



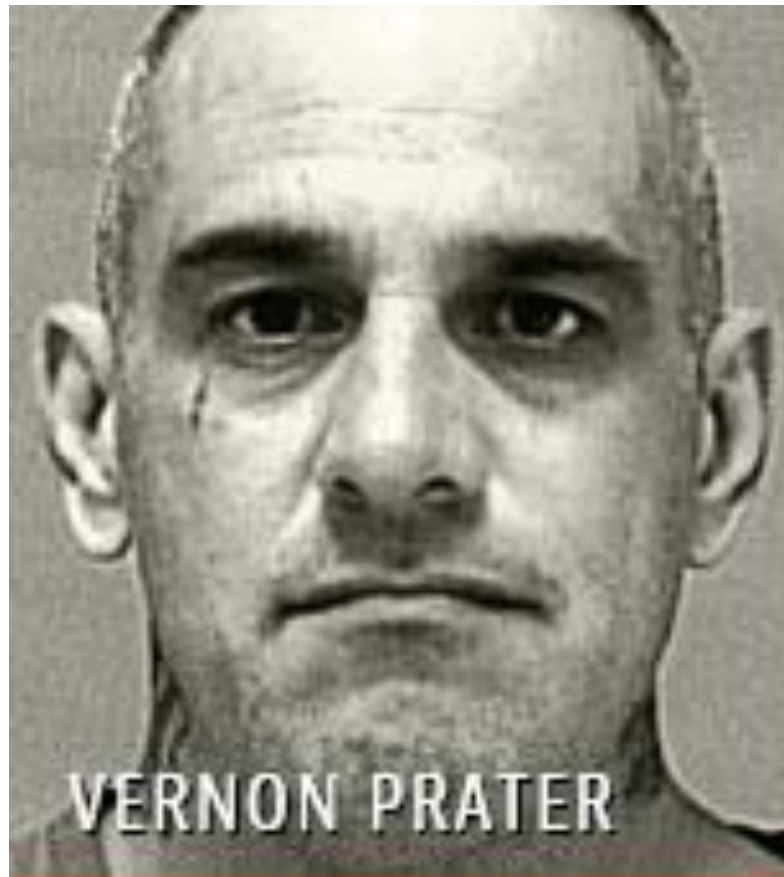
Responsible AI at Microsoft

Apr 25, 2024

Ome Sivadith
National Technology Officer

COMPAS is a risk assessment tool for recidivism that skews heavily against African-Americans but has success rate of 20% in their predictions. (Propublica)

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>



VERNON PRATER

LOW RISK

3

Current offence: Shop lifting
Prior offences: 2 armed robberies, 1 attempted armed robbery
Subsequent offences: 1 grand theft



BRISHA BORDEN

HIGH RISK

8

Current offence: Taking a bike on a spin without permission
Prior offences: 4 juvenile misdemeanors
Subsequent offences: None

Generative AI introduce new harms



Ungrounded outputs & errors



Jailbreaks & prompt injection attacks



Harmful content & code



Copyright infringement



Manipulation and human-like behavior

Gave Her Too Much Miralax!!

Umcat · Jun 11, 2019

Jun 11, 2019



Umcat

TCS Member

Thread starter

Kitten

Joined: May 26, 2019

Messages: 4

Purraise: 2

Our cat is severely constipated and also has early stage of kidney disease. She is on Lactulose now and pooping but her poop is as hard as a rock without Lactulose. Our vet recommended us to give MiraLax to her instead of Lactulose because it's easier to give and also it can be bought without prescription. The vet didn't tell us how much MiraLax to give her, so I searched online about dosage. This (screenshot below) is how the search result came up. I read the first link and gave her the dosage, but it turned out this dosage was about a different medicine! The Google result confused me! I should have searched more...

I thought I was started with medium dosage, 2 teaspoons a day. But it seems like it's like 4 times more than the max dosage of MiraLax for cats!!

It's ok if she just gets diarrhea, but she also has kidney problem, too, so I am really worried. What can I

how much miralax to give a cat



All Shopping Images News Videos More Settings Tools

About 59 thousand results in 0.64 seconds

Mix one to four teaspoons with your **cat's** food every 12 to 24 hours. - **Miralax** is another laxative and stool softener. Mix 1/4 tsp once a day with wet **cat** food.

