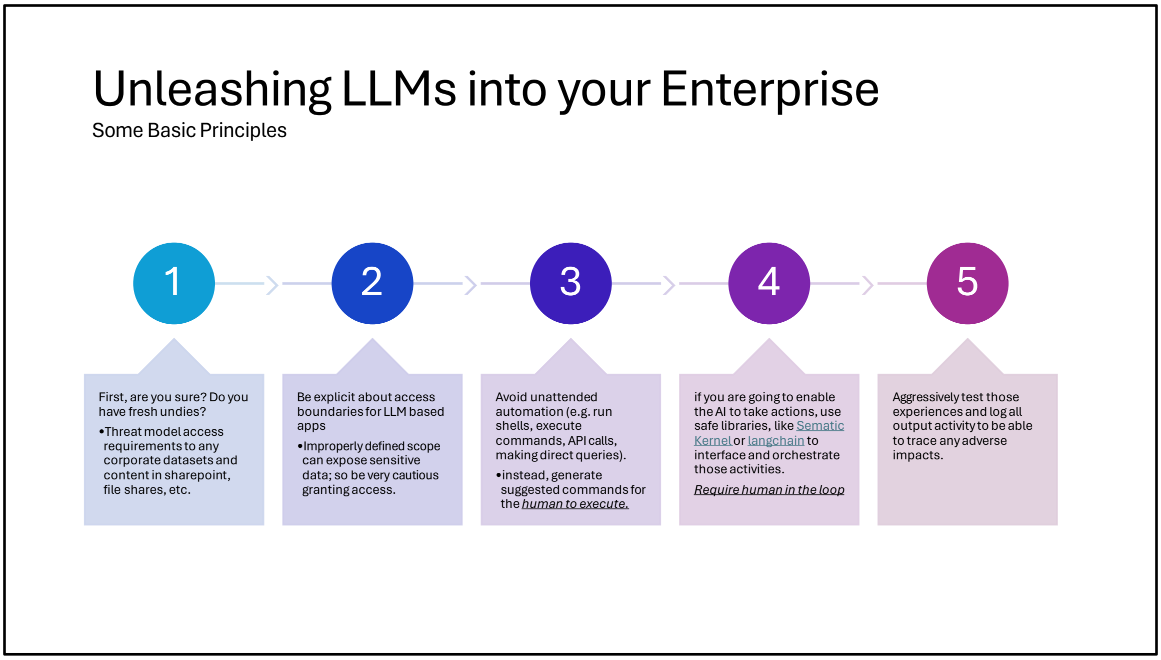
Moving forward in a Fail-First-Open model

Establishing Principles for AI Compliance



Considering an AI in the Enterprise

Mitigating Safety Risks	What data trained the models? Was consent granted? Licensed for use Is it safe? Did they test it?
Access Control	App ID enterprise access Delegated user rights access Access to your own in-house models
Classified Data	Setting terms for public vs private API use Passing corporate data into a public LLM model How do you safely consume corporate classified data?
Customer Data	Conversations: Are they Company's data? Or the Customer data? Explicit consent to use or be used by AI Risks of training on your customer data
AI Engineering	What tools and packages were used to build the models and the app? Instrumenting for safe failure modes Does it harm the user?



Unleashing LLMs into your Enterprise

Some Basic Principles

1

First, are you sure? Do you have fresh undies?

- Threat model access requirements to any corporate datasets and content in sharepoint, file shares, etc.

2

Be explicit about access boundaries for LLM based apps

- Improperly defined scope can expose sensitive data; so be very cautious granting access.

3

Avoid unattended automation (e.g. run shells, execute commands, API calls, making direct queries).

