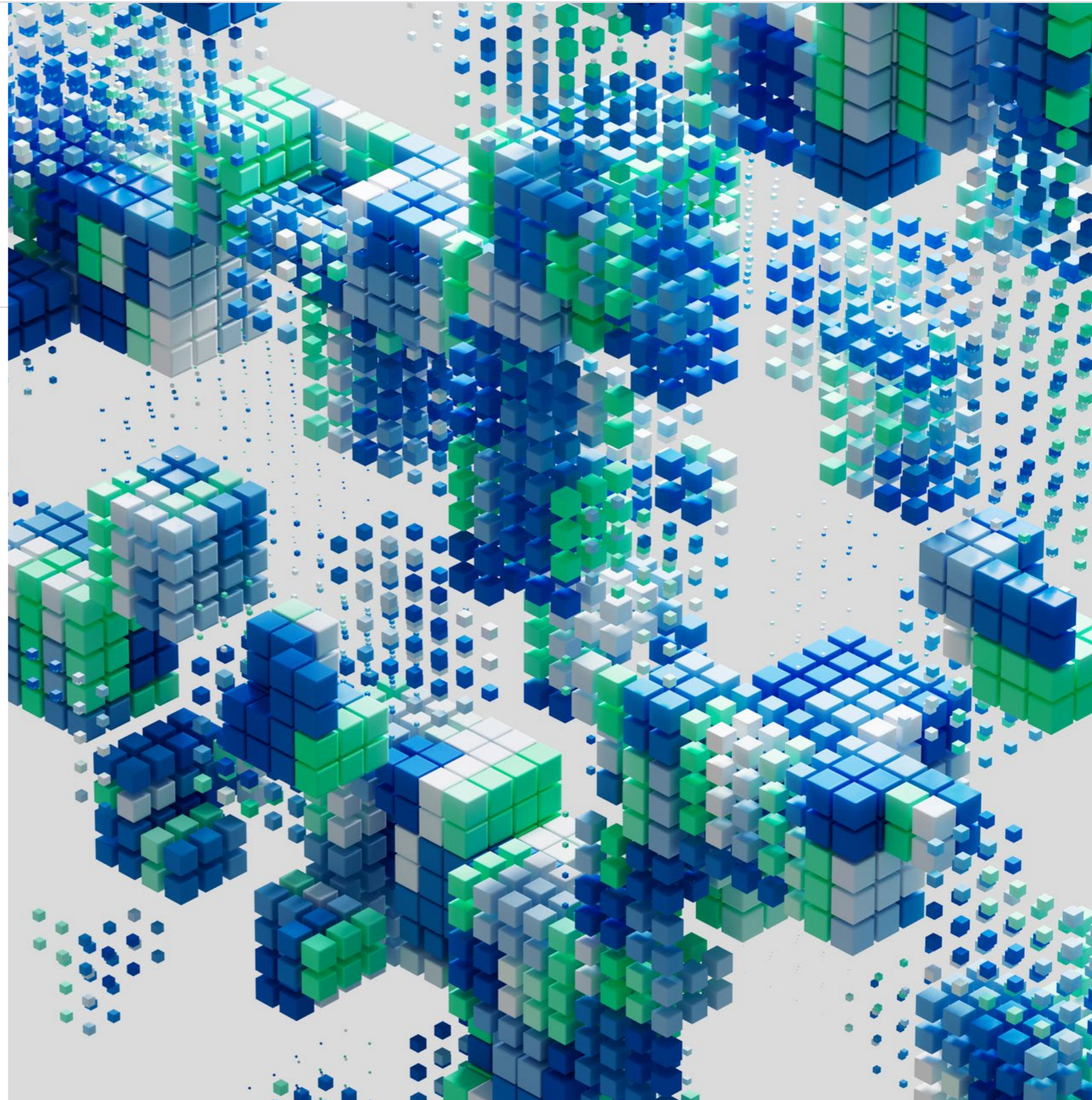




# Responsible AI Progress Report

Published in February 2025



# Foreword

AI is a transformational technology that offers both a unique opportunity to meet our mission, and the chance to expand scientific discovery and tackle some of the world's most important problems. At Google we believe it's crucial that we continue to develop and deploy AI responsibly, with a focus on making sure that people, businesses, and governments

around the world can benefit from its extraordinary potential while at the same time mitigating against its potential risks.

In 2018, we were one of the first in the industry to adopt [AI Principles](#), and since then, we've published annual AI responsibility reports detailing our progress. This year's report shares

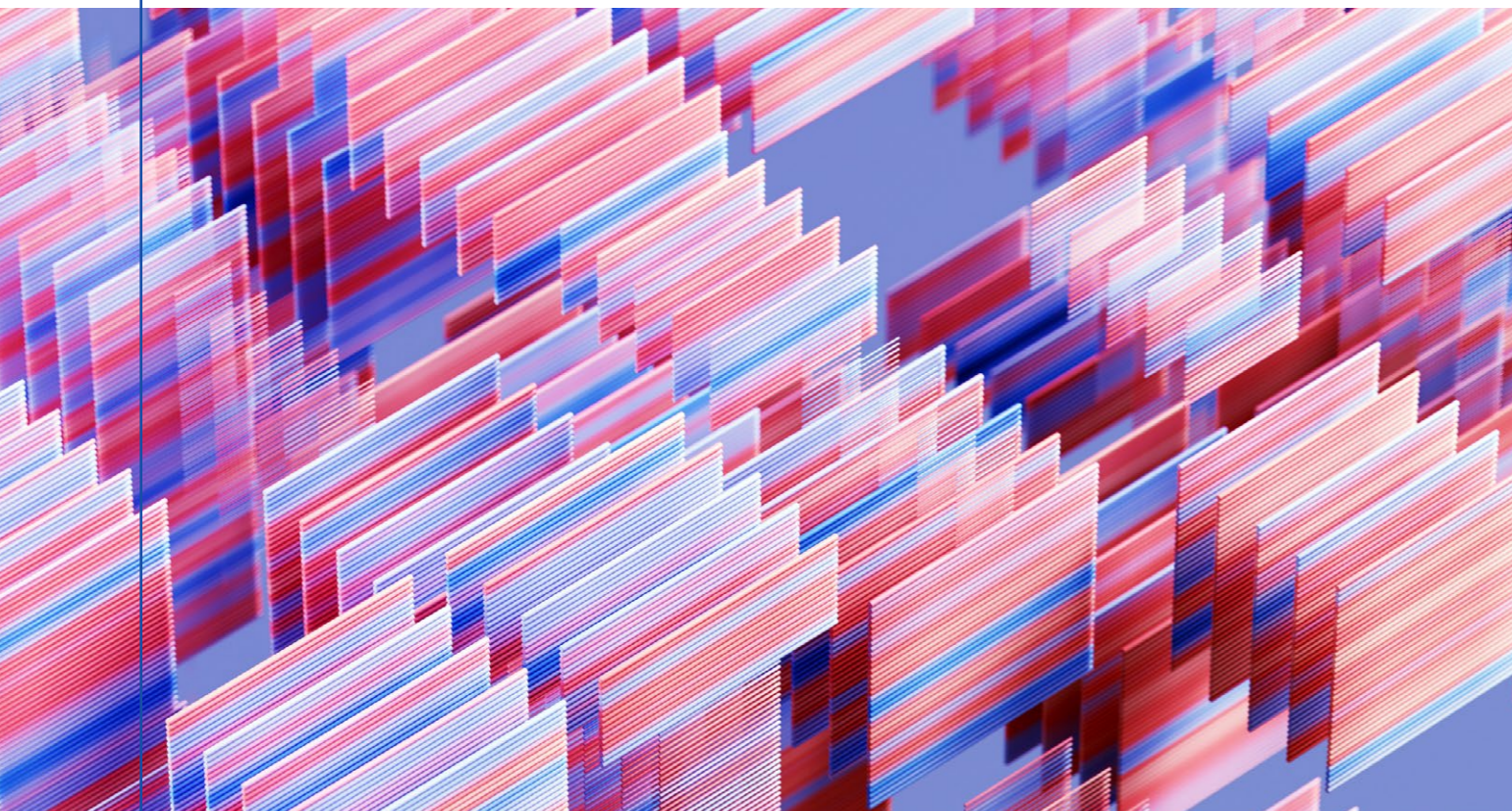
information from our latest research and practice on AI safety and responsibility topics. It details our methods for governing, mapping, measuring, and managing AI risks aligned to the [NIST framework](#), as well as updates on how we're operationalizing responsible AI innovation across Google. We also provide more specific insights and best practices on topics ranging from our rigorous red teaming and evaluation processes to how we mitigate risk using techniques, including better safety tuning and filters, security and privacy controls, provenance technology in our products, and broad AI literacy education.