[8 Ways to Help Your Constipated Cat | petMD](#)

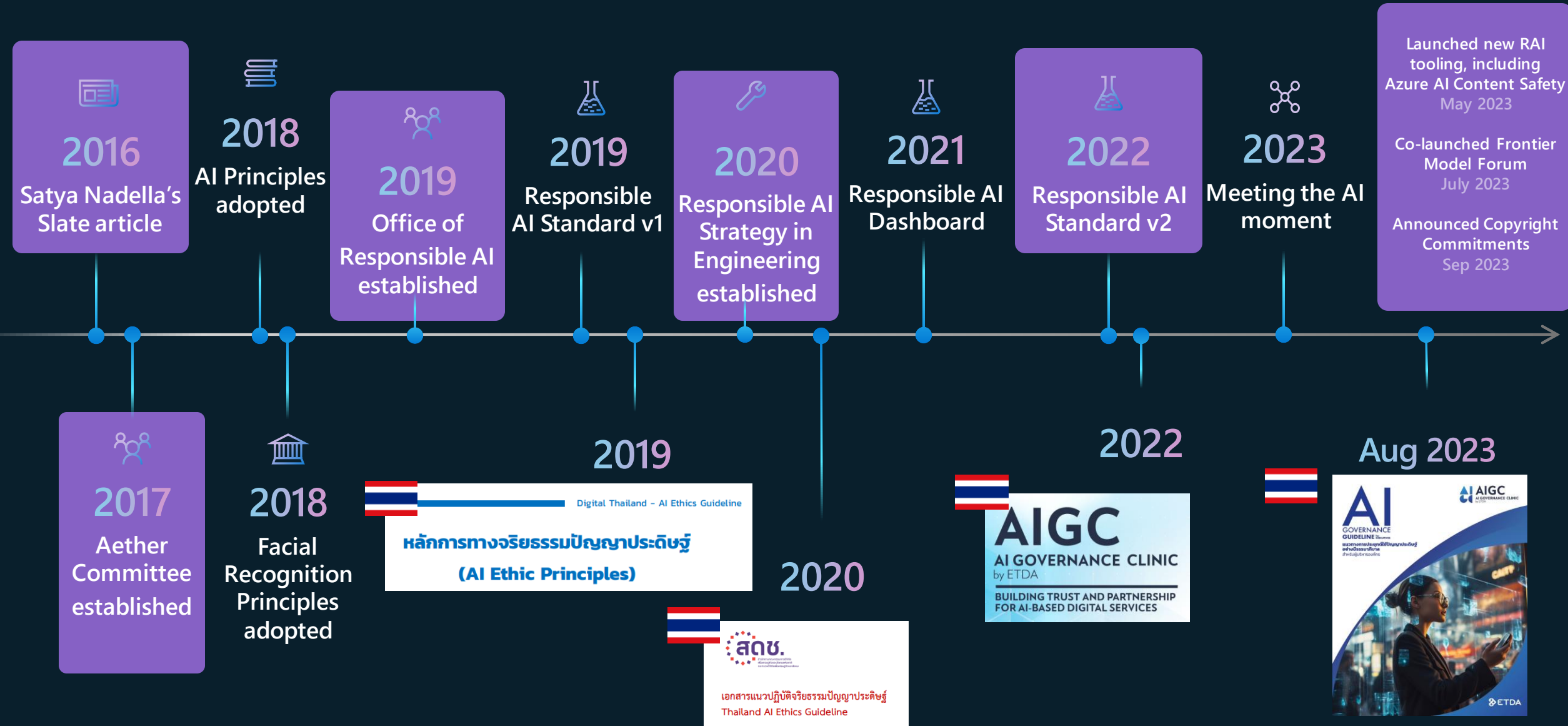
<https://www.petmd.com/cat/care/8-ways-help-your-constipated-cat>