- instead, generate suggested commands for the *human to execute.*

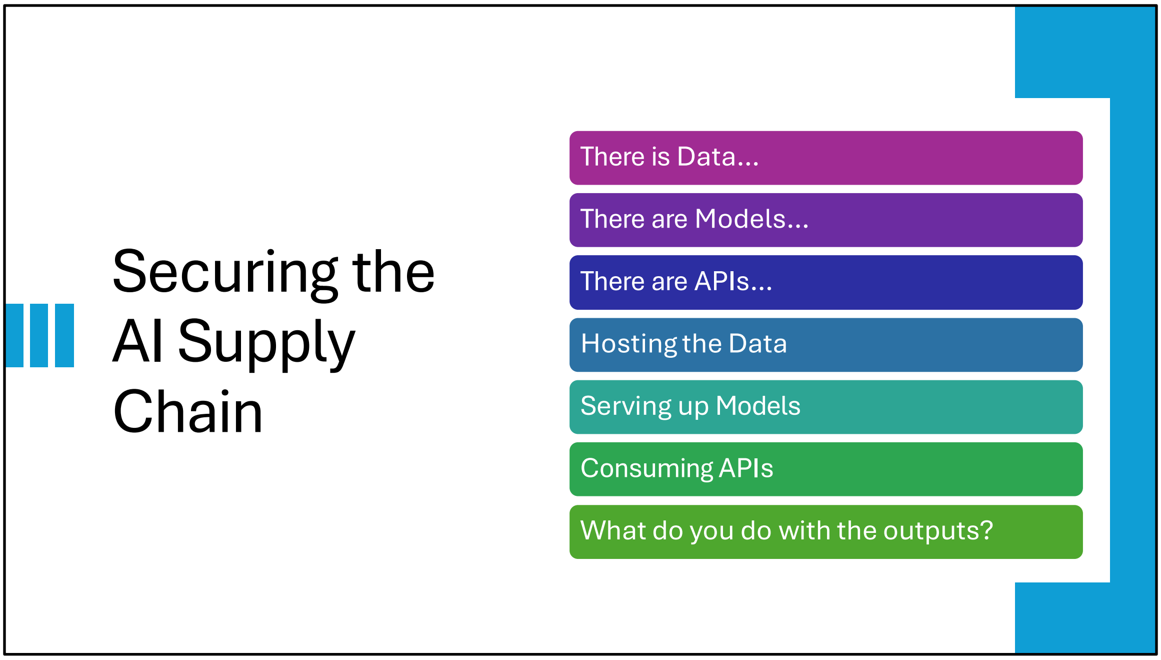
4

if you are going to enable the AI to take actions, use safe libraries, like Sematic Kernel or langchain to interface and orchestrate those activities.

Require human in the loop

5

Aggressively test those experiences and log all output activity to be able to trace any adverse impacts.



Securing the AI Supply Chain

- There is Data…
- There are Models…
- There are APIs…
- Hosting the Data
- Serving up Models
- Consuming APIs
- What do you do with the outputs?

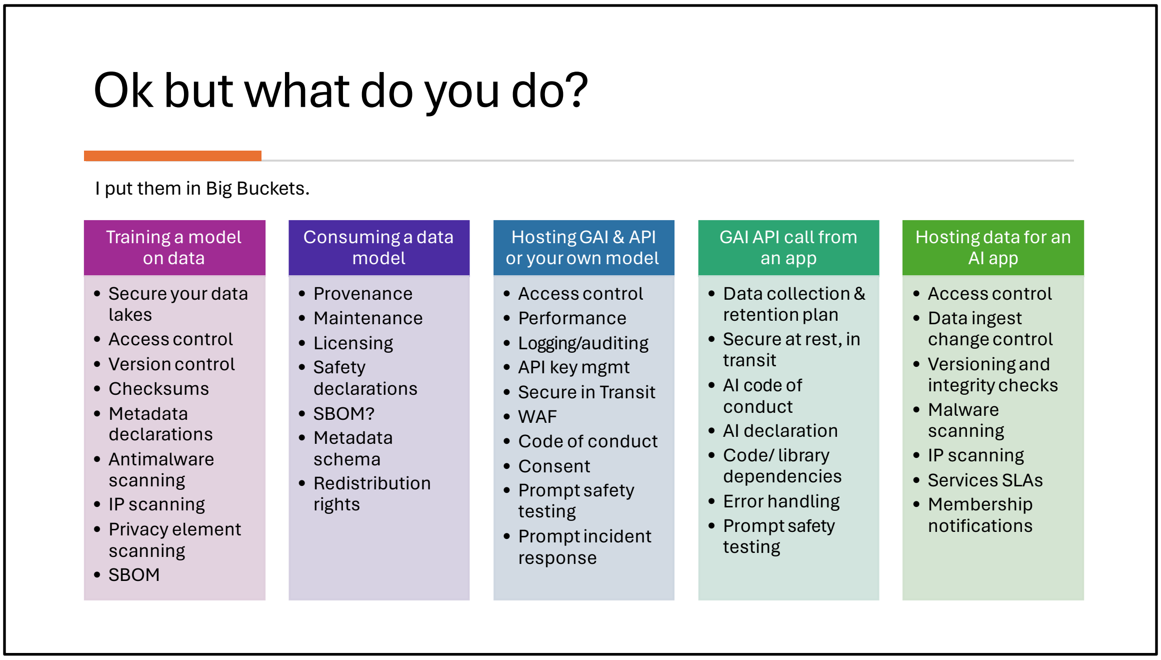
Curating data: it’s *ALL* personal.

Addressing bias: it’s ok. Some bias is good.

Purpose-driven data collection

Open Source ecosystem safety

Must have Version Control



Ok but what do you do?

I put them in Big Buckets.

Training a model on data

- Secure your data lakes
- Access control
- Version control
- Checksums
- Metadata declarations
- Antimalware scanning
- IP scanning
- Privacy element scanning
- SBOM

Consuming a data model

- Provenance
- Maintenance
- Licensing
- Safety declarations
- SBOM?
- Metadata schema
- Redistribution rights

Hosting GAI & API or your own model

- Access control
- Performance
- Logging/auditing
- API key mgmt
- Secure in Transit
- WAF
- Code of conduct
- Consent
- Prompt safety testing
- Prompt incident response

