Microsoft

Transparency Note: Azure Content Safety

Updated <04/04/2023>

Table of Contents

What is a Transparency Note?

An AI system includes not only the technology, but also the people who will use it, the people who will be affected by it, and the environment in which it is deployed. Creating a system that is fit for its intended purpose requires an understanding of how the technology works, what its capabilities and limitations are, and how to achieve the best performance. Microsoft's Transparency Notes are intended to help you understand how our AI technology works, the choices system owners can make that influence system performance and behavior, and the importance of thinking about the whole system, including the technology, the people, and the environment. You can use Transparency Notes when developing or deploying your own system, or share them with the people who will use or be affected by your system.

Microsoft's Transparency Notes are part of a broader effort at Microsoft to put our AI Principles into practice. To find out more, see the Microsoft AI principles.

The basics of Azure Content Safety

Introduction

Azure Content Safety is a part of Azure Cognitive Services that detects material that is potentially offensive, risky, or otherwise undesirable. Azure Content Safety works on new functionalities that offer state-of-the-art text, image, and multimodal APIs and Interactive Studio that detects problematic content. Azure Content Safety helps make applications and services safer by redacting harmful user-generated and AI-generated content.

Key terms

Content categories

Category	**Description**
Hate	*Hate* refers to any content that attacks or uses pejorative or discriminatory language with reference to a person or identity group on the basis of certain differentiating attributes including but not limited to race, ethnicity, nationality, gender identity and expression, sexual orientation, religion, immigration status, ability status, personal appearance, and body size.
Sexual	*Sexual* describes language that is related to anatomical organs and genitals; romantic relationships; acts portrayed in erotic or affectionate terms; pregnancy; physical sexual acts, including those portrayed as an assault or as a forced sexual, violent act against one's will; prostitution; pornography; and abuse.
Violence	*Violence* describes language that is related to physical actions that are intended to hurt, injure, damage, or kill someone or something; or describes weapons, guns, and related entities, such as manufacturers, associations, or legislation.
Self-Harm	*Self-harm* describes language that is related to physical actions that are intended to purposely hurt, injure, or damage one's body or to kill oneself.
Severity levels	Every content flag the service applies also comes with a risk level rating. The risk level is meant to indicate the severity of the consequences of showing the flagged content.

Category	Description
	Severity Label *0 Safe* *2 Low* *4 Medium* *6 High*

Capabilities

System behavior

NEED ADDITIONAL CONTENT HERE – PLEASE REVIEW SOME EXAMPLES SUCH AS AZURE OPEN AI TN: Use cases for Azure OpenAI - Azure Cognitive Services | Microsoft Learn

Best practices for improving system performance

Show and tell when designing prompts. With text and code models, make it clear to the model what kind of outputs you expect through instructions, examples, or a combination of the two. If you want the model to rank a list of items in alphabetical order or to classify a paragraph by sentiment, show it that's what you want.

Keep your application on topic. Carefully structure prompts and image inputs to reduce the chance of producing undesired content, even if a user tries to use it for this purpose. For instance, you might indicate in your prompt that a chatbot only engages in conversations about mathematics and otherwise responds "I'm sorry. I'm afraid I can't answer that." Adding adjectives like "polite" and examples in your desired tone to your prompt can also help steer outputs. With image models, you might indicate in your prompt or image input that your application generates only conceptual images. It might otherwise generate a pop-up notification that explains that the application is not for photorealistic use or to portray reality. Consider nudging users toward acceptable queries and image inputs, either by listing such examples up front or by offering them as suggestions upon receiving an off-topic request. Consider training a classifier to determine whether an input (prompt or image) is on topic or off topic.

Provide quality data. With text and code models, if you're trying to build a classifier or get the model to follow a pattern, make sure that there are enough examples. Be sure to proofread your examples—the model is usually smart enough to see through basic spelling mistakes and give you a response, but it also might assume this is intentional and it could affect the response. Providing quality data also includes giving your model reliable data to draw responses from in chat and question answering systems.

Measure model quality. As part of general model quality, consider measuring and improving fairness-related metrics and other metrics related to responsible AI in addition to traditional accuracy measures for your scenario. Consider resources like this checklist when you measure the fairness of the system. These measurements come with limitations, which you should acknowledge and communicate to stakeholders along with evaluation results.