Microsoft's Approach to AI

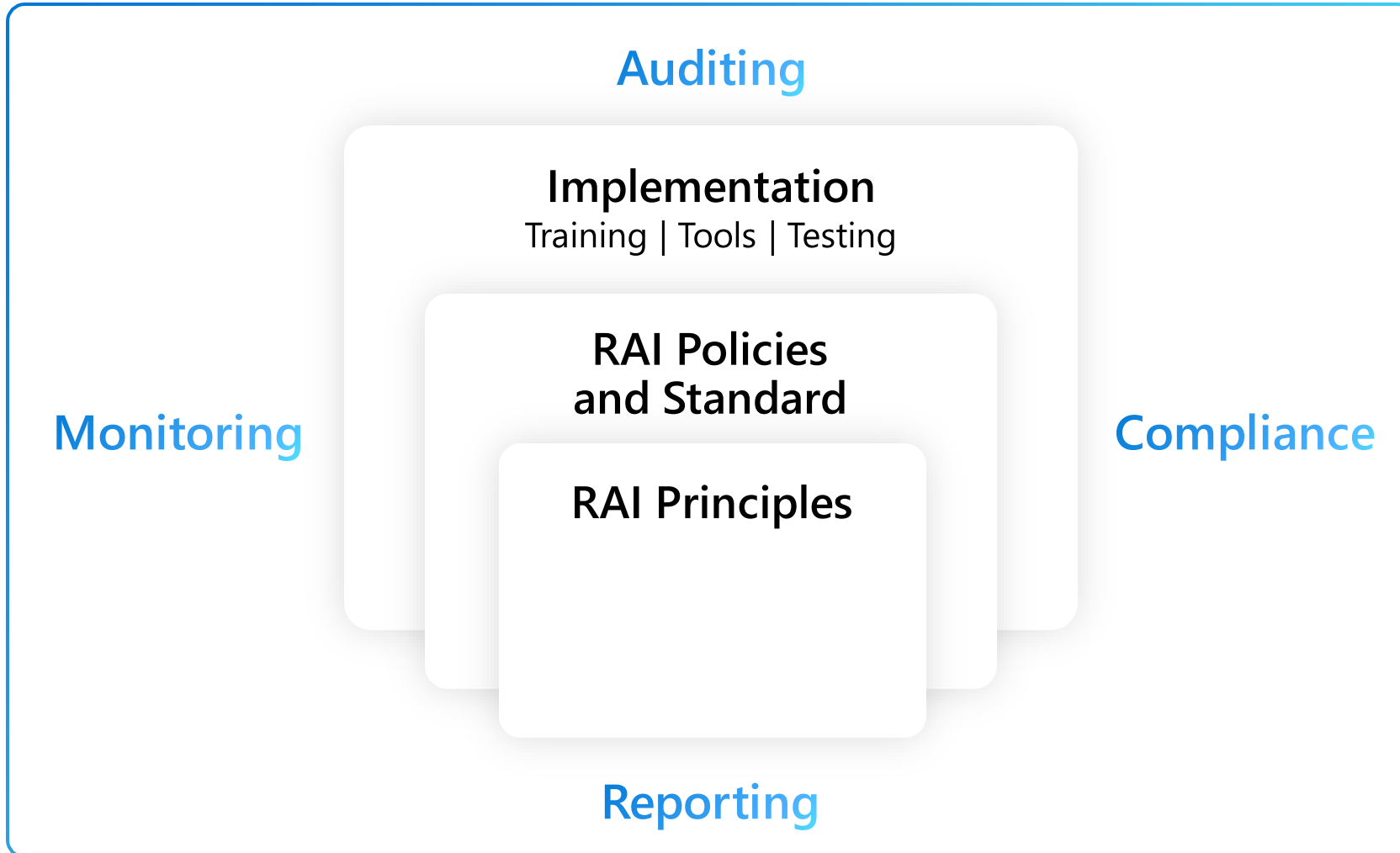


We have made huge investments in AI because we're **optimistic about what it can do to help people, industry and society**, and because we're committed to **bringing technology and people together to realize the promises of AI responsibly**.