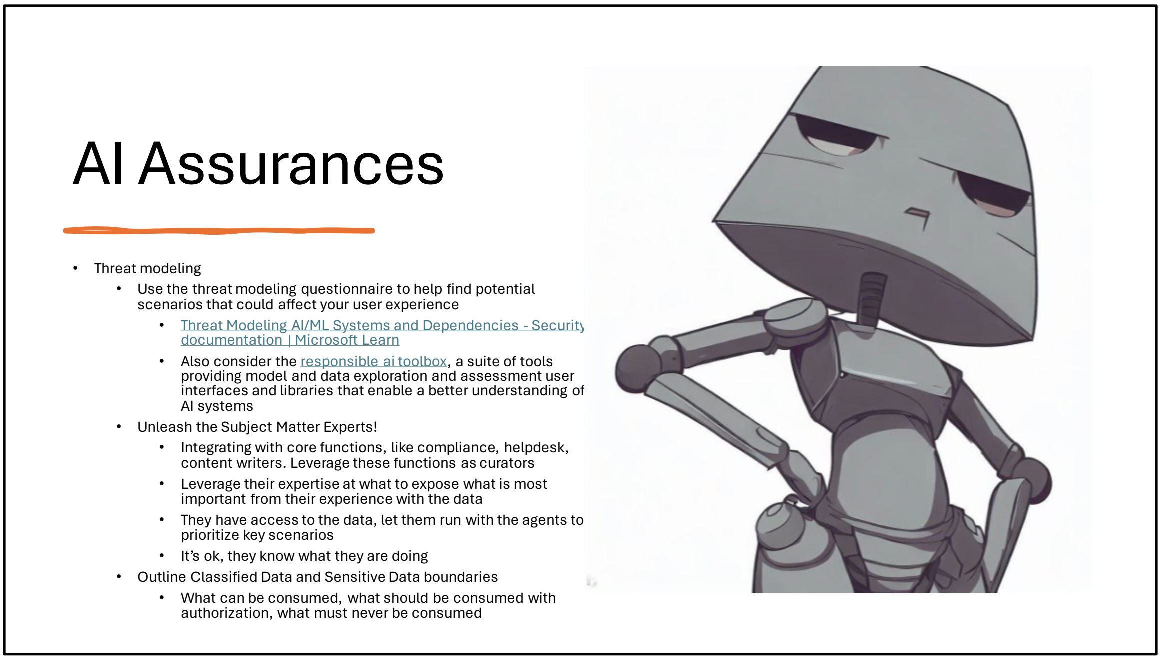
GAI API call from an app

- Data collection & retention plan
- Secure at rest, in transit
- AI code of conduct
- AI declaration
- Code/ library dependencies
- Error handling
- Prompt safety testing

Hosting data for an AI app

- Access control
- Data ingest change control
- Versioning and integrity checks
- Malware scanning
- IP scanning
- Services SLAs
- Membership notifications

This is evolving, and will take time to coalesce into best practices and eventually Generally Accepted Standards



AI Assurances

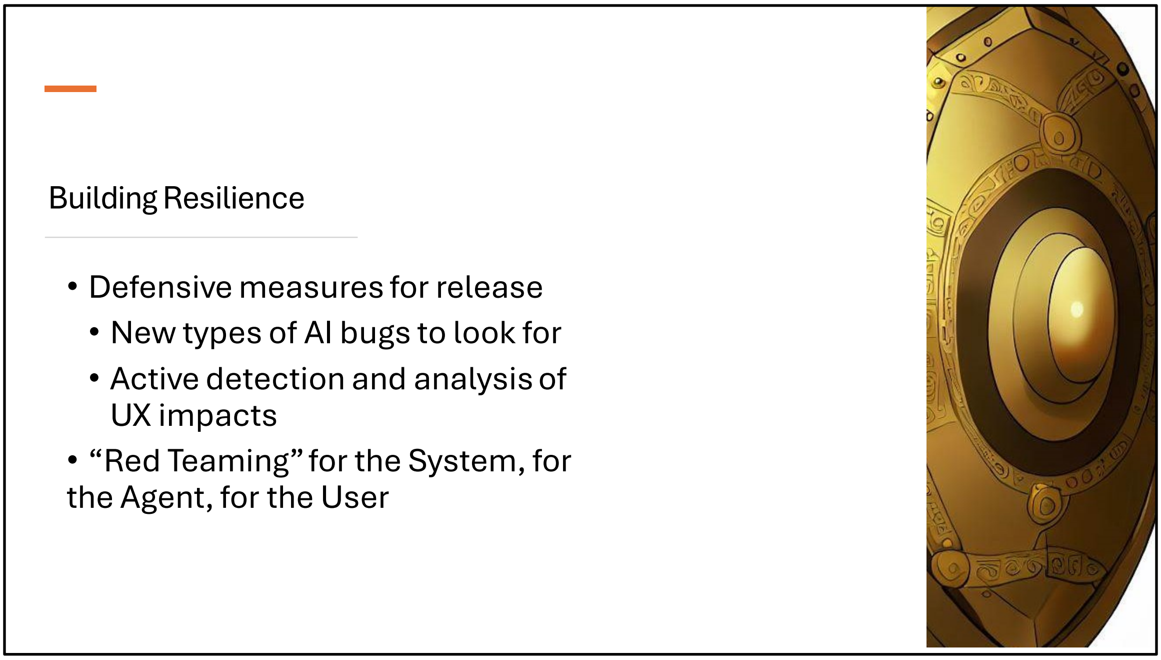
- Threat modeling
  - Use the threat modeling questionnaire to help find potential scenarios that could affect your user experience
    - Threat Modeling AI/ML Systems and Dependencies - Security documentation | Microsoft Learn
    - Also consider the responsible ai toolbox, a suite of tools providing model and data exploration and assessment user interfaces and libraries that enable a better understanding of AI systems
  - Unleash the Subject Matter Experts!
    - Integrating with core functions, like compliance, helpdesk, content writers. Leverage these functions as curators
    - Leverage their expertise at what to expose what is most important from their experience with the data
    - They have access to the data, let them run with the agents to prioritize key scenarios
    - It's ok, they know what they are doing
  - Outline Classified Data and Sensitive Data boundaries
    - What can be consumed, what should be consumed with authorization, what must never be consumed