Limit the length, structure, and rate of inputs and outputs. Restricting the length or structure of inputs and outputs can increase the likelihood that the application will stay on task and mitigate, at least in part, any potentially unfair, unreliable, or offensive behavior. Other options to reduce the risk of misuse include (i) restricting the source of inputs (for example, limiting inputs to a particular domain or to authenticated users rather than being open to anyone on the internet) and (ii) implementing usage rate limits.

Encourage human review of outputs prior to publication or dissemination. With generative AI, there is potential for generating content that might be offensive or not related to the task at hand, even with mitigations in place. To ensure that the generated output meets the task of the customer, consider building ways to remind customers to review their outputs for quality prior to sharing widely. This can reduce many different harms, including offensive material, disinformation, and more.

Implement additional scenario-specific mitigations. Refer to the mitigations outlined in Evaluating and integrating Azure OpenAI for your use including content moderation strategies. These do not represent every mitigation that might be required for your application, but they point to the general minimum baseline we check for when approving use cases for Azure OpenAI Service.

See also the template for structure and what should be covered in this section. You can also go into more detail about the severity levels.

Different types of analysis are available in Azure Content Safety:

Type	**Functionality**
Text Detection API	Scans text for hate, sexual, violence, and self-harm content, with multi-severity risk levels.
Image Detection API	Scans images for hate, sexual, violence, and self-harm content, with multi-severity risk levels.
Multimodal Detection API	Scans both images and text (including separate text or text from OCR from an image) for hate content, with multi-severity risk levels.
Azure Content Safety Studio	Azure Content Safety Studio is an online tool that you can use to visually explore, understand, and evaluate the Azure Content Safety service. The studio provides a platform for you to experiment with the different Azure Content Safety classifications and to interactively sample returned data without writing any code.

Use cases

Intended uses cases

Azure Content Safety can be used in multiple scenarios. The system's intended uses include:

- **Social media platforms:** Content safety systems are commonly used by social media customers to prevent the spread of harmful and inappropriate content, such as hate speech, cyberbullying, and pornography.
- **E-commerce websites:** E-commerce customers use Azure Content Safety ~~content safety systems~~ to screen product listings and reviews for inappropriate content, such as fake reviews and offensive language.
- **Gaming platforms:** Gaming platforms use content safety systems to detect cheating, hacking, and other forms of misconduct, as well as to prevent inappropriate behavior in chats and forums.
- **News websites:** Content safety systems are used by news websites to ensure that user comments remain civil and respectful, and to prevent the spread of fake news and hate speech.
- **Video-sharing platforms**: Video-sharing platforms use content safety systems to detect and remove inappropriate content, such as violence, hate speech, and pornography.

Considerations when choosing other use cases

We encourage customers to leverage Azure Content Safety in their innovative solutions or applications. However, here are some considerations when choosing a use case:

- **Compliance:** Depending on the nature of your application or solution, you might need to comply with certain regulations or standards that are related to content safety, including the Children's Online Privacy Protection Act (COPPA) or the General Data Protection Regulation (GDPR).
- **Customization:** Different applications and solutions might have different requirements when it comes to content safety. It's important to choose a solution that can be customized to meet your specific needs and that can integrate with your existing workflows and processes. Azure Content Safety may allow you to set the threshold for detection, help you mitigate some risks, however, we have the limitation for customization.
- **Transparency:** Some of your users might want to understand how your application or solution moderates content. When you choose to use Azure Content Safety, it's important to ensure that the service provides transparency and clear communication with your users about how content is moderated and why certain content might be flagged or removed.

Unsupported uses

- **Illegal activities**: Azure Content Safety should not be used to support or facilitate illegal activities, such as the distribution of child pornography or the promotion of hate speech.

Limitations

Technical limitations, operational factors, and ranges

Technical limitations: Content safety systems have some technical limitations that can affect their effectiveness. Some of these limitations are:

Accuracy: ~~Our~~ systems are not always 100% accurate, and there is a risk of false positives or false negatives. When you choose to use Azure Content Safety, it's important to evaluate its accuracy and to ensure that it meets your specific needs.

Language barriers: Content safety systems might not be able to detect inappropriate content in languages that they are not programmed to understand.

Image recognition: Content safety systems might not be able to detect inappropriate content in images that are not clear or that have been edited.