Our approach to AI responsibility has evolved over the years to address the dynamic nature of our products, the external environment, and the needs of our global users. Since 2018, AI has evolved into a general-purpose technology used daily by billions of people and countless organizations and businesses. The broad establishment of responsibility frameworks has been an important part of this evolution. We've been encouraged by progress on AI governance coming from bodies like the G7 and the International Organization for Standardization, and also

frameworks emerging from other companies and academic institutions. Our updated [AI Principles](#) — centered on bold innovation, responsible development, and collaborative partnership — reflect what we're learning as AI continues to advance rapidly.

As AI technology and discussions about its development and uses continue to evolve, we will continue to learn from our research and users, and innovate new approaches to responsible development and deployment. As we do, we remain committed to sharing what we learn with the broader ecosystem through the publication of reports like this, and also through continuous engagement, discussion, and collaboration with the wider community to help maximize the benefits of AI for everyone.

**Laurie Richardson**  
Vice President, Trust & Safety, Google



# Summary of our responsible AI approach

We have developed an approach to AI governance that focuses on responsibility throughout the AI development lifecycle. This approach is guided by our [AI Principles](#), which emphasize bold innovation, responsible development, and collaborative progress. Our ongoing work in this area reflects key concepts in industry guidelines like the [NIST AI Risk Management Framework](#).

## Govern

Our [AI Principles](#) guide our decision-making and inform the development of our different [frameworks and policies](#), including the Secure AI Framework for security and privacy, and the Frontier Safety Framework for evolving model capabilities and mitigations. Additional policies address design, safety, and prohibited uses.

Our [pre- and post-launch processes](#) ensure alignment with these Principles and policies through clear requirements, mitigation support, and leadership reviews. These cover [model and application](#) requirements, with a focus on safety, privacy, and security. [Post-launch monitoring](#) and assessments enable continuous improvement and risk management.

We regularly publish external model cards and technical reports to provide [transparency](#) into model creation, function, and intended use. And we invest in [tooling for model and data lineage](#) to promote transparency and accountability.

## Map

We take a scientific approach to mapping AI risks through research and expert consultation, codifying these inputs into a risk taxonomy.

A core component is [risk research](#), encompassing emerging AI model capabilities, emerging risks from AI, and potential AI misuse. This research, which we have published in over 300 papers, directly informs our [AI risk taxonomy](#), launch evaluations, and mitigation techniques.

Our approach also draws on [external domain expertise](#), offering new insights to help us better understand emerging risks and complementing in-house work.

## Measure

We have developed a rigorous approach to measuring AI model and application performance, focusing on [safety, privacy, and security benchmarks](#). Our approach is continually evolving, incorporating new measurement techniques as they become available.

[Multi-layered red teaming](#) plays a critical role in our approach, with both internal and external teams proactively testing AI systems for weaknesses and identifying emerging risks. Security-focused red teaming simulates real-world attacks, while content-focused red teaming identifies potential vulnerabilities and issues. [External partnerships](#) and [AI-assisted red teaming](#) further enhance this process.

[Model and application evaluations](#) are central to this measurement approach. These evaluations assess alignment with established frameworks and policies, both before and after launch.

[AI-assisted evaluations](#) help us scale our risk measurement. [AI autoraters](#) streamline evaluation and labeling processes. [Synthetic testing](#) data expedites scaled measurement. And [automatic testing for security vulnerabilities](#) helps us assess code risks in real time.

## Manage

We deploy and evolve mitigations to [manage content safety, privacy, and security](#), such as safety filters and jailbreak protections.

We often [phase our launches](#) with audience-specific testing, and conduct [post-launch monitoring of user feedback](#) for [rapid remediation](#).


We work to [advance user understanding](#) of AI through innovative developments in provenance technology, our research-backed explainability guidelines, and AI literacy education.

To support the broader [ecosystem](#), we provide research funding, as well as tools designed for developers and users. We also promote [industry collaboration](#) on the development of standards and best practices.

# Summary of our responsible AI outcomes to date

Building AI responsibly requires collaboration across many groups, including researchers, industry experts, governments, and users.


We are active contributors to this ecosystem, working to maximize AI's potential while safeguarding safety, privacy, and security.



**300+**  
research papers on AI responsibility and safety topics

Partnered on AI responsibility with outside groups and institutions like the **Frontier Model Forum**, the **Partnership on AI**, the **World Economic Forum**, **MLCommons**, **Thorn**, the **Coalition for Content Provenance and Authenticity**, the **Digital Trust & Safety Partnership**, the **Coalition for Secure AI**, and the **Ad Council**