Our Responsible AI Journey



Responsible AI Governance Framework



Microsoft's AI Principles



Fairness



Reliability
& Safety



Privacy &
Security



Inclusiveness



Transparency



Accountability

Our ecosystem



Sensitive Uses

A rule-making and oversight process



**Consequential
Impact**



**Physical or
Psychological Injury**



**Threat to
Human Rights**

[← Return to Blog Home](#)

Microsoft Research Blog

Advancing transparency: Updates on responsible AI research

Published January 10, 2024

By [Mihaela Vorvoreanu](#), Director, UX Research and Education, Aether; Kathy Walker, Allovus Design


Share this page



Editor's note: *All papers referenced here represent collaborations throughout Microsoft and across academia and industry that include authors who contribute to Aether, the Microsoft internal advisory body for AI ethics and effects in engineering and research.*

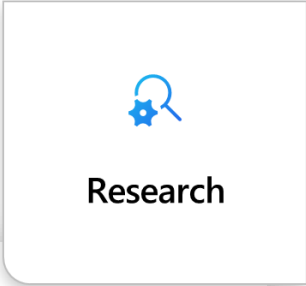
<https://www.microsoft.com/en-us/research/blog/advancing-transparency-updates-on-responsible-ai-research/>

Research Areas

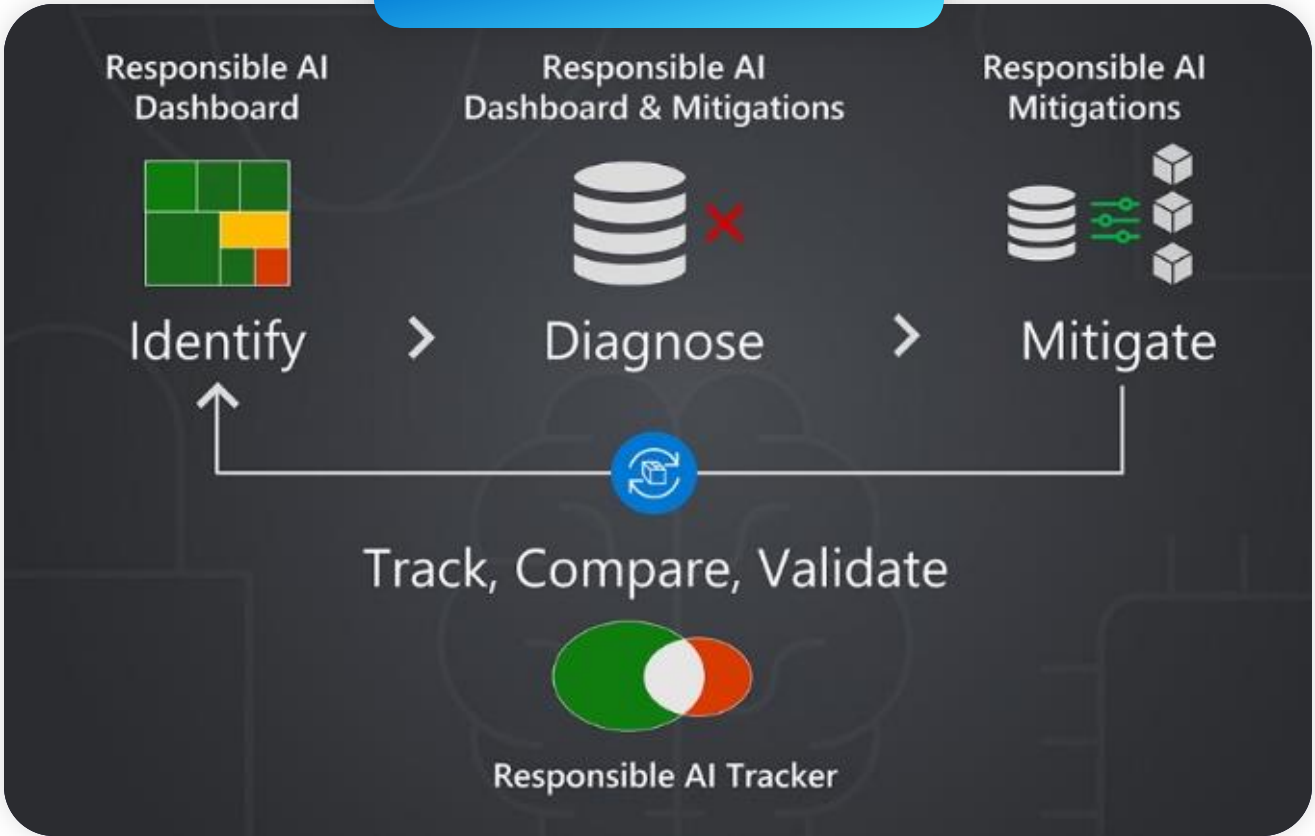
 Artificial intelligence

Tools & Practices:

Pioneering responsible AI practices



Responsible AI Toolbox



Guidelines for Human-AI Interaction

HAX is based on this set of foundational best practices for human interaction with AI systems.

