

End-to-end responsibility

A lifecycle approach to AI

Foreword: The AI Responsibility Lifecycle

Introduction

01 Research

02 Design

03 Govern

04 Share

Foreword:

The AI Responsibility Lifecycle

Since its earliest days as a startup, Google has deployed innovative artificial intelligence to make our products more useful, secure, and safe for users around the world. As people begin to use AI in their daily lives, we believe our approach to AI as a company must be both bold and responsible. That's why we are building AI in ways that seek to maximize benefits and minimize risks.

As teams across Google work to make AI even more helpful for our users, we are taking a responsible approach from start to finish. Building AI responsibly is a collective effort that benefits from transparency. For five years we've published an annual report on our progress implementing our AI Principles.

Building on our previous efforts, this paper describes our AI Responsibility Lifecycle: a four-phase process (Research, Design, Govern, Share) that guides responsible AI development at Google. The initial Research and Design phases foster innovation, while the Govern and Share phases focus on risk assessment, testing, monitoring, and transparency. We share insights into how we developed this process, with recent examples and practical tips for implementation.

Our approach to this work will evolve along with the technology. We look forward to continuing to collaborate with experts across the AI ecosystem to learn and help others build upon our advances.

Regards,

Laurie Richardson
Vice President, Trust & Safety

Introduction

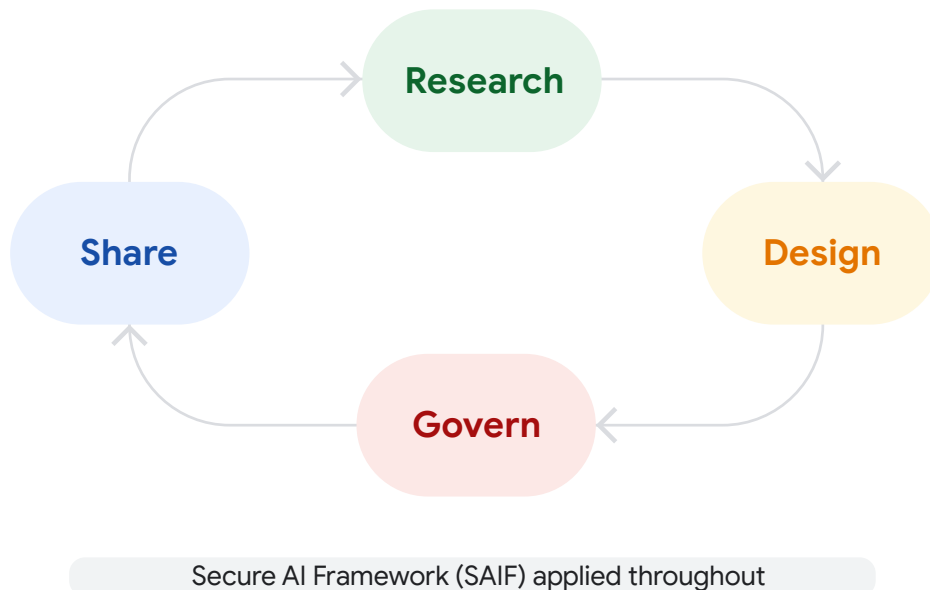
Building on a history of safety

AI is an emerging technology, and we're constantly learning as it evolves. As we work together with others across the industry, we apply lessons from our longstanding approach to improving safety, security, and privacy for our users.

The end-to-end framework that teams at Google use to identify and address potential harms is multidimensional and complex. It requires continuous refinement as we learn from both industry-wide challenges and product-specific risks.

Our holistic approach uses expertise from across Google and is informed by feedback from both internal and external testers, and subject-matter experts. We use what we learn about new applications, extensions and risks to inform how we innovate. The aim is to build safer products that maximize the positive benefits of AI to society. Our goal is to share our thoughts on emerging best practices for generative AI responsibility with others across the AI ecosystem. In this paper, we discuss examples of learnings we've had in the first half of 2024, and look ahead to what's next.

AI Responsibility Lifecycle



01 Research

Context

A history of open research

Since 2012, our researchers have published more than 300 papers on the topic of Responsible AI (available on [Google’s Research publications repository](#) and [Google DeepMind’s publications archive](#)). This includes topics such as user perceptions of AI, data protection, and adversarial testing, and informs responsible model and product innovation.

Research approaches

Our research draws on in-house expertise, including computer scientists, social scientists, and user-experience researchers. We’ve developed novel research approaches such as [Society-Centered AI](#) — as an extension of [Human-Centered AI](#) — to explore the impact of AI on society in areas such as accessibility and health. And, we guide our research efforts with context and insights from bodies like licensed medical associations, together with frameworks such as the [United Nations Universal Declaration of Human Rights](#).