**\$120 million**  
for AI education and training around the world

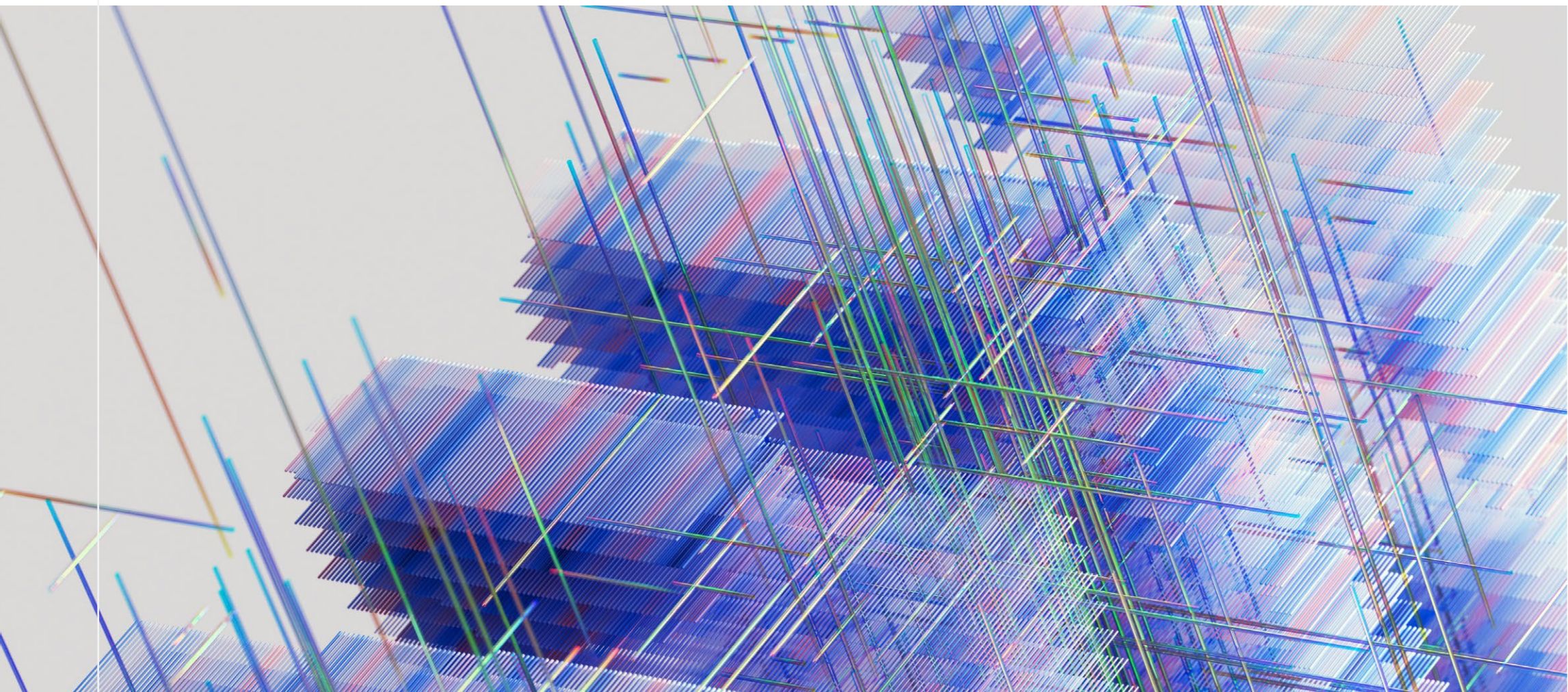


Certified Gemini app, Google Cloud, and Google Workspace through the **ISO/IEC 42001 process**

Achieved "mature" rating for Google Cloud AI in a third-party evaluation of readiness through the **NIST AI Risk Management Framework governance and ISO/IEC 42001 compliance**



**19,000**  
security professionals have taken the SAIF Risk Self Assessment to receive a personalized report of AI risks relevant to their organization



# Govern: Full-stack AI governance

We take a full-stack approach to AI governance — from responsible model development and deployment to post-launch monitoring and remediation.

Our policies and principles guide our decision-making, with clear requirements at the pre- and post-launch stages, leadership reviews, and documentation.

## Policies and principles

Our governance process is grounded in our principles and frameworks:

**AI Principles.** We established and evolve our [AI Principles](#) to guide our approach to developing and deploying AI models and applications. Core to these Principles is pursuing AI efforts where the likely overall benefits substantially outweigh the foreseeable risks.

**Model safety framework.** The [Frontier Safety Framework](#), which we recently updated, helps us to proactively prepare for potential risks posed by more powerful future AI models. The Framework follows the emerging approach of Responsible Capability Scaling proposed by the [U.K.'s AI Safety Institute](#).

**Content safety policies.** Our [policies](#) for mitigating harm in areas such as child safety, suicide, and self-harm have been informed by years of research, user feedback, and expert consultation. These policies guide our models and products to minimize certain types of harmful outputs. Some individual applications, like the Gemini app, also have their own [policy guidelines](#). We also prioritize neutral and inclusive design principles, with a goal of minimizing unfair bias. And we have [Prohibited Use Policies](#) governing how people can engage with our AI models and features.

**Security and privacy framework.** Our [Secure AI Framework](#) focuses on the security and privacy dimensions of AI.

**Application-specific development frameworks.** In addition to Google-wide frameworks and policies, several of our applications have [specific frameworks](#) to guide their day-to-day development and operation.

Our [approach to the Gemini app](#) guides our day-to-day development of the app and its behavior. We believe the Gemini app should:

### 1. Follow your directions

Gemini's top priority is to serve you well.

### 2. Adapt to your needs

Gemini strives to be the most helpful AI assistant.

### 3. Safeguard your experience

Gemini aims to align with a set of [policy guidelines](#) and is governed by Google's [Prohibited Use Policy](#).

### Pre- and post-launch reviews

We operationalize our principles, frameworks, and policies through a system of launch requirements, leadership reviews, and post-launch requirements designed to support continuous improvement.

**Model requirements.** Governance requirements for models focus on filtering training data for quality, model performance, and adherence to policies, as well as documenting training techniques in technical reports and model cards. These processes also include safety, privacy, and security criteria.

**Application requirements.** Launch requirements for applications address risks and include testing and design guidance. For example, an application that generates audiovisual content is required to incorporate a robust provenance solution like [SynthID](#). These requirements are based on the nature of the product, its intended user base, planned capabilities, and the types of output involved. For example, an application made available to minors may have additional requirements in areas like parental supervision and age-appropriate content.

**Leadership reviews.** Executive reviewers with expertise in responsible AI carefully assess evaluation results, mitigations, and risks before making a launch decision. They also oversee our frameworks, policies, and processes, ensuring that these evolve to account for new modalities and capabilities.

**Post-launch requirements.** Our governance continues post-launch with assessments for any issues that might arise across products. Post-launch governance identifies unmitigated residual and emerging risks, and opportunities to improve our models, applications, and our governance processes.

**Launch infrastructure.** We are evolving our infrastructure to streamline AI launch management, responsibility testing, and mitigation progress monitoring.

### Documentation

We foster transparency and accountability throughout our AI governance processes.

**Model documentation.** External model cards and technical reports are published regularly as transparency artifacts. Technical reports provide details about how our most advanced AI models are created and how they function. This includes offering clarity on the intended use cases, any potential limitations of the models, and how our models are developed in collaboration with safety, privacy, security, and responsibility teams. In addition, we publish model cards for our most capable models and open models. These cards offer summaries of technical reports in a “nutrition label” format to surface vital information needed for downstream developers or to help policy leaders assess the safety of a model.

**Data and model lineage.** We are investing in robust infrastructure to support data and model lineage tracking, enabling us to understand the origins and transformations of data and models used in our AI applications.

Our responsible AI approach reflects key concepts in industry guidelines like the NIST AI Risk Management Framework — govern, map, measure, and manage.



# Case study: Promoting AI transparency with model cards

[Model cards](#) were introduced in a [Google research paper](#) in 2019 as a way to document and provide transparency about how we evaluate models.

That paper proposed some basic model card fields that would help provide model end users with the information they need to evaluate how and when to use a model. Many of the fields first proposed remain vital categories of metadata that are found in model cards across the industry today.

Previous iterations of our model cards, such as one to [predict 3D facial surface geometry](#) and one for an [object detection model](#), conveyed important information about those respective models.

However, as generative AI models have advanced, we have adapted our most recent model cards, such as the card for our highest quality [text-to-image model Imagen 3](#), to reflect the rapidly evolving landscape of AI development and deployment. While these model cards still contain some of the same categories of metadata we originally proposed in 2019, they also prioritize clarity, practical usability, and include an assessment of a model’s intended usage, limitations, risks and mitigations, and ethical and safety considerations.

As models continue to evolve, we will work to recognize the key commonalities between models in these model cards. By identifying these commonalities, while also remaining flexible in our approach, we can use model cards to support a shared understanding and increased transparency around how models work.

## Model Card

### Model Details

Basic information about the model.

- Person or organization developing model
- Model date
- Model version
- Model type
- Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
- Paper or other resource for more information
- Citation details
- License
- Where to send questions or comments about the model

### Intended Use

Use cases that were envisioned during development.

- Primary intended uses
- Out-of-scope use cases

### Factors

Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed as required.