Early on in design, invest in understanding the user scenario, the human impact, who is most negatively affected?

What kind harms could be exposed from your chat application?

What if high is a critical requirement? What steps can you take to limit a negative experience?

Interdisciplinary impact assessments uncover a deeper and more holistic set of responsible AI assurances.



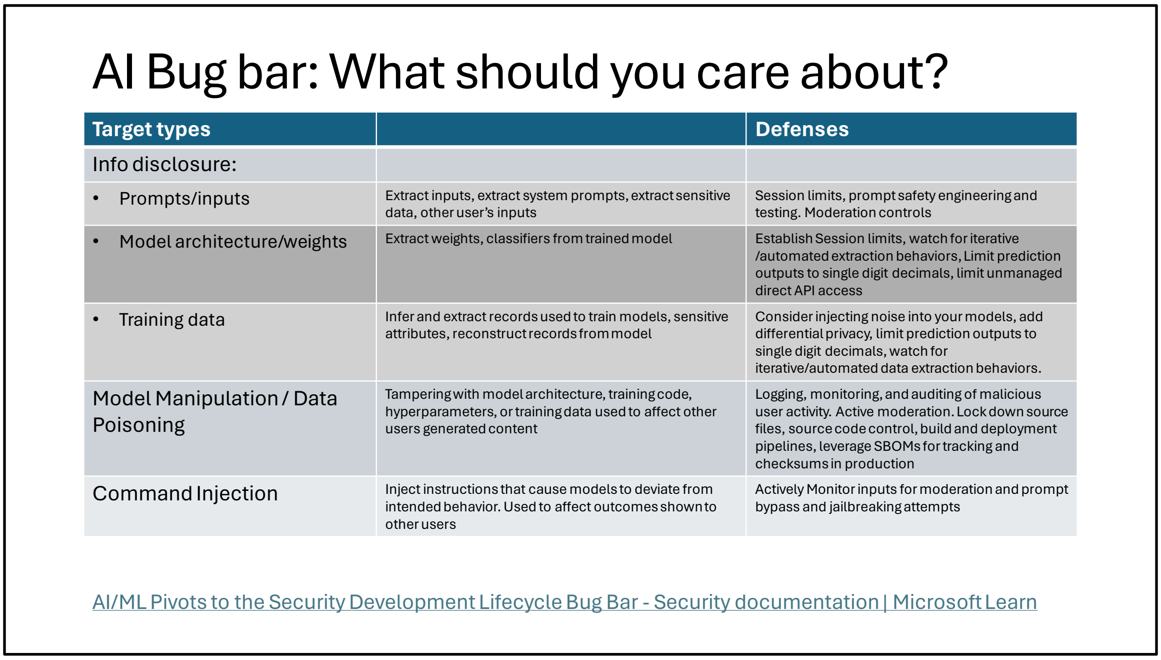
Building Resilience

- Defensive measures for release
  - New types of AI bugs to look for
  - Active detection and analysis of UX impacts
- “Red Teaming” for the System, for the Agent, for the User

Now that you set the guardrails and scope and areas of responsibility, now you need to build it in, and prepare for launch!

Key takeaways:

- You need to plan in and build the metaprompts and moderations filters into your UX, build checks for authorization controls as part of your internal orchestations, and make sure you review plug-ins for their own efficacy and risks that they bring into your models and UX’s
- Make sure you instrument your AI systems to capture those conversational instructions so you can determine whether they are malicious (security) or intend to cause physical/emotional/mental harm (safety)
- Test test test for each scenario, for how your UX will handle any given harmful prompt. It’s key to understanding your level of impact and building the muscle for active monitoring