Evolving nature of content: Content safety systems might struggle to keep up with the evolving nature of online content and as new types of inappropriate content emerge.

Operational factors:

Content safety systems also have some operational factors that need to be considered for their effectiveness. Some of these factors are:

Volume of content: Content safety systems might struggle to handle large volumes of content. This can lead to delays in detecting inappropriate content.

Time sensitivity: Some types of inappropriate content require immediate action. Content safety systems might be unable to identify these types of content quickly and alert moderators.

Contextual analysis: Content safety systems might be unable to analyze content in context to determine whether it is inappropriate. For example, certain words might be appropriate in some contexts but not in others.

Ranges for content safety systems:

Content safety systems can cover a range of content. For example:

Text-based content: This includes social media posts, comments, and messages.

Image-based content: This includes photos and videos.

System performance

NEED ADDITIONAL CONTENT HERE TO HELP THE READER UNDERSTAND WHAT THIS MEANS FOR YOUR SYSTEM – PLEASE REVIEW SOME EXAMPLES SUCH AS SPEAKER RECOGNITION Characteristics and limitations of Speaker Recognition - Azure Cognitive Services | Microsoft Learn

See also the template for structure and what should be covered in this section.

Error Type	**Definition**	**Example**
True Positive	The model correctly identifies inappropriate content.	A social media post that contains hate speech is flagged by the model and removed.
False Positive	The model incorrectly identifies appropriate content as inappropriate.	A social media post that discusses a political topic is flagged as inappropriate due to the presence of certain keywords.

Error Type	Definition	Example
True Negative	The model correctly identifies appropriate content.	A social media post that contains a picture of a dog is not flagged as inappropriate by the model.
False Negative	The model fails to identify inappropriate content.	An image that contains nudity is not flagged as inappropriate by the model .

Best practices for improving system performance

Do:

- Monitor the system's performance regularly to ensure that the tradeoff is appropriate for the use case. For example, tradeoff for precision and recall.
- Adjust the risk levels for blocking based on user feedback and observed trends in content safety.
- Consider the impact of the system's performance on different user populations and adjust accordingly.For example, in gaming industry, the adult plan and the family plan for content safety may different.
- Take steps to mitigate any unintended consequences of adjusting the risk levels for blocking, such as over-removal of content or the spread of harmful content.

Don't:

- Set the risk levels for blocking too high, which might lead to a large number of false positives, which can negatively affect user experience.
- Set the risk levels for blocking too low, which might lead to a large number of false negatives, which can harm users and communities.
- Ignore feedback from users and communities about the system's performance.
- Over-rely on the system's automated decision-making capabilities without ensuring appropriate human oversight and intervention.

Evaluation of Azure Content Safety

Evaluation methods

The methods that are used to evaluate a system for content safety considerations typically involve analyzing large datasets of harmful content and evaluating the system's ability to accurately identify and flag potentially harmful or inappropriate content.

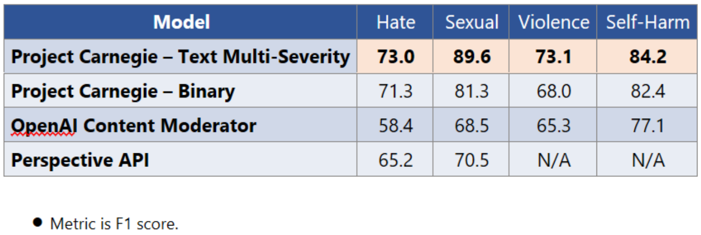
It’s important to evaluate how the system performs across different demographic and geographic areas. The groups of people that are included in the evaluation depend on the type of content that is being evaluated and the intended audience of the system. For example, if the system is designed to monitor social media posts, the dataset might include a diverse range of users from various geographic locations, backgrounds, and age groups. However, if the system is designed for use in a specific industry or niche market, the dataset might be limited to users who are in that specific group.

The evaluation itself might involve a combination of automated testing and manual review by content safety experts to ensure that the system is effectively identifying potentially harmful or inappropriate content. The results of the evaluation are then used to improve the system and to optimize its performance for real-world use.

Evaluation results

~~We~~ found that the system was able to accurately identify and flag potentially harmful or inappropriate content, and it was effective in a variety of settings and use cases.