[Learn more about guidelines >](#)

Initially	During interaction	When wrong	Over time
<p>1 Make clear what the system can do.</p>	<p>3 Time services based on context.</p>	<p>5 Match relevant social norms.</p>	<p>7 Support efficient invocation.</p>
<p>2 Make clear how well the system can do what it can do.</p>	<p>4 Show contextually relevant information.</p>	<p>6 Mitigate social biases.</p>	<p>8 Support efficient dismissal.</p>
			<p>9 Support efficient correction.</p>
			<p>10 Scope services when in doubt.</p>
			<p>11 Make clear why the system did what it did.</p>
			<p>12 Remember recent interactions.</p>
			<p>13 Learn from user behavior.</p>
			<p>14 Update and adapt cautiously.</p>
			<p>15 Encourage granular feedback.</p>
			<p>16 Convey the consequences of user actions.</p>
			<p>17 Provide global controls.</p>
			<p>18 Notify users about changes.</p>



Responsible AI Maturity Model

The Responsible AI Maturity Model (RAI MM) is a framework to help organizations identify their current and desired levels of RAI maturity. The RAI MM contains 24 empirically derived dimensions that are key to an organization's RAI maturity.





Governing AI: A Blueprint for the Future

May 25, 2023



Foreword: How Do We Best Govern AI?



Brad Smith, Vice Chair
and President, Microsoft

**“Don’t ask what computers can do, ask
what they should do.”**



Policy

Meeting the moment: combating AI deepfakes in elections through today's new tech accord

Feb 16, 2024 | [Brad Smith - Vice Chair & President](#)



Policy

Commitments to Help Combat Deceptive Use of AI in 2024 Elections

Addressing deepfake creation

- 1 Advance content authenticity through provenance and watermarking
- 2 Strengthen safety architecture for content creation tools

Detecting and responding to deceptive deepfakes

- 3 Detect the distribution of deepfakes
- 4 Address deepfakes that are detected, including by removing them
- 5 Share information and best practices across the tech sector

Transparency and resilience

- 6 Provide transparency to the public
- 7 Engage with civil society, academics, and experts
- 8 Foster public awareness and resilience



Engineering

Azure AI Content Safety Service

Detect and assign severity scores to unsafe content


Works on human/AI generated content

Available as a Service (API) and integrated across Azure AI

Get started with Azure AI Content Safety Studio

Run moderation tests


Explore, try out, and view sample code for different types of content.



Moderate text content

Run moderation tests on text contents. Assess the test results with detected severities. Experiment with different threshold levels.


[Try it out](#)



Moderate image content

Run moderation tests on image contents. Assess the test results with detected severities. Experiment with different threshold levels.

[Try it out](#)




Moderate multi-modal content

Coming soon

Run moderation tests on image and text combined contents. Assess the test results with detected severities.

What else would you like to do?


Monitor online activity and data on your own content source. Learn about how you will be able to build your own custom filter solution soon.



Monitor online activity

Stay on top of your Azure Content Safety usage with our real-time dashboard and gain immediate insights.