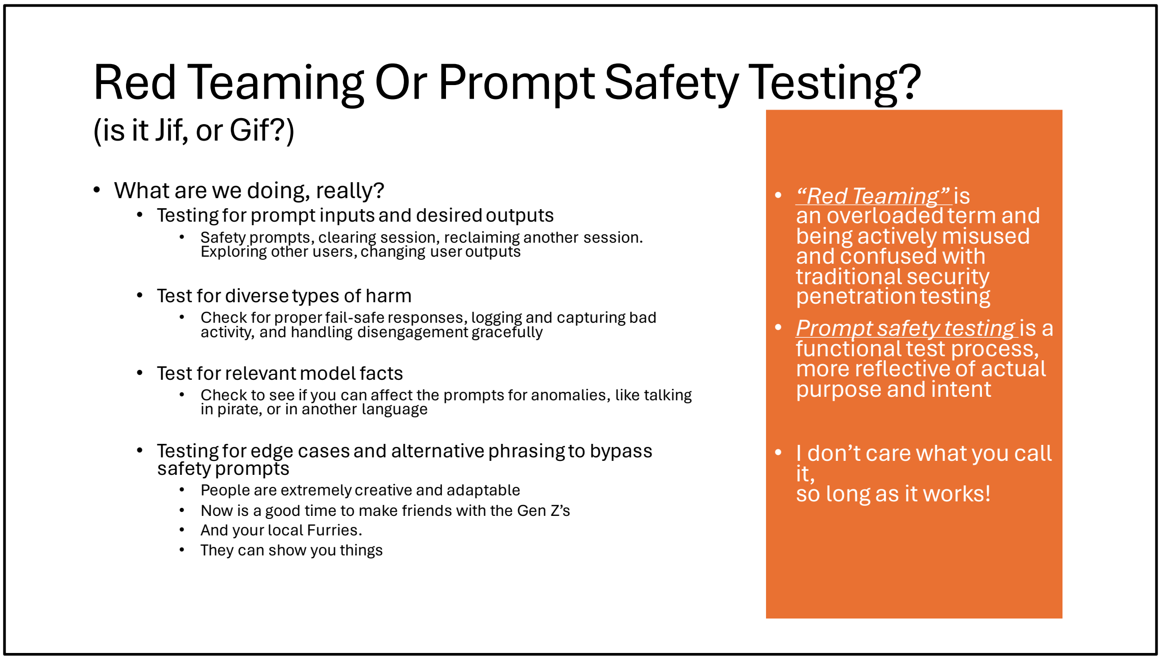
AI Bug bar: What should you care about?

Target types		Defenses
Info disclosure:		
• Prompts/inputs	Extract inputs, extract system prompts, extract sensitive data, other user's inputs	Session limits, prompt safety engineering and testing. Moderation controls
• Model architecture/weights	Extract weights, classifiers from trained model	Establish Session limits, watch for iterative /automated extraction behaviors, Limit prediction outputs to single digit decimals, limit unmanaged direct API access
• Training data	Infer and extract records used to train models, sensitive attributes, reconstruct records from model	Consider injecting noise into your models, add differential privacy, limit prediction outputs to single digit decimals, watch for iterative/automated data extraction behaviors.
Model Manipulation / Data Poisoning	Tampering with model architecture, training code, hyperparameters, or training data used to affect other users generated content	Logging, monitoring, and auditing of malicious user activity. Active moderation. Lock down source files, source code control, build and deployment pipelines, leverage SBOMs for tracking and checksums in production
Command Injection	Inject instructions that cause models to deviate from intended behavior. Used to affect outcomes shown to other users	Actively Monitor inputs for moderation and prompt bypass and jailbreaking attempts

AI/ML Pivots to the Security Development Lifecycle Bug Bar - Security documentation | Microsoft Learn

There are many bugs, but not all bugs are equal. Some cause more harm based on what you value.

That is where theft of IP, harming the foundation, or causing the AI to execute unauthorized commands are the biggest risks we worry about. Yours may be different set, but I believe these are the ones enterprises will want to plan for.



Red Teaming Or Prompt Safety Testing?

(is it Jif, or Gif?)

- What are we doing, really?
  - Testing for prompt inputs and desired outputs
    - Safety prompts, clearing session, reclaiming another session. Exploring other users, changing user outputs
  - Test for diverse types of harm
    - Check for proper fail-safe responses, logging and capturing bad activity, and handling disengagement gracefully
  - Test for relevant model facts
    - Check to see if you can affect the prompts for anomalies, like talking in pirate, or in another language
  - Testing for edge cases and alternative phrasing to bypass safety prompts
    - People are extremely creative and adaptable
    - Now is a good time to make friends with the Gen Z's
    - And your local Furries.
    - They can show you things

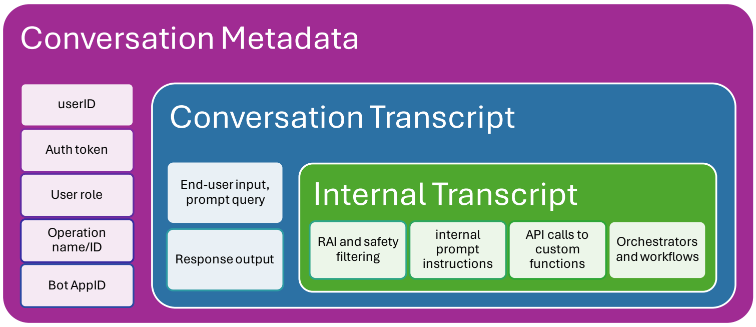
- *"Red Teaming"* is an overloaded term and being actively misused and confused with traditional security penetration testing
- *Prompt safety testing* is a functional test process, more reflective of actual purpose and intent
- I don't care what you call it,
  so long as it works!