More than a decade of research on responsible AI topics



Applied research: Models

Teams across Google are working to apply and implement these research learnings into our models, products, and risk governance. For example, the 2019 paper “[Model Cards for Model Reporting](#)” defined “nutrition labels” for AI models which were then adopted at Google and across the industry. The 2023 paper on “[Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction](#)” proposed a taxonomy of harms for AI that have informed our pre-launch risk assessments ([see page 18](#)). And, a 2023 Google DeepMind paper examining novel risks that may arise in the future from more powerful general-purpose models (e.g. cyber security threats) and proposing methods for [an early warning system](#) for evaluating models against these risks is being used to inform decisions about responsible model training, deployment, and security.

Applied research: Products

Engineering teams across Google are also using technical research to implement responsible design choices directly in our products. In February, we trained neural network language models in more than 7 languages to enhance privacy for users of the Gboard mobile keyboard. This used techniques outlined in a 2023 paper “[Federated Learning of Gboard Language Models with Differential Privacy](#)” and was one of the largest deployments of user-level differential privacy at Google.

“As generative AI becomes more accessible to more people around the world, [user-experience](#) research on the topic of AI responsibility will be more important than ever. This is why we continue to research novel safety concerns and responsible solutions.”

Recent progress

Applied research: Emerging risks

At the same time as developing advanced models, our researchers are also exploring advanced AI safety evaluations to identify new risks. A 2024 Google DeepMind paper on “[Holistic Safety and Responsibility Evaluations of Advanced AI Models](#)” examined how we have deployed different safety evaluation techniques for our leading models. The findings include the importance of distinguishing between ‘development’ and ‘assurance’ evaluations which apply at different points of the model development lifecycle.

Aligning AI innovation to human values and ethics

Since the beginning of 2024, we have published new papers on “[Controlled Decoding from Language Models](#)” and “[Interactively Critiquing Large Language Models by Converting Feedback into Principles](#)” to align model outcomes to desired behaviors. We published papers on “[Helping or Herding? Reward Model Ensembles Mitigate but do not Eliminate Reward Hacking](#)” for helpfulness and harmlessness, and “[Gradient-Based Language Model Red Teaming](#)” to automatically find prompts that trigger a language model to output unsafe responses.

We also published “[Patchscopes: A Unifying Framework for Inspecting Hidden Representations of Language Models](#),” which describes a new technique to provide natural language explanations of a model’s internal hidden representations — in other words, how the model is generating “meaning” from inputs. This can be used to investigate hallucinations, aid the exploration of multimodal (image and text) representations, and investigate how models build predictions in more complex scenarios.

Additionally, we’re exploring approaches that can be used to mitigate against new harms. In April, we published a study which looked at how to mitigate the risks of generative AI taking advantage of people’s cognitive biases or misrepresentations of information: “[A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI](#).” Also in April, Google DeepMind released “[The Ethics of Advanced AI Assistants](#),” mapping the moral and technical implications of AI assistants as a potentially transformative technology.

Understanding people’s expectations of AI

We also engage in ongoing user research with people who may be directly or indirectly impacted by AI. In January, we released the results of our [global AI survey](#) (with Ipsos) of more than 17,000 people across 17 countries. Looking ahead 25 years, the majority of respondents around the world said they believe AI will be a force for good in areas like healthcare, education and quality of life. However, 1 in 5 workers surveyed have concerns on how they might adapt to a new, AI-enabled economy as the technology advances.

Partnering on research and framework development

We participate in working groups within global organizations including [MLCommons](#), the [World Economic Forum’s AI Governance Alliance](#), the [Coalition for Content Provenance and Authenticity \(C2PA\)](#), [Thorn](#), [Partnership on AI](#), and the [UK AI Safety Institute](#). We have jointly launched responsibility frameworks for [safe deployment](#), [synthetic media](#), and [data enrichment sourcing](#), and a [set of papers](#) on AI governance. We have also signed voluntary commitments including the [Tech Accord to Combat Deceptive Use of AI in 2024 Elections](#) and the [Safety by Design Generative AI principles for child safety](#) developed by Thorn and All Tech is Human.

What’s next?

As generative AI becomes more accessible to more people around the world, [user-experience research](#) on the topic of AI responsibility will be more important than ever. This is why we continue to research novel safety concerns and responsible solutions.

And, as we state in our [AI Principles](#), we’re committed to sharing AI knowledge by publishing papers like this one, [practical advice based on our research](#), and [tools](#) to help researchers and developers explore emerging best practices.

