Microsoft Azure

AI-powered content safety

Build a leading content moderation strategy with AI

Azure AI Content Safety

User-generated content is at the core of many digital platforms

Across industries, more organizations are offering social features for users to interact with each other and with brands. This is true of social media, and it is becoming increasingly relevant in industries such as gaming, ecommerce and advertising. This trend may be largely driven by the fact that **user-generated content is the most trustworthy form of content** according to consumers.[1] By providing a platform for customers and users to engage, companies can build brand connection and awareness, promote trust and authenticity, and drive digital engagement. But if users have a negative experience on a digital platform, they are quick to leave—**in response to being harassed, almost a third of users (29%) stop or reduce their use of platforms altogether**. And the number of Americans who have experienced online harassment has been hovering around 42% since 2020.[2]

Hosting user-generated content is becoming increasingly fraught

Growth of inappropriate content

The amount of hate speech, bullying, and child sexual abuse material (CSAM) posted to platforms every day is growing.[3,4]

Regulatory pressures

Governments are passing new regulations on content. Since 2021, over 250 US state bills have been introduced to regulate online content.[5]

Demands for transparency

Users are demanding transparency around content moderation standards and enforcement procedures.[6]

Surge of complex content

Content forms like memes, live chat, and text-on-video push the boundaries of moderation capabilities.[7]

The new reality of generative AI

Content on digital platforms will be reshaped by generative AI. It can be used to quickly produce high-quality content and speed time to value, but if used unethically and irresponsibly, generative AI could manipulate online reviews or risk propagation of fake identities and information.[8]

It's no surprise that content safety is becoming a top priority

Traditional approaches to content moderation face known challenges

Lack semantic understanding

Most moderation tools are based on keyword identification and pattern matching and aren't capable of grasping nuance or context.

Most models are not multi-lingual

Many moderation tools are less effective in languages apart from English—limiting their ability to help companies moderate content in languages around the world.

Resulting in more content that requires human review

The human toll of moderation increases

Content moderation teams are essential to online safety, but the surging amount of content adds to the psychologically taxing nature of moderating for violent, hateful, and inappropriate content.

The cost of digital safety grows

The cost of scaling human content moderation globally to fill the gap left by today's tools can be prohibitive—leading organizations to make difficult choices around safety, resources, and user-generated content.

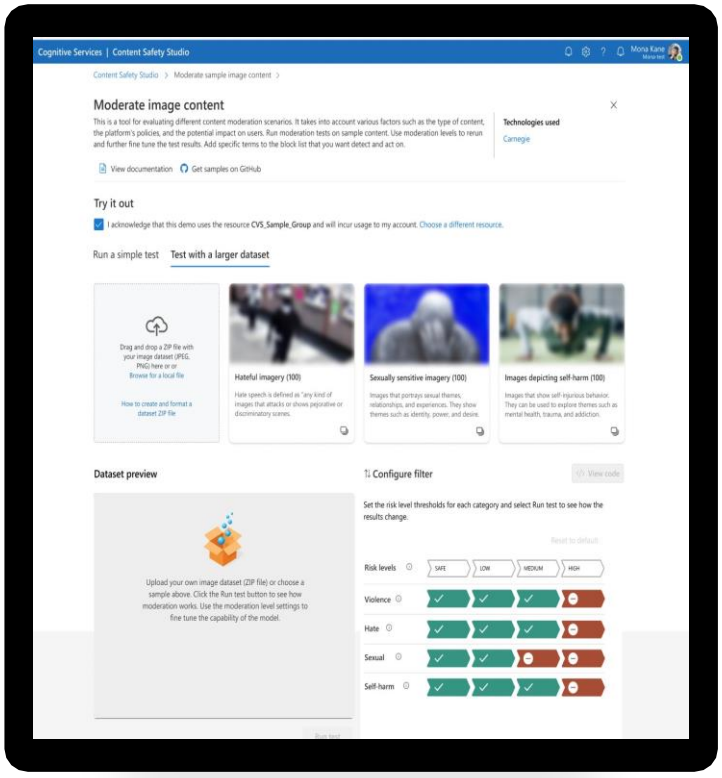
AI offers a more sophisticated approach to content moderation

AI models can better understand context and nuance—making AI able to review content in ways that are much closer to human review. But like all new and powerful technology, AI can be used for both desirable and undesirable purposes and, if left unchecked, AI can have unintended outcomes. **That's why responsibility must be at the heart of any AI-based technology used for content moderation.**

Introducing Azure AI Content Safety

Azure AI Content Safety uses AI to help you create safer online spaces. With cutting edge AI models, it can detect hateful, violent, sexual, and self-harm content and assign it a severity score, allowing businesses to prioritize what content moderators review. Unlike most solutions used today, **Azure AI Content Safety can handle nuance and context**, which eases the load on human content moderator teams.