[Go to dashboard](#)



Build a custom solution

Coming soon

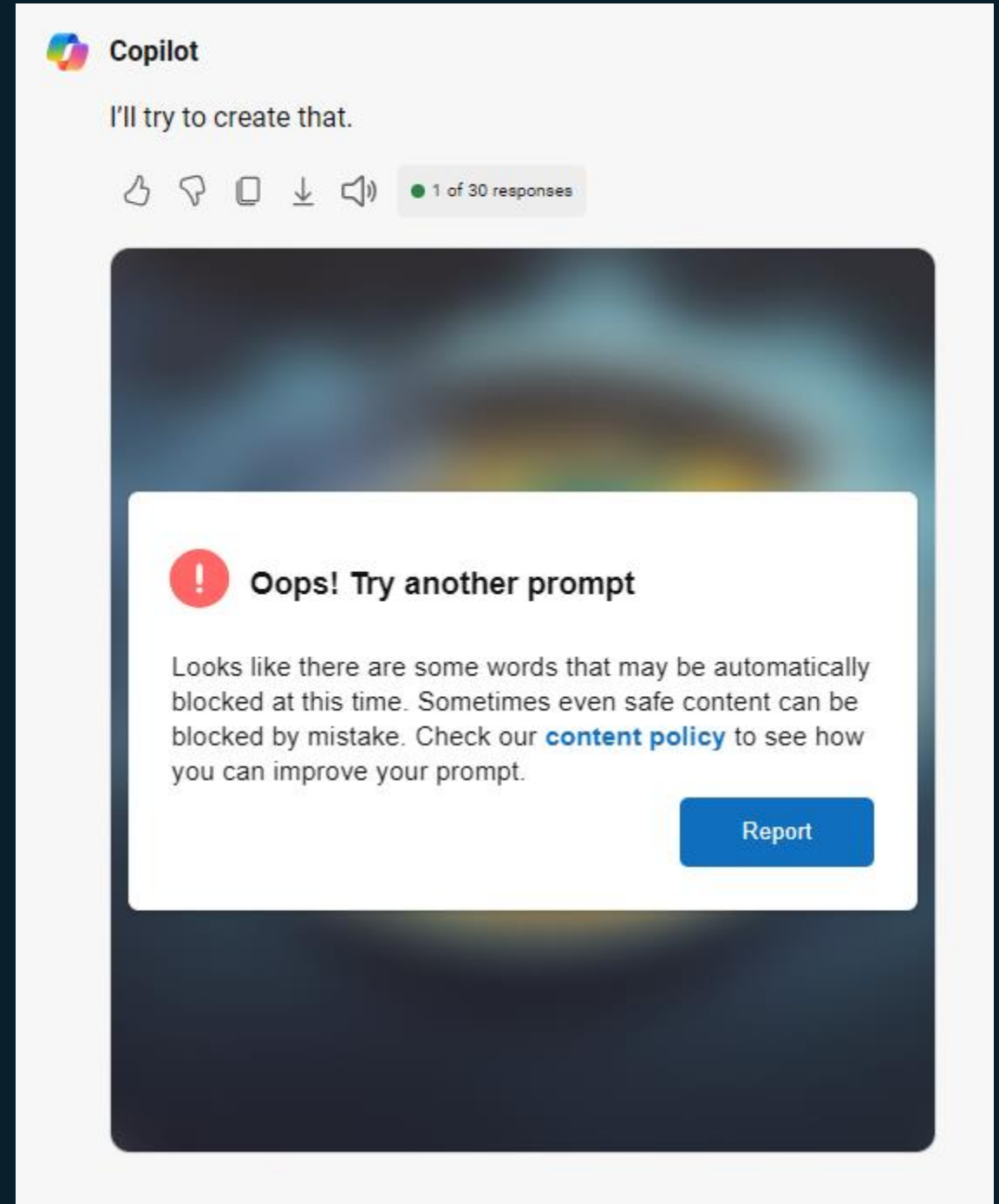
Customize the moderation levels and set up blocklist to provide a moderation experience that suits your needs.

Responsible AI in Content Safety

AI Content Safety

Harmful Content

Prompt: Create image of guy and girl in bikini



The screenshot shows the Microsoft Copilot interface. At the top, the Copilot logo is visible. Below it, the text "I'll try to create that." is displayed. Underneath, there are icons for thumbs up, thumbs down, copy, download, and a speaker icon, along with a "1 of 30 responses" indicator. The main content area is blurred, but a white error message box is overlaid in the center. The message box contains a red exclamation mark icon, the text "Oops! Try another prompt", and a paragraph explaining that some words may be automatically blocked. A blue "Report" button is located at the bottom right of the message box.