02 Design

Context

Building upon a history of content safety and product quality

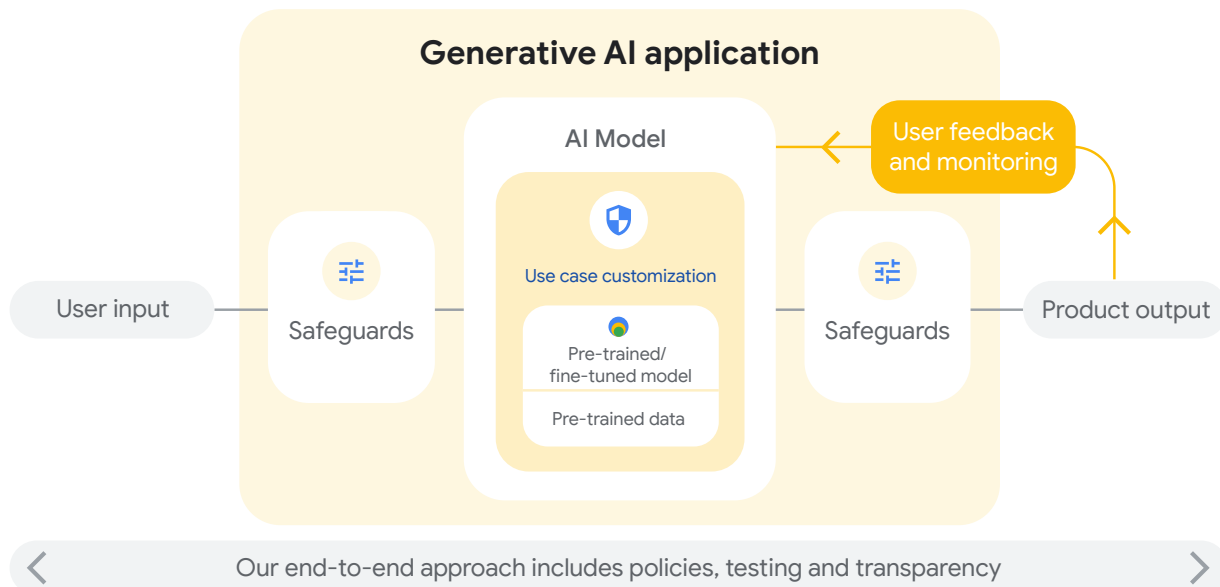
Our approach to responsibility by design is guided by our [AI Principles](#) and builds upon Google's previous experience with keeping users safe on our platforms. We've applied these foundations to evolve our [content safety](#) and [product quality frameworks](#), and to develop a set of additional [generative AI prohibited use policies](#). These policies set out the rules of the road for AI-generated content when people use our services.

As we build generative AI services, our technical approaches to enforce policies at scale include techniques like fine tuning and other layered protections. We also use feedback from people to tune the model, known as [reinforcement learning from human feedback](#) (RLHF). Other layered protections are deployed both when a person inputs a prompt and again when the model provides the output. Policy improvements are informed by ongoing user [feedback](#) and monitoring.

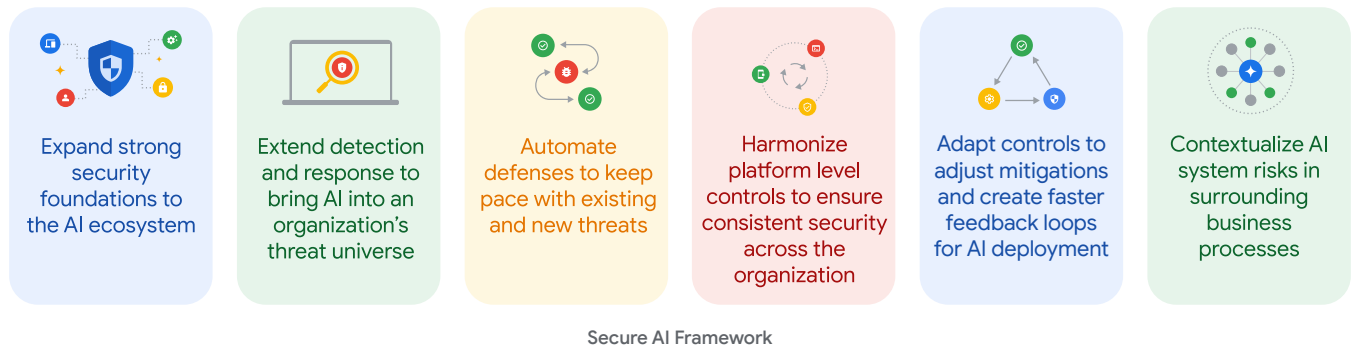
Ensuring the security and integrity of AI systems

Responsibility by design also involves building security into our products from the very beginning. We've codified this approach in our [Secure AI Framework \(SAIF\)](#). For example, hacking techniques like code injection have existed for some time and are used to attack databases. With generative AI, bad actors can use "malicious prompts" to carry out an attack, a process known as "prompt injection." Applying SAIF, we build on our existing security knowledge and adjust mitigations to these new threats.