- Relevant factors
- Evaluation factors

### Metrics

Metrics should be chosen to reflect potential real-world impacts of the model.

- Model performance measures
- Decision thresholds
- Variation approaches

### Evaluation Data

Details on the dataset(s) used for the quantitative analyses in the card.

- Datasets
- Motivation
- Preprocessing

### Training Data

May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.

### Quantitative Analyses

- Unitary results
- Intersectional results

### Ethical Considerations

### Caveats and Recommendations

The model card fields suggested in our 2019 research paper [“Model Cards for Model Reporting.”](#)



# Map: Identifying and understanding risks

We take a scientific approach to mapping AI risks through research and expert consultation, codifying these inputs into a risk taxonomy. Our mapping process is

fundamentally iterative, evolving alongside the technology, and adapting to the range of contexts in which people use AI models or applications.

## Risk research

We've published more than 300 papers on responsible AI topics, and collaborated with research institutions around the world. Recent areas of focus include:

**Research on novel AI capabilities.** We research the potential impact of emerging AI capabilities such as new modalities and agentic AI, to better understand if and how they materialize, as well as identifying potential mitigations and policies.

**Research on emerging risks from AI.** We also invest in research on the potential emerging risks from AI in areas like biosecurity, cybersecurity, self-proliferation, dangerous capabilities, misinformation, and privacy, to evolve our mitigations and policies.

**Research on AI misuse.** Mapping the potential misuse of generative AI has become a core area of research, and contributes to how we assess and evaluate our own models in these risk areas, as well as potential mitigations. This includes recent research into how government-backed threat actors are trying to use AI and whether any of this activity represents novel risks.

## External domain expertise

We augment our own research by working with external domain experts and trusted testers who can help further our mapping and understanding of risks.

**External expert feedback.** We host workshops and demos at our Google Safety Engineering Centers around the world and industry conferences, garnering insights across academia, civil society, and commercial organizations.

**Trusted testers.** Teams can also leverage external trusted testing groups who receive secure access to test models and applications according to their domain expertise.

## Risk taxonomy

We've codified our mapping work into a taxonomy of potential risks associated with AI, building on the NIST AI Risk Management Framework and informed by our experiences developing and deploying a wide range of AI models and applications. These risks span safety, privacy, and security, as well as transparency and accountability risks such as unclear provenance or lack of explainability. This risk map is designed to enable clarity around which risks are most relevant to understand for a given launch, and what might be needed to mitigate those risks.



### A selection of our latest research publications focused on responsible AI

#### June 2024

- [Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data](#)
- [Beyond Thumbs Up/Down: Untangling Challenges of Fine-Grained Feedback for Text-to-Image Generation](#)

#### July 2024

- [On Scalable Oversight with Weak LLMs Judging Strong LLMs](#)
- [Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders](#)
- [ShieldGemma: Generative AI Content Moderation Based on Gemma](#)

#### August 2024

- [Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2](#)
- [Imagen 3](#)

#### September 2024

- [Knowing When to Ask - Bridging Large Language Models and Data](#)
- [Operationalizing Contextual Integrity in Privacy-Conscious Assistants](#)
- [A Toolbox for Surfacing Health Equity Harms and Biases in Large Language Models](#)

#### October 2024

- [New Contexts, Old Heuristics: How Young People in India and the US Trust Online Content in the Age of Generative AI](#)
- [All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI](#)
- [Gaps in the Safety Evaluation of Generative AI](#)
- [Insights on Disagreement Patterns in Multimodal Safety Perception across Diverse Rater Groups](#)
- [STAR: SocioTechnical Approach to Red Teaming Language Models](#)

#### November 2024

- [A New Golden Age of Discovery: Seizing the AI for Science Opportunity](#)

#### December 2024

- [Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice](#)

#### January 2025

- [Adversarial Misuse of Generative AI](#)
- [How we Estimate the Risk from Prompt Injection Attacks on AI Systems](#)

## Case study: Mapping and addressing risks to safely deploy AlphaFold 3

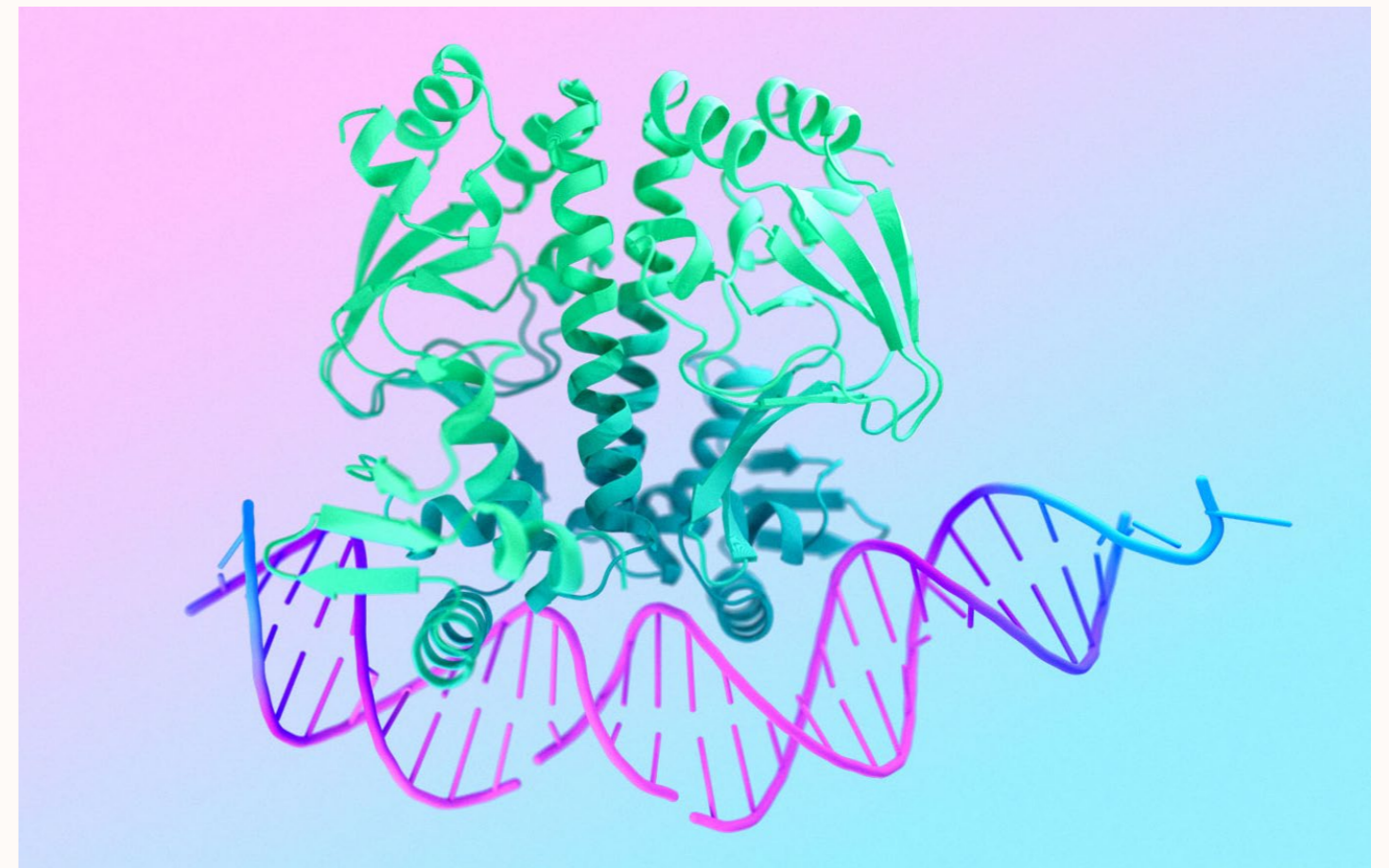
In May 2024, Google DeepMind released [AlphaFold 3](#), an AI model capable of predicting molecular structures and interactions and how they interact, which holds the promise of transforming scientists' understanding of the biological world and accelerating drug discovery. Scientists can access the majority of its capabilities, for free, through our [AlphaFold Server](#), an easy-to-use research tool, or via [open code and weights](#).