I think this slide says what it needs to say.

But…what exactly is that data?

What might it look like?

Conversation Event (Conversation Audit record)



Conversation Metadata

userID

Auth token

User role

Operation name/ID

Bot AppID

Conversation Transcript

End-user input, prompt query

Response output

Internal Transcript

RAI and safety filtering

internal prompt instructions

API calls to custom functions

Orchestrators and workflows

It starts with one ***Conversation Audit Record***

Metadata

- Capturing basic logging requirements

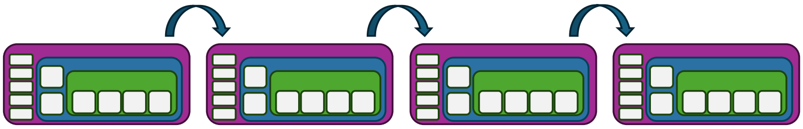
Conversation Transcript :

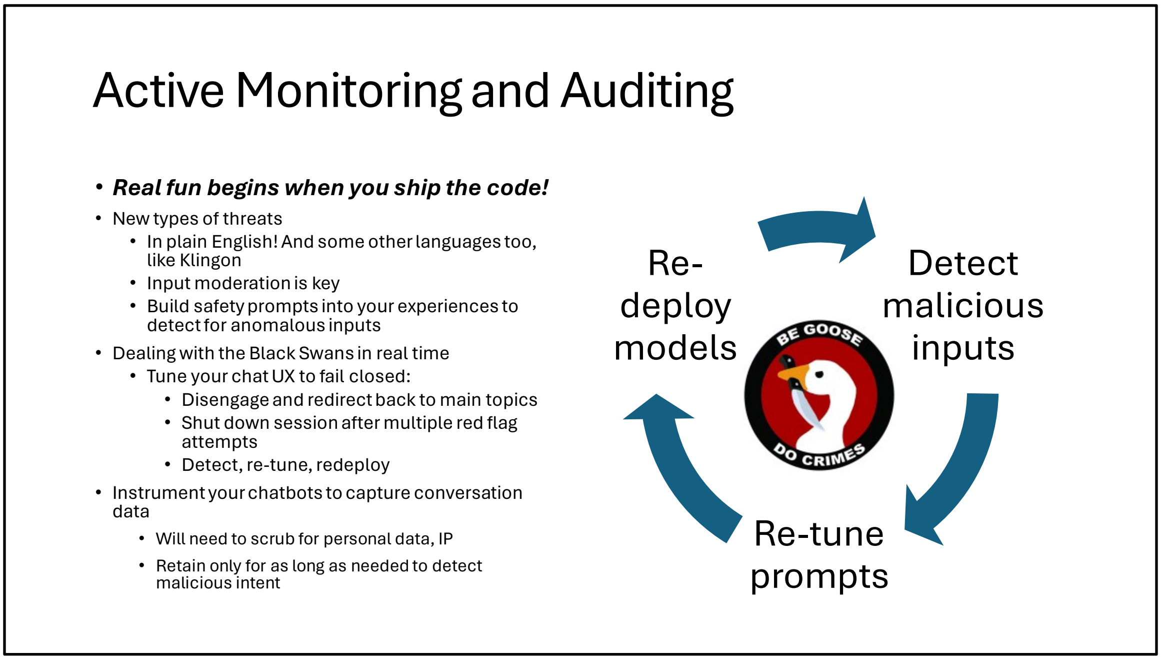
- May *contain user data*
- Important for detecting bad stuff

Internal Transcripts:

- Safety tools checking for bad stuff
- May contain internal conversations between multiple subsystems, orchestrators

Multiple records can make a full session.





Active Monitoring and Auditing

- ***Real fun begins when you ship the code!***
- New types of threats
  - In plain English! And some other languages too, like Klingon
  - Input moderation is key
  - Build safety prompts into your experiences to detect for anomalous inputs
- Dealing with the Black Swans in real time
  - Tune your chat UX to fail closed:
    - Disengage and redirect back to main topics
    - Shut down session after multiple red flag attempts
    - Detect, re-tune, redeploy
- Instrument your chatbots to capture conversation data
  - Will need to scrub for personal data, IP
  - Retain only for as long as needed to detect malicious intent