A responsible approach to building applications



Six core elements of SAIF



Providing context to users

We've been working for many years to provide helpful context to people who want to understand more about the information they find on Google. For instance, in 2021, we announced [About This Result](#), a feature to help people understand and evaluate the context of the results they find; and last year, we introduced a similar initiative for [images](#) to help users understand whether an image is reliable or not.

As part of our commitment to user context, we developed [SynthID](#) to detect and watermark AI-generated content made with our services. The technology enables our models to attach an invisible mark to the content they generate that denotes it was created with Google AI.

“Our approach to responsibility by design is guided by our [AI Principles](#) and builds upon Google’s previous experience with keeping users safe on our platforms. We’ve applied these foundations to evolve our [content safety](#) and [product quality frameworks](#), and to develop a set of additional [generative AI prohibited use policies](#). These policies set out the rules of the road for AI-generated content when people use our services.”

Recent progress

Evolving our technology

In May, we announced that we have expanded SynthID to watermarking and identifying text and video generated by the Gemini app and web experience, in addition to images and audio. These technical solutions are still nascent and case-specific, but represent a step toward offering scalable solutions for researchers and others to identify synthetic AI-generated content.

Updating our policies on synthetic content

To help keep users informed about video content, we introduced a [policy for YouTube creators](#) requiring them to disclose content that is meaningfully altered or synthetically generated when it seems realistic. Last September, we also updated our [political ads policies](#), requiring election advertisers to clearly disclose when their ads contain synthetic images, videos and / or audio that realistically portrays events.

Partnering on provenance standards

Tools like SynthID are a first step. It will take many services working together across the AI ecosystem to help users answer questions about how content was made. That's why Google recently joined the Steering Committee of the [Coalition for Content Provenance Authenticity \(C2PA\)](#), in order to help drive the development and adoption of [Content Credentials](#) — a new type of tamper-resistant metadata that provides an interoperable method to share information about how content was made and edited over time.

What's next?

As AI safety evolves with new models and products, we continue to adapt and tailor our approach for different settings. Whether enabling enterprise customers across industries, or as a partner in boosting everyday creativity or productivity, our priority is to be both helpful and responsible. This means putting safeguards in place from the outset. We view this process as iterative. We will continue learning as the technology develops, and will apply emerging best practices in future updates, features and products.

03 Govern

Context

Assessing impact throughout the product development life cycle

We assess the potential risk and impact of the AI models we're building at both the model level, and at the point of embedding them into a product or service. Reviewers who conduct the risk assessments understand that potential risks and impacts might be different at the model level and at the application level, and consider mitigations accordingly.

At the model level, we conduct impact assessments to identify and document the benefits and harms associated with different potential uses of our models. We draw from various sources in producing impact assessments, including a wide range of academic literature, external expertise and our in-house ethics and safety research. Models are then [instruction tuned](#) and [fine tuned](#) for application into products and services.

Systematically improving models and products with red teaming

[Red teaming](#), also referred to as adversarial testing, is a technique where “ethical hackers” intentionally violate policies for the purpose of discovering and addressing vulnerabilities which could harm users. With the rise of generative AI, it has become a useful tool to help teams systematically improve models and products, and to inform launch decisions. We've established a dedicated [Google AI security red team](#) focused on testing for security and privacy risks. We also host internal, company-wide “Hack-AI-thons” to draw on the experience of hundreds of security and safety experts at Google.

To expand on these efforts to address content safety risks, we've built a new team to use adversarial testing techniques to identify new and unexpected patterns on generative AI products. This team explores innovative uses of AI to augment and expand existing testing efforts.

We also engage in external red-teaming, including at forums like the [DEF CON conference for ethical hacking](#). We're applying and growing external testing methods such as [The Adversarial Nibbler Challenge](#), which engages users to understand potential harms.

Applying the adversarial approach we use during pre-launch responsibility evaluations to post-launch evaluations helps us improve model performance based on user feedback and helps us identify emerging risks. Adversarial testing for safety and fairness reveals how guardrails are working after an application goes to market so we can continuously improve.

Security expertise at scale

Across the development and deployment lifecycle of our AI technologies, we use robust security and safety controls, which we adapt to risks for specific products and users. This is important as it enables early inclusion of prevention and detection controls, augmented by adversarial testing and red-teaming.

In order to detect security attacks that occur against foundational models and applications we monitor input and output, production traffic, and insider access patterns. We also use threat intelligence to stay abreast of novel attacks. For urgent issues, we defend against not only an individual attack but also similar copycat attacks in the future. Our models are developed, trained, and stored within Google’s infrastructure, supported by our global teams of security engineers.

Identifying abuse and protecting against potential harms

Internal controls

- Ongoing red teaming
- Global testing
- Bug reporting

External monitoring

- Bug reporting