We carried out extensive research throughout AlphaFold 3's development to understand how it might help or pose risks to biosecurity. Over the course of AlphaFold's development, we consulted with more than 50 external experts across various fields, including DNA synthesis, virology, and national security, to understand their perspectives on the potential benefits and risks.

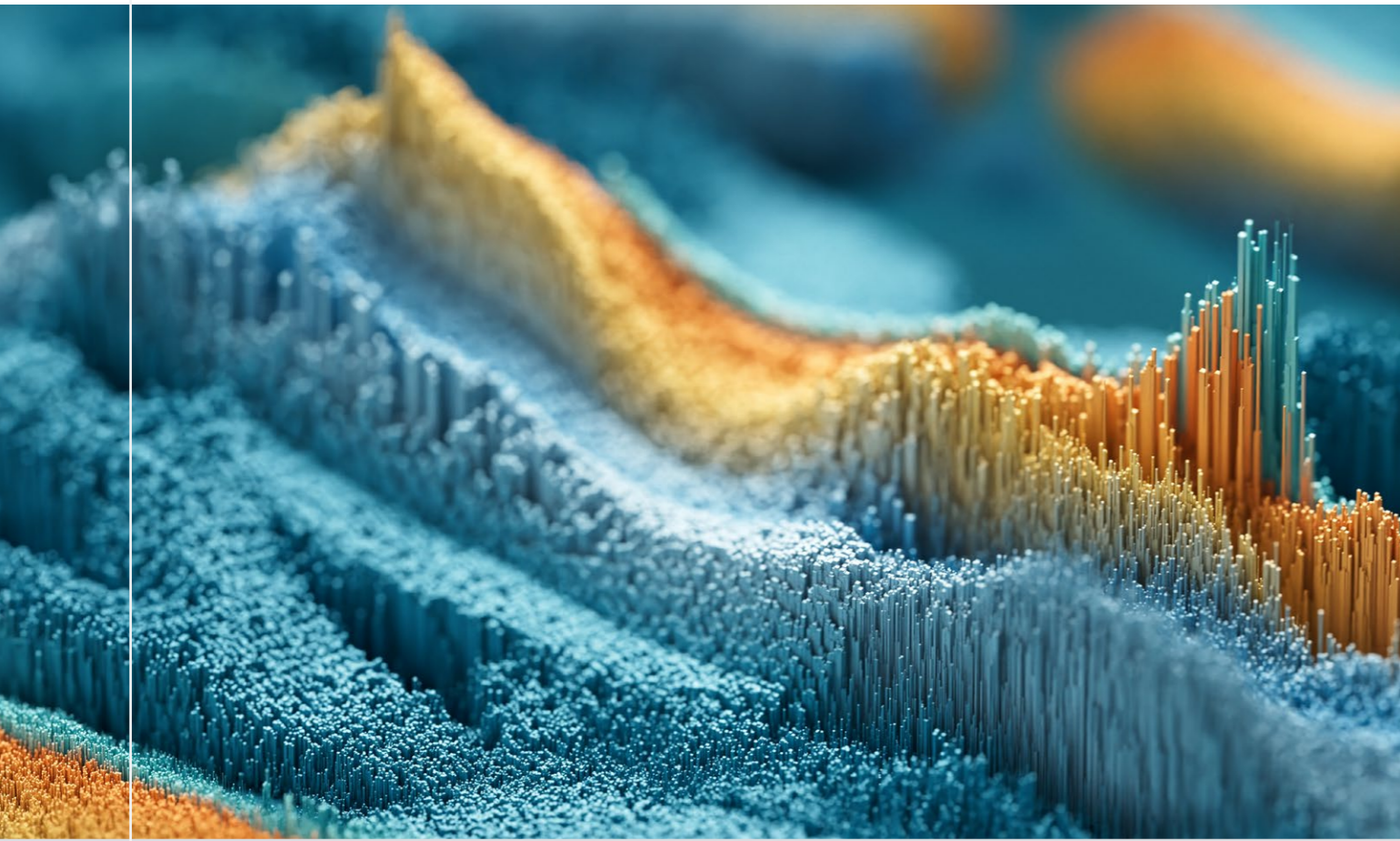
An ethics and safety assessment was conducted with external experts, in which potential risks and benefits of AlphaFold 3 were identified and analyzed, including their potential likelihood and impact. This assessment was grounded in the specific technical capacities of the model and compared the model to other resources like the [Protein Data Bank](#) and other AI biology tools. The assessment was then reviewed by a council of senior internal experts in AI responsibility and safety, who provided further feedback.

As with all Google DeepMind models, AlphaFold 3 was developed, trained, stored, and served within Google's infrastructure, supported by security teams, engineers, and researchers. Quantitative and qualitative techniques are used to monitor the adoption and impact of AlphaFold 3. We partnered with the [European Bioinformatics Institute of the European Molecular Biology Laboratory \(EMBL\)](#) to launch free tutorials on how to best use AlphaFold that more than 10,000 scientists have accessed. We are currently expanding the course and partnering with local capacity builders to accelerate the equitable adoption of AlphaFold 3.

To continue to identify and map emerging risks and benefits from AI to biosecurity, we contribute to civil society and industry efforts such as the U.K. [National Threat Initiative's AI-Bio Forum](#) and the [Frontier Model Forum](#), as well as engaging with government bodies.



**AlphaFold is accelerating breakthroughs in biology with AI, and has revealed millions of intricate 3D protein structures, helping scientists understand how life's molecules interact.**



# Measure: Assessing risks and mitigations

After identifying and understanding risks through mapping, we systematically assess our AI systems through red teaming exercises. We evaluate how well our models and applications perform, and how effectively our risk mitigations work,

based on benchmarks for safety, privacy, and security. Our approach evolves with developments in the underlying technology, new and emerging risks, and as new measurement techniques emerge, such as AI-assisted evaluations.

## Multi-layered red teaming

Red teaming exercises, conducted both internally and externally, proactively assess AI systems for weaknesses and areas for improvement. Teams working on these exercises collaborate to promote information sharing and industry alignment in red teaming standards.

**Security-focused red teaming.** Our [AI Red Team](#) combines security and AI expertise to simulate attackers who might target AI systems. Based on [threat intelligence](#) from teams like the Google Threat Intelligence Group, the AI Red Team explores and identifies how AI features can cause security issues, recommends improvements, and helps ensure that real-world attackers are detected and thwarted before they cause damage.

**Content-focused red teaming.** Our Content Adversarial Red Team (CART) proactively identifies weaknesses in our AI systems, enabling us to mitigate risks before product launch. CART has conducted over 150 red teaming exercises across various products. Our internal AI tools also assist human expert red teamers and increase the number of attacks they're able to test for.

**External red teaming partnerships.** Our external red teaming includes live hacking events such as [DEF CON](#) and [Escal8](#), targeted research grants, [challenges](#), and [vulnerability rewards programs](#) to complement our internal evaluations.