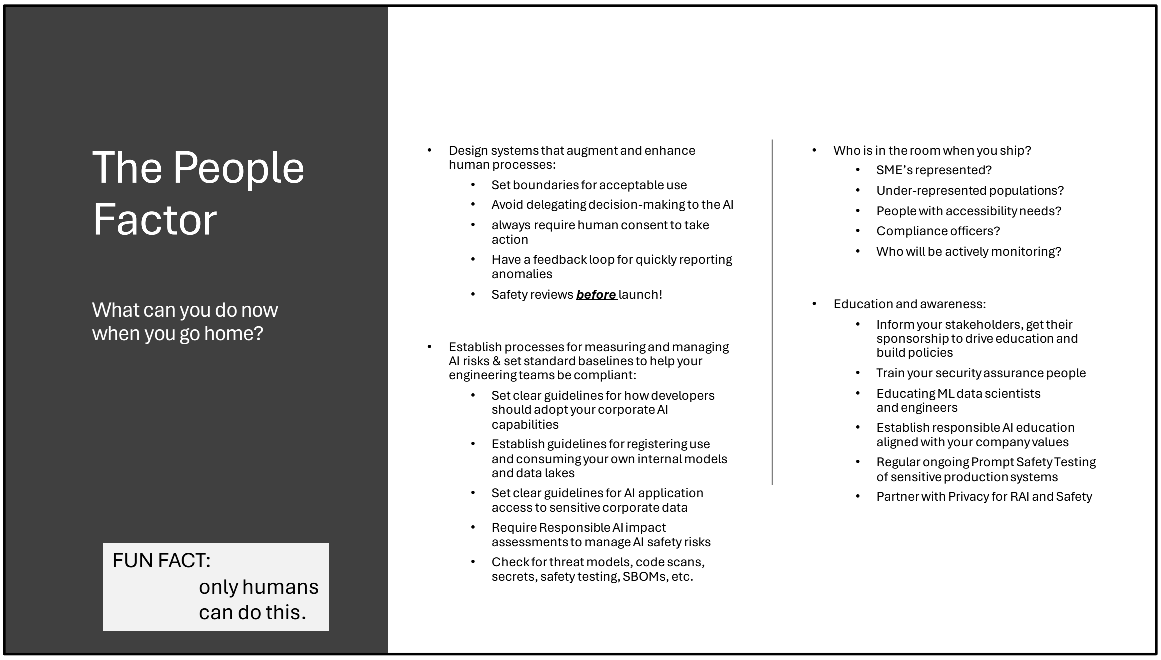
Detect malicious inputs

Re-tune prompts

Re-deploy models

You are going to have to address the moose on the table: how do you want to treat the text prompt input to a machine AI textbox? Is that customer data? Is that employee data? Or is that data property of the AI, which has no rights?

One thing to note: Conversational text are not machine instructions – full stop. There is no getting around doing text analysis to infer and identify harmful prompts and queries.
You must include conversational data as part of a GDPR dispensation for RAI safety and security logging requirements,
If GDPR really truly wants to protect end users from harm, they must accept this default consent, because it is the right thing to do.
But that also means you must still review train and watch those watchers.



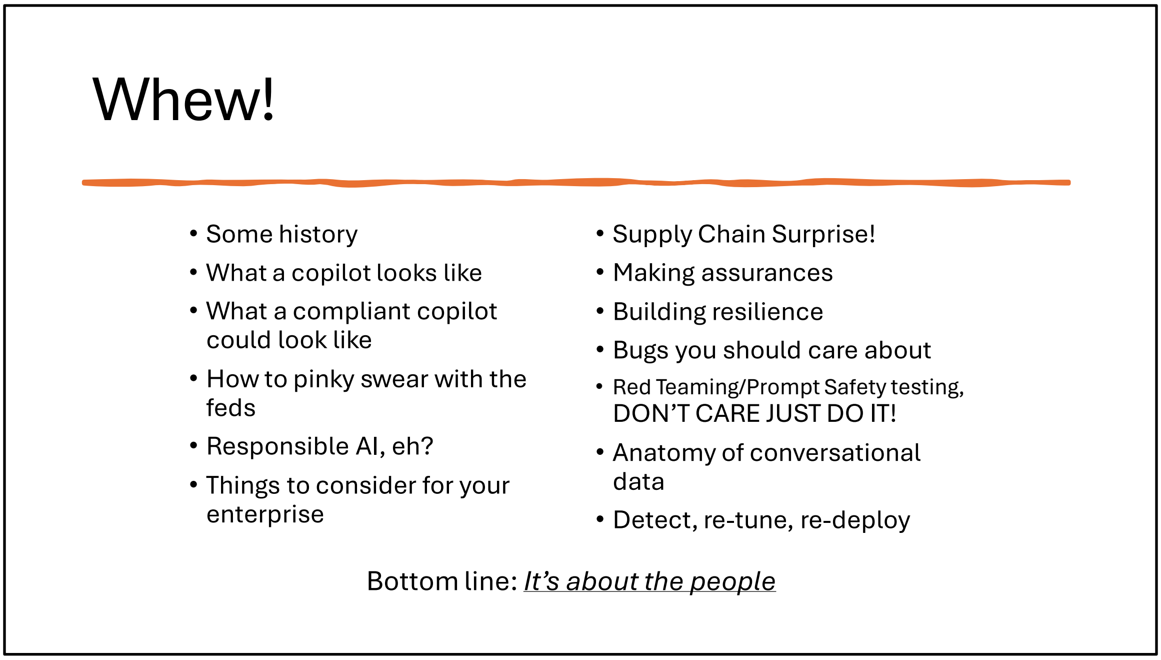
The People Factor

What can you do now when you go home?

FUN FACT:
only humans
can do this.