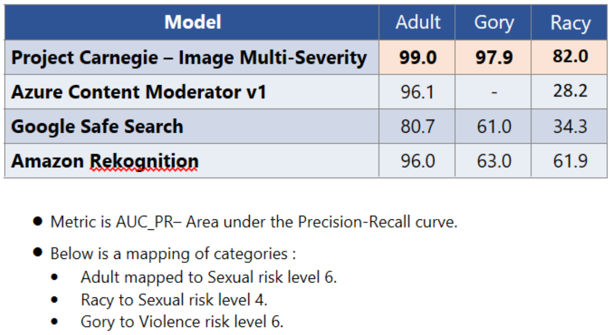
Text multi-severity model:



Model	Hate	Sexual	Violence	Self-Harm
Project Carnegie – Text Multi-Severity	**73.0**	**89.6**	**73.1**	**84.2**
Project Carnegie – Binary	71.3	81.3	68.0	82.4
OpenAI Content Moderator	58.4	68.5	65.3	77.1
Perspective API	65.2	70.5	N/A	N/A

- Metric is F1 score.

Image multi-severity model:



Model	Adult	Gory	Racy
Project Carnegie – Image Multi-Severity	**99.0**	**97.9**	**82.0**
Azure Content Moderator v1	96.1	-	28.2
Google Safe Search	80.7	61.0	34.3
Amazon Rekognition	96.0	63.0	61.9

- Metric is AUC_PR– Area under the Precision-Recall curve.
- Below is a mapping of categories :
  - Adult mapped to Sexual risk level 6.
  - Racy to Sexual risk level 4.
  - Gory to Violence risk level 6.

~~Fairness considerations~~

~~<If any of the Fairness Goals from RAIS v2 applies to this system, include this section. Otherwise, you do not need to include this section.~~

~~Two or three paragraphs:~~

- ~~A summary of our evaluation of the system in relation to potential fairness harms. A disaggregated analysis of results for different demographic groups, identified demographic groups for which performance may not meet any target minimum performance level, any remaining performance disparities or differences between the rates at which resources and opportunities are allocated between identified demographic groups that may exceed the target, any justifiable factors that account for these performance levels and differences. See requirements F1.9, F2.9 and F3.7 in RAIS v2.~~
- ~~Disclosure of risks related to representational harms (stereotyping, demeaning, and erasing outputs). When possible, propose mitigations people can put in place to address these risks (e.g., consider your use case carefully, have a human in the loop, use a blocklist, use an allow list, etc.).~~
- ~~If Fairness evaluations are not very complete, language such as "What we know so far, with regard to Fairness evaluations" can be used.>~~

Evaluating and integrating Azure Content Safety for your use

- **Appropriate human oversight for the system is critical to ensure that it is being used effectively and responsibly.** This includes ensuring that the people who are responsible for oversight understand the system's intended uses, how to interact with the system effectively, how to interpret system behavior, and when and how to intervene in or override the system. Considerations such as UX and UI design and the use of risk levels can inform human oversight strategies and help prevent over-reliance on system outputs. For example, for a product like a content safety system, it is important to provide content moderators with the training and resources that they need to effectively oversee the system. This might involve providing access to training materials and documentation, as well as ongoing support from content safety experts.
- **Remind users that they are accountable for final decisions and final content.**
- **Establish feedback channels for users and affected groups.** AI-powered products and features require ongoing monitoring and improvement. Establish channels to collect questions and concerns from users as well as from people who are affected by the system. For example, build feedback features into the user experience. Invite feedback on the usefulness and accuracy of outputs, and give users a separate and clear path to report outputs that are problematic, offensive, biased, or otherwise inappropriate.

Learn more about responsible AI

Microsoft AI principles

Microsoft responsible AI resources

Microsoft Azure Learning courses on responsible AI

Learn more about Azure Content Safety

<Insert links here to relevant resources for learning more about the product or feature including contracts (OST/trust center), marketing materials, and technical documentation.>

Contact us

Give us feedback on this document <document author: link to an appropriate feedback channel such as an alias with several team or division members or a help email for the product or feature>

About this document

© <year> Microsoft Corporation. All rights reserved. This document is provided "as-is" and for informational purposes only. Information and views expressed in this document, including URL and other Internet Web site references, may change without notice. You bear the risk of using it. Some examples are for illustration only and are fictitious. No real association is intended or inferred.

Published: 04/04/2023

Last updated: 04/04/2023