**AI-assisted red teaming.** To enhance our approach, we have developed forms of [AI-assisted red teaming](#) — training AI agents to find potential vulnerabilities in other AI systems, drawing on work from gaming breakthroughs like [AlphaGo](#). For example, we recently shared [details](#) of how we used AI-assisted red teaming to understand how vulnerable our systems may be to indirect prompt injection attacks, and to inform how we mitigate the risk.

## Model and application evaluations

A core component of our measurement approach is running evaluations for models and applications. These evaluations primarily focus on known risks, in contrast to red teaming, which focuses on known and unknown risks.

**Model evaluations.** A subset of the mapped risks is relevant to test at the model level. For example, as we prepared to launch [Gemini 1.5 Pro](#), we evaluated the model for risks such as self-proliferation, offensive cybersecurity, child safety harms, and persuasion. We also develop new evaluations in key areas — such as our work on [FACTS Grounding](#), which is a benchmark for evaluating how accurately LLMs ground their responses in provided source material and avoid hallucinations.

**Application evaluations.** These evaluations are designed to assess the extent to which a given application follows the frameworks and policies that apply to that application. This pre-launch testing generally covers a wide range of risks spanning safety, privacy, and security, and this portfolio of testing results helps inform launch decisions. We also invest in systematic post-launch testing that can take different forms, such as running regression testing for evaluating an application's ongoing alignment with our frameworks and policies, and cross-product evaluations to identify whether known risks for one application may have manifested in other applications.

## AI-assisted evaluations

As AI continues to scale, it's critical that our ability to measure risks scales along with it. That's why we're investing in automated testing solutions, which can run both before launch and on an ongoing basis after release.

**AI autoraters.** At the model layer, Gemini 2.0's reasoning capabilities have enabled major advances in automating evals and developing training data to mitigate identified risks. We have also published [research](#) on the future use of more capable models to help evaluate and rate less capable models. At the application layer, we have been investing in applied AI to triage and label content to streamline and scale evals.

**AI-generated testing data.** We've been investing in "few shot" learning where an AI creates a testing set based on inputs from experts. This significantly accelerates testing when compared with human creation of testing sets.

## Case study: Evaluating Gemma, a family of open models

Our [Gemma](#) models are a family of lightweight, state-of-the-art open models built from the same research and technology used to create the [Gemini](#) family of models.

As part of making Gemma pre-trained models safe and reliable, we used automated techniques to filter out certain personal information and other sensitive data from training sets. Additionally, we used extensive fine-tuning and reinforcement learning from human feedback to align our instruction-tuned models with responsible behaviors. To understand and reduce the risk profile for Gemma models, we conducted robust evaluations including manual red teaming, automated adversarial testing, and assessments of model capabilities for dangerous activities. On top of robust internal evaluations, we also evaluate against well-known academic safety benchmarks. These evaluations are outlined in our [model cards](#) for Gemma models and include:

**Text-to-text content safety.** Human evaluation on prompts covering safety policies, including child sexual abuse and exploitation, harassment, violence and gore, and hate speech.

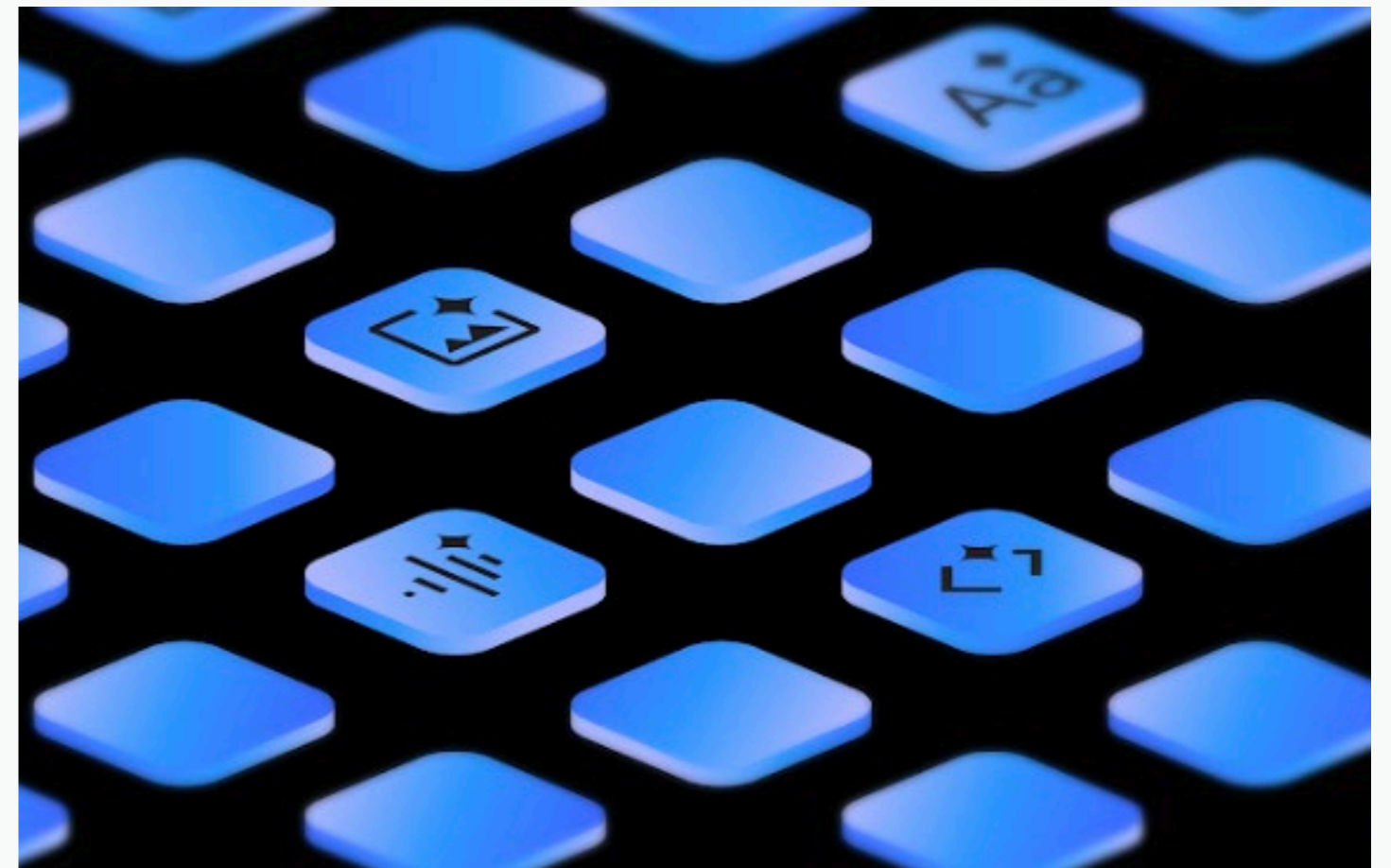
**Text-to-text representational harms.** Benchmarks against relevant academic datasets such as [WinoBias](#) and [BBQ dataset](#).

**Memorization.** Automated evaluation of memorization of training data, including the risk of personally identifiable information exposure.