Copilot

I'll try to create that.

1 of 30 responses

Oops! Try another prompt

Looks like there are some words that may be automatically blocked at this time. Sometimes even safe content can be blocked by mistake. Check our [content policy](#) to see how you can improve your prompt.

Report

Jailbreak risk detection

Detect and filter User Prompts designed to provoke the Generative AI model into exhibiting behaviors it was trained to avoid or to break the rules set in the System Message



Optional filter in Azure
OpenAI Service



Feature in Azure AI Content Safety
and integrated across Azure AI

Protected material detection

Mitigation to defend customers against certain third-party intellectual property claims related to large language model outputs

Protected material detection for text

- Identifies text in language model output that matches known text content
- Example: song lyrics, articles, recipes, selected web content

Protected material detection for code

- Identifies source code in language model output that matches a set of source code from public repositories and retrieves citation and license information in annotations for the public repositories that contain those code snippets
- Example: public GitHub repository code

Announcing

Prompt Shield for Indirect Attacks

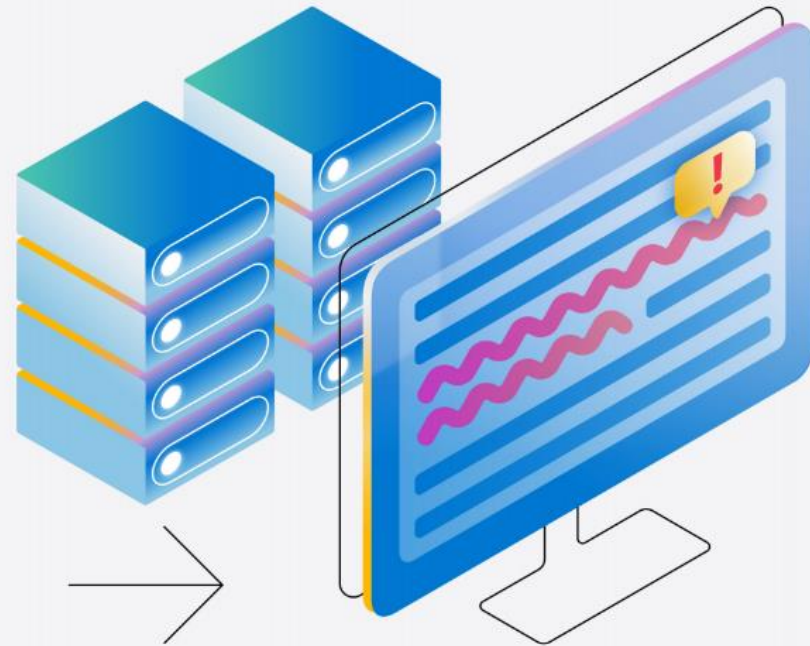
Detect and block prompt injection attacks hidden in data and grounding sources



Announcing

Groundedness detection

Detect and block model outputs that lack grounding



Responsible AI investments and safeguards for Facial Recognition

Published date: July 15, 2022

We're committed to helping developers and organizations use AI responsibly by protecting the rights and safety of customers. We're announcing three service updates to support the use of Azure Cognitive Services

2. Feature Retirements in Face API

We are retiring Face API attributes that predict emotion, gender, age, smile, facial hair, hair, and makeup. Read more about this decision [here](#).

We will also retire the Snapshot API, which allowed biometric data transfer from one Face subscription to another.

Existing customers have until **June, 30 2023** to use the emotion, gender, age, smile, facial hair, hair, and makeup attributes and the Snapshot API through Face API.

Three critical elements for progress

1. Leadership must be committed and involved
2. Build inclusive governance models and actionable guidelines
3. Invest in and empower your people

Learn more about Microsoft's approach to responsible AI

Learn more about AI Principles and explore resources like the Responsible AI Impact Assessment Guide

[Responsible AI Practices](#)

Skill up with self-guided and instructor-led courses

Visit Microsoft Learn and find courses on generative AI for business leaders, developers, and ML professionals

[AI Learning HUB](#)

Try responsible AI tools and services for free in Azure

Sign up for Azure and start using the model catalog and prompt flow in Azure AI

[Azure AI](#)

Thank You