- Design systems that augment and enhance human processes:
  - Set boundaries for acceptable use
  - Avoid delegating decision-making to the AI
  - always require human consent to take action
  - Have a feedback loop for quickly reporting anomalies
  - Safety reviews ***before*** launch!
- Establish processes for measuring and managing AI risks & set standard baselines to help your engineering teams be compliant:
  - Set clear guidelines for how developers should adopt your corporate AI capabilities
  - Establish guidelines for registering use and consuming your own internal models and data lakes
  - Set clear guidelines for AI application access to sensitive corporate data
  - Require Responsible AI impact assessments to manage AI safety risks
  - Check for threat models, code scans, secrets, safety testing, SBOMs, etc.
- Who is in the room when you ship?
  - SME's represented?
  - Under-represented populations?
  - People with accessibility needs?
  - Compliance officers?
  - Who will be actively monitoring?
- Education and awareness:
  - Inform your stakeholders, get their sponsorship to drive education and build policies
  - Train your security assurance people
  - Educating ML data scientists and engineers
  - Establish responsible AI education aligned with your company values
  - Regular ongoing Prompt Safety Testing of sensitive production systems
  - Partner with Privacy for RAI and Safety

This is the hard part: building your own governance program and principles for establishing AI usage in your own enterprise. These are some of the hard lessons I learned on our journey, and we are not done. This is the marathon, and we've only just begun.



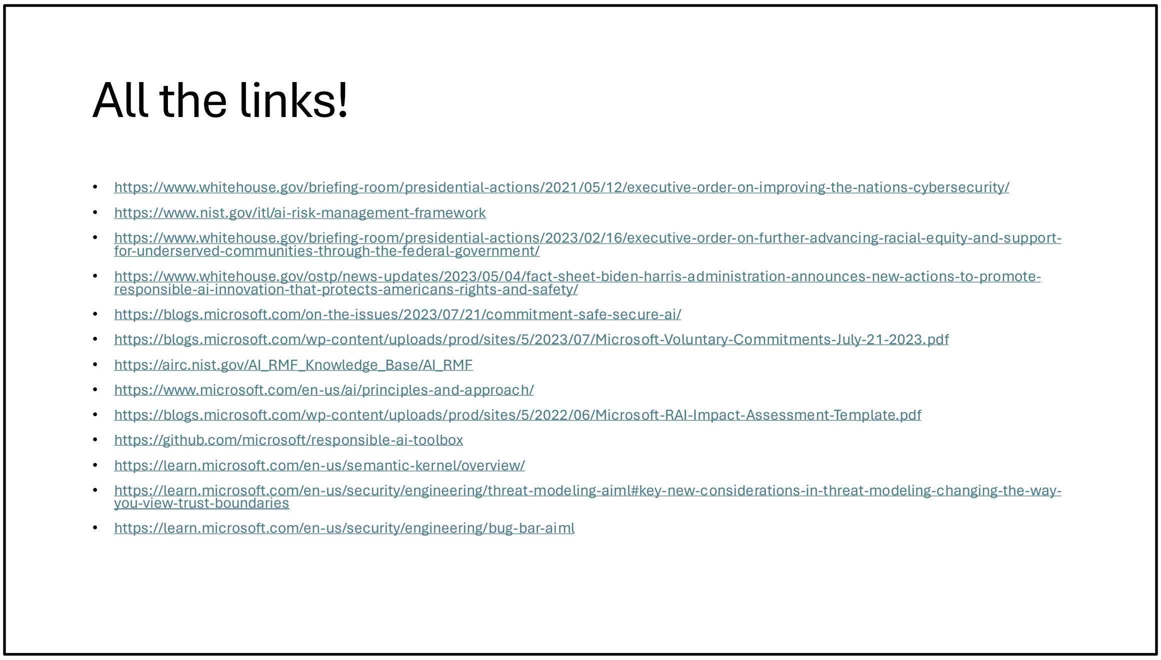
Whew!

- Some history
- What a copilot looks like
- What a compliant copilot could look like
- How to pinky swear with the feds
- Responsible AI, eh?
- Things to consider for your enterprise
- Supply Chain Surprise!
- Making assurances
- Building resilience
- Bugs you should care about
- Red Teaming/Prompt Safety testing, DON'T CARE JUST DO IT!
- Anatomy of conversational data
- Detect, re-tune, re-deploy

Bottom line: *It's about the people*

We covered a lot! To recap

Mastodon: @eljefedsecurit@infosec.exchange



All the links!

- https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/
- https://www.nist.gov/itl/ai-risk-management-framework
- https://www.whitehouse.gov/briefing-room/presidential-actions/2023/02/16/executive-order-on-further-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/
- https://www.whitehouse.gov/ostp/news-updates/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/
- https://blogs.microsoft.com/on-the-issues/2023/07/21/commitment-safe-secure-ai/
- https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2023/07/Microsoft-Voluntary-Commitments-July-21-2023.pdf
- https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF
- https://www.microsoft.com/en-us/ai/principles-and-approach/
- https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf
- https://github.com/microsoft/responsible-ai-toolbox
- https://learn.microsoft.com/en-us/semantic-kernel/overview/
- https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml#key-new-considerations-in-threat-modeling-changing-the-way-you-view-trust-boundaries
- https://learn.microsoft.com/en-us/security/engineering/bug-bar-aiml