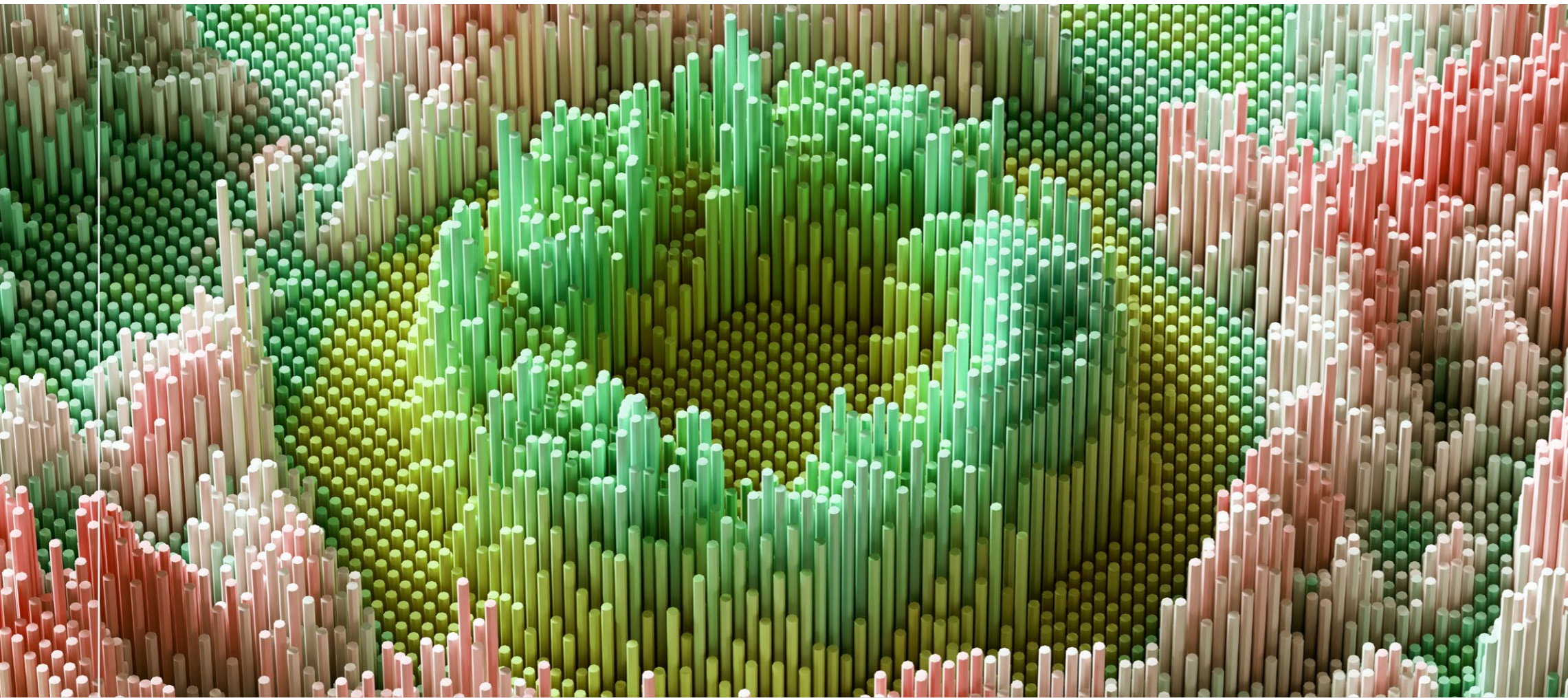
**Large-scale harm.** Tests for “dangerous capabilities,” such as chemical, biological, radiological, and nuclear risks.

The results of these internal and external ethics and safety evaluations are within acceptable thresholds for meeting internal policies for categories such as child safety, content safety, representational harms, memorization, and large-scale harms.

In addition, a Gemma model achieved strong external [results](#) in the [AILuminate v1.0 benchmark](#) from [MLCommons](#). This benchmark assesses the safety of text-to-text interactions with a general purpose AI chat model by a user with malicious or vulnerable intent.



**Gemma is built for responsible AI development from the same research and technology used to create Gemini models.**



# Manage: Mitigating risks

We take a multi-faceted approach to risk mitigation. We implement content safety, security, and privacy mitigations; employ phased launches; empower users with transparency, labeling and training;

harness user feedback; and deploy ongoing monitoring to continuously improve. In addition, we support the wider ecosystem with AI safety tools and standards.

## Managing content safety, security, and privacy

**Managing content safety.** We leverage the expertise our Trust & Safety teams have honed over decades of abuse fighting to establish model and application-level mitigations for a wide range of content safety risks. A critical piece of our safety strategy is a pre-launch risk assessment that identifies which applications have sufficiently great or novel risks that require specialized testing and controls. We also employ guardrails in our models and products to reduce the risk of generating harmful content, for example:

- **Safety filters.** We build safety classifiers to prevent our models from showing users harmful outputs such as suicide content or pornography.
- **System instructions.** We steer our models to produce content that aligns with our [safety guidelines](#) by using system instructions — prompts that tell the model how to behave when it responds to user inputs.

- **Safety tuning.** We fine-tune our models to produce helpful, high-quality answers that align to our [safety guidelines](#).

**Managing security.** We use the SAIF framework to mitigate known and novel AI security risks. The latter category includes risks such as [data poisoning](#), [model exfiltration](#), and [rogue actions](#). We apply security controls, or repeatable mitigations, to these risks. For example, for [prompt injections](#) and [jailbreaks](#), we apply robust filtering and processing of inputs and outputs. Additionally, thorough training, tuning, and evaluation processes help fortify the model against prompt injection attacks. For data poisoning, we implement data sanitization, secure AI systems, enable access controls, and deploy mechanisms to ensure data and model integrity. We have published a full list of our controls for AI security risks. In addition, we continue to research new ways to help mitigate a model's susceptibility to security attacks. For example, we've developed an [AI agent that auto-detects](#) real-world code for security risks.

**Managing privacy.** We have invested deeply in mitigations for privacy risks, as well as researching new risks that might emerge from evolving capabilities like agentic. For examples, our paper on [how AI assistants can better protect privacy by using a "contextual integrity" framework](#) to steer AI assistants to only share information that is appropriate for a given context.

### Phased launches, monitoring, and rapid remediation

**Phased launches.** A gradual approach to deployment is a critical risk mitigation. We have a multi-layered approach — starting with testing internally, then releasing to trusted testers externally, then opening up to a small portion of our user base (for example, Gemini Advanced users first). We also phase our country and language releases, constantly testing to ensure mitigations are working as intended before we expand. And finally, we have careful protocols and additional testing and mitigations required before a product is released to under 18s. To give an example, as [Gemini 2.0's](#) multimodality increases the complexity of potential outputs, we have been careful to release it in a phased way via trusted testers and subsets of countries.

**Monitoring and rapid remediation.** We design our applications to promote user feedback on both quality and safety, through user interfaces that encourage users to provide thumbs up/down and give qualitative feedback where appropriate. Our teams monitor user feedback via these channels closely, as well as feedback delivered through other channels. We have mature incident management and crisis response capabilities to rapidly mitigate and remediate where needed, and feed this back into our risk identification efforts. Importantly, teams are enabled to have rapid-remediation mechanisms in place to block content flagged as illegal.

### Advancing user understanding: provenance, explainability, and AI literacy

**Provenance.** Outputs of our generative AI products typically carry watermarking (via our [SynthID](#) technology) and, when it comes to imagery, relevant metadata (per IPTC standards). As an example, About This Image in Google Image Search started identifying and labeling AI-generated images with SynthID in 2023, alongside other image metadata. We've [open-sourced SynthID](#) to make it easier for any developer to apply watermarking for their own generative AI models, and shared our [analysis](#) of how labeling AI-generated content helps people make informed decisions about the content they see online. [Google Search](#), [Ads](#), and [YouTube](#) are also implementing the latest version of the [Coalition for Content Provenance and Authenticity \(C2PA\)](#)'s authentication standard. And moving forward, we plan to continue investing in the deployment of C2PA across our services.