Azure AI Content Safety isn't one-size-fits-all—**it can be customized to help businesses implement their policies.** Plus, its multi-lingual models enable it to **understand many languages simultaneously.**



1

Azure AI Content Safety classifies harmful content into four categories:

Hate Sexual Self-harm Violence

2

Next, it returns a severity level for each category from 0 – 6:

Hate: 0 – **2** – 4 – 6
Sexual: **0** – 2 – 4 – 6
Self-harm: 0 – 2 – **4** – 6
Violence: **0** – 2 – 4 – 6

3

Then, it surfaces content based on the severity level:

High risk: Auto blocked

Medium risk: Sent to moderator and prioritized by risk level, topic, and user reputation

Low risk: Auto approved

Microsoft's commitment to responsible AI

At Microsoft, we are focused on creating advanced AI tools that drive responsible transformation. **Azure AI Content Safety is one of many AI solutions that Microsoft is building to help address pressing societal needs.**

We believe that AI can be harnessed for good if we are committed to a people-centric approach, if we engage with industry working groups to align AI with the world's greatest needs, and if we remain dedicated to providing transparency about how our AI technology works.

With Azure AI Content Safety, we are holding fast to our responsible AI principles and using this groundbreaking technology to help promote safety online. In fact, we are already leveraging Azure AI Content Safety in Microsoft products such as Azure OpenAI, Azure ML, GitHub Copilot, Bing, and more to help detect potentially harmful content. We are continuing to expand the reach of this technology so it can help provide users across all kinds of platforms with safer experiences.

Azure AI Content Safety in generative AI

Azure AI Content Safety classification models power Azure OpenAI content filters—enabling the filters to identify and flag harmful content.

The user-generated prompt and the AI-generated response are both analyzed. If the system detects violating material in either, the generation process stops, and an error message is created.

Our responsible AI principles

Fairness
AI systems should treat 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

Learn more about **Microsoft's commitment to responsible AI**

Analyze text with semantic understanding and multi-lingual models

With semantic understanding and natural language processing techniques, Azure AI Content Safety can address the meaning and context of language in ways that more closely match the accuracy of human review, and it can do so in multiple languages. Azure AI Content Safety can analyze multilingual text in both short form and long form to detect and flag harmful content.

Industry spotlight: Social media

As new social media platforms emerge and expand, content moderation solutions are key to longevity and keeping users happy. Social media moves fast, and it's essential that content moderation tools can act in real time. Azure AI Content Safety can monitor content in posts, threads, and chats to help keep online social communities protected from harmful content.

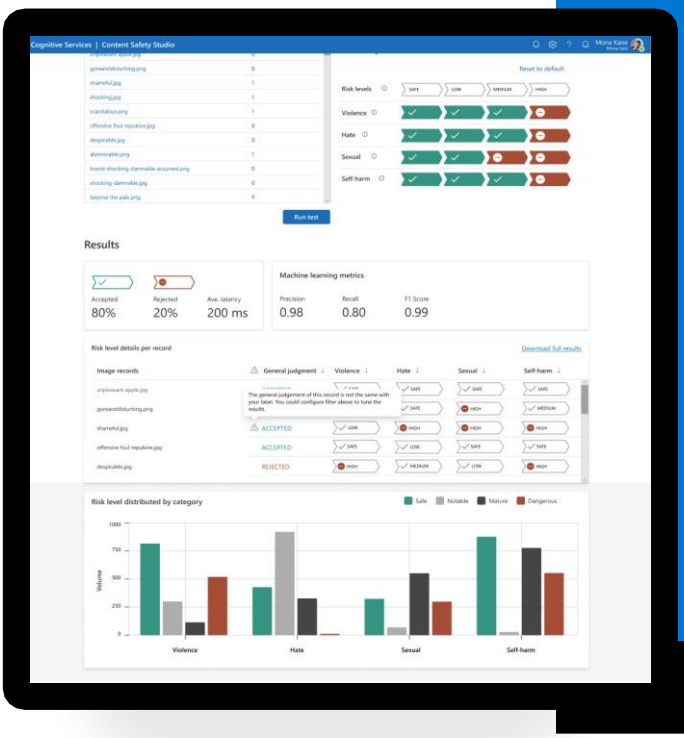
Detect images using vision models

Industry spotlight: Gaming

Content moderation for gaming is challenging due to visual and text-based elements, live chat, and at times violent content. Azure AI Content Safety's advanced computer vision tools help monitor avatars, usernames, images, and chat, making platforms safer and supporting human moderators.

In an increasingly visual culture, computer vision models are a cornerstone of content moderation. Azure AI Content Safety is powered by Microsoft's Florence foundation model, which advances state-of-the-art computer vision technologies. Florence has been trained with billions of text-image pairs, and it can be easily adapted for various computer vision tasks—enabling Azure AI Content Safety to identify explicit images in real time to minimize the load on moderators.

Meet with an **Azure AI sales specialist** about how you can monitor text and image content with Azure AI Content Safety



> “Azure AI Content Safety allows us to develop workflows that are as diverse as the communities, languages and regions we serve across the world, enabling our users to safely express themselves and interact with their favorite public figures and celebrities on the platform.”
>
> —Aprameya Radhakrishna, CEO and Cofounder, KOO

Improve retention and increase engagement

User-generated content is important to user engagement, but it can also lead to a platform exodus if safety isn’t property addressed. Azure AI Content Safety can help reduce the presence of harmful content to help create a safer environment for users.

Reduce human toll of moderation and speed time to resolution

Moderation needs intensify when user-generated content grows on digital platforms. Azure AI Content Safety can help reduce a human moderator’s exposure to harmful content by more accurately identifying inappropriate content and supporting faster takedown.

Learn more about the benefits of implementing Azure AI Content Safety by talking to an **Azure AI sales specialist**

Microsoft Azure

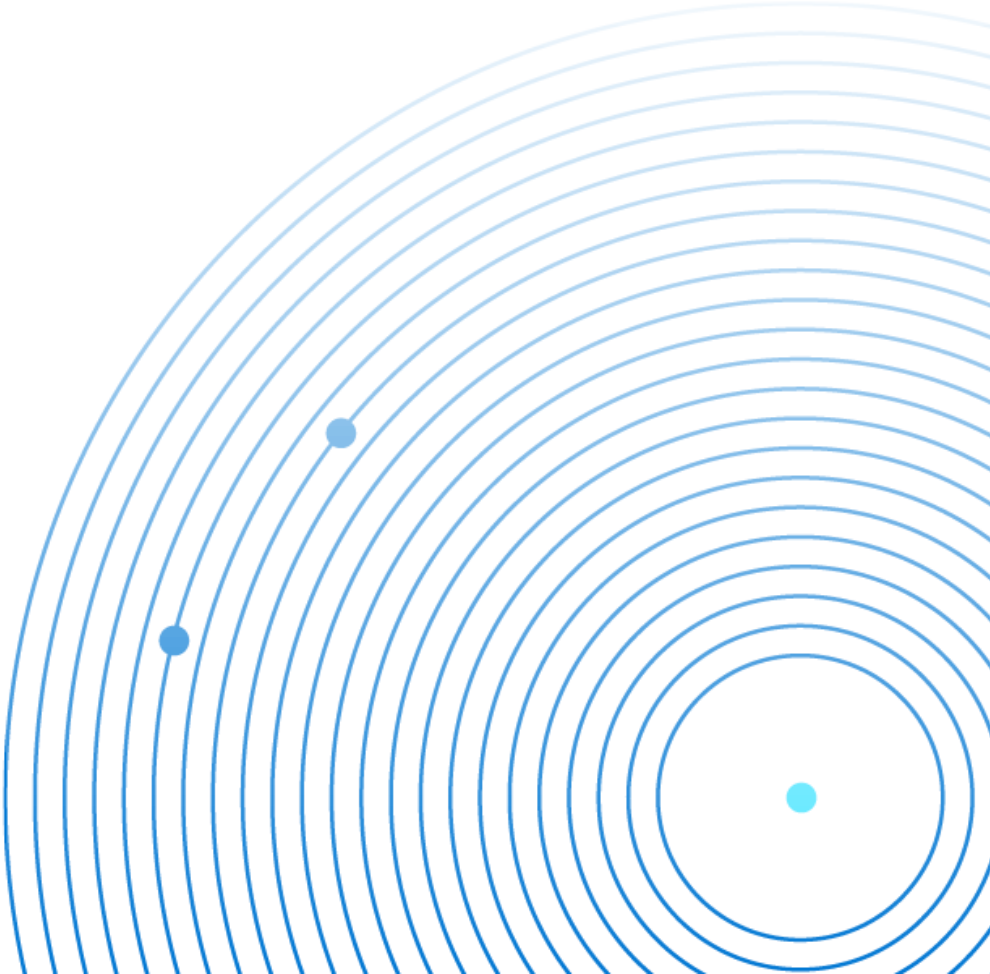
Learn more about how Azure AI Content Safety can help create safer communities online:

Get started

- Sign up to get started for free
- Try it yourself in the Azure AI Content Safety Studio
- Speak with an Azure AI sales specialist

Learn more

- Discover more on the product page
- Review the documentation
- Watch the demo video
- Read the announcement blog



1. The state of social & user-generated content | TINT, 2023
2. Online hate and harassment | ADL, 2022
3. Hate speech's rise on Twitter is unprecedented | The New York Times, 2022
4. CSAM annual report | IWF, 2022
5. State content moderation landscape and look to 2023 | DISCO, 2022
6. Transparency is essential for effective content moderation | Brookings, 2022
7. Content moderation – future trends and challenges | Imagga, 2022
8. IDC, Generative Artificial Intelligence: A New Chapter for Enterprise Business Applications, doc #US50471523, March 2023