### We've open-sourced SynthID to make it easier for any developer to apply watermarking

**Explainability.** Explainability is about helping people understand how an AI application operates. Products use disclaimers to set clear expectations — such as reminding people that the AI-generated outputs may contain inaccuracies and that they should take steps to verify information generated by the tool. These disclosure policies are backed up by [research](#), and codified into [explainability guidelines](#) for our teams.

**AI literacy.** To complement the transparency mitigations we implement, it is also critical that governments and industry continue to educate people about how to use AI, and its limitations. We have committed \$120 million for AI education and training around the world. We have also launched AI training for [businesses](#), [developers](#), and [younger learners](#). With Raspberry Pi Foundation, we also co-developed [Experience AI](#), an educational program that offers cutting-edge resources on AI for teachers and students aged 11–14.

### Ecosystem enablement: funding, tools, and standards

**Enabling the ecosystem with research funding.** With our Frontier Model Forum partners, we co-founded the [AI Safety Fund \(AISF\)](#), which provides grants to researchers to help identify, evaluate, and mitigate risks and improve the safe deployment of AI for the benefit of society. Currently, AISF is [prioritizing](#) three critical research areas: biosecurity, cybersecurity, and AI agent evaluation and synthetic content (including AI agent identity verification systems and AI agent safety evaluations).

**Enabling the ecosystem with mitigation tools.** We believe that we also need to complement our mitigation efforts by offering mitigations for the ecosystem.

- We released [ShieldGemma](#) — a series of state-of-the-art safety classifiers that developers can apply to detect and mitigate harmful content in AI model input and outputs. Specifically, ShieldGemma is designed to target hate speech, harassment, sexually explicit content, and dangerous content.
- We offer an existing suite of safety classifiers in our [Responsible Generative AI Toolkit](#), which includes a methodology to build classifiers tailored to a specific policy with limited number of datapoints, as well as existing Google Cloud off-the-shelf classifiers served via API.
- We share AI interpretability tools to help researchers improve AI safety. Our research teams are continually exploring new ways to better understand how models behave. For example, we recently announced [Gemma Scope](#), a new set of tools enabling researchers to “peer inside” the workings of our Gemma 2 model to see how it parses and completes tasks. We believe that this kind of interpretability could open up new opportunities to identify and mitigate safety risks at the model behavior level.
- We launched the [SAIF Risk Self Assessment](#), a questionnaire-based tool that generates a checklist to guide AI practitioners responsible for securing AI systems. The tool will immediately provide a report highlighting specific risks such as data poisoning, prompt injection, and model source tampering, tailored to the submitter's AI systems, as well as suggested mitigations, based on the responses they provided.

### Establishing ecosystem mitigation standards.

- In 2024, we joined [Partnership on AI's](#) working group responsible for understanding progress on the transparency documentation practices of model providers. The working group will help provide valuable insights to policymakers and standards bodies working on codes of practice related to AI.
- We are a founding member of [MLCommons](#), an engineering consortium focused on AI benchmarks, including the new [AILuminate benchmark v1.0](#). This is the first AI safety benchmark produced with open academic, industry, and civil society input and operated by a neutral non-profit with AI benchmarking experience. AILuminate combines a hazard assessment standard, more than 24,000 prompts, online testing with hidden prompts, a proprietary mixture of expert evaluators, and clear grade-based reporting.
- We are also a partner in the World Economic Forum's [AI Governance Alliance](#), a multi-stakeholder initiative to promote transparent development and deployment of AI systems and establish global frameworks and standards for AI governance.
- We introduced the [Coalition for Secure AI](#), which works to advance security measures for addressing the unique risks that come with AI, both for issues that arise in real time and those over the horizon.

# Case study: Managing the safe deployment of NotebookLM

NotebookLM is an AI-powered research and writing assistant designed to help users understand complex information. The development of NotebookLM prioritized responsible AI practices by focusing on identifying potential risks, implementing mitigation strategies, and adopting a phased launch approach.

The phased launch approach consisted of several stages, including:

- **Initial internal testing.** The product was first tested internally by the immediate team, then by a broader group within Google Labs.
- **Trusted tester program.** A group of 50 trusted testers participated in a diary study to record their experiences and provide feedback.
- **Waitlist launch.** After the trusted tester phase, NotebookLM was announced at Google I/O, with a waitlist launch limited to users in the U.S.

- **Global expansion.** Following the U.S. launch and further improvements, NotebookLM was made available in over 200 countries and territories.
- **NotebookLM Business pilot program.** A pilot program for NotebookLM Business was launched for organizations, universities, and businesses.
- **NotebookLM Plus subscription.** A premium version of NotebookLM was launched with enhanced features and higher usage limits.

This phased approach and the incorporation of user feedback have been central to the iterative development of NotebookLM. The team used user feedback to refine the product and implement safety measures before expanding to larger audiences. This strategy allowed for iterative improvements to the tool's safety and effectiveness.

The screenshot displays the NotebookLM interface for a notebook titled "Science in motion". On the left, a sidebar lists 12 sources related to physics, such as "Action and Reaction", "Balanced and Unbalance...", and "Newton's Three Laws...". The main content area features a "Notebook guide" section with options like "Table of contents", "Study guide", and "FAQ". Below this is a "Summary" section providing a comprehensive overview of the sources. To the right, an "Audio Overview" section includes a play button and a progress bar. At the bottom, there is a "Help me understand" section with three prompts: "How do Newton's three laws of motion explain how objects move and interact?", "What are the different types of forces, and how do they affect objects?", and "What is the relationship between speed, velocity, acceleration, and momentum?". The interface also includes a "View chat" button, a "12 sources" indicator, and a "Start typing..." input field.

The development of NotebookLM prioritized responsible AI practices.

# Case study: Offering SynthID to the ecosystem

Google DeepMind developed [SynthID](#), a technology designed to identify AI-generated content by embedding digital watermarks directly into AI-generated images, audio, text, or video. We prioritized the development of SynthID as a tool to manage the risk of misuse of generative AI, particularly the risk of contributing to misinformation and misattribution.

SynthID uses a variety of deep learning models and algorithms for watermarking and identifying AI-generated content. Watermark detection via SynthID can output three possible states: watermarked, not watermarked, or uncertain. This detector can be customized by setting threshold values to achieve a specific false positive and false negative rate for each. The open-sourcing of our SynthID text watermarking tool — developed in-house and used by the Gemini app and web experience — contributes to the responsible use of AI. It makes it easier for any developer to apply watermarking for their generative AI models, so they can detect what text outputs have come from their own LLMs. The open source code is available on [Hugging Face](#), and we've added it to our [Responsible Generative AI Toolkit](#) for developers.



**SynthID** helps identify AI-generated content by embedding an imperceptible watermark on text, images, audio, and video content generated by our models.



# Conclusion

We believe that being bold in AI means being responsible from the start. Our approach to responsible AI is comprehensive, proactive, and aligned with industry standards, including the [NIST AI Risk Management Framework](#).

We will continue to govern AI development with robust internal governance, risk assessments, and continuous updates to our processes, embedding our updated AI Principles in our responsibility work. As AI continues to evolve, we are committed to remaining at the forefront of responsible AI practices. That means continuing to invest in research, collaborate with external experts and institutions, and engage with the wider community to inform how AI is developed and used in a way that benefits society and upholds our core values.

AI is a dynamic field, and responsible AI work has no finish line. We believe that through bold innovation and responsible development, coupled with an ecosystem that helps others to innovate, we can create a future where AI is a force for good, enabling scientific progress and widespread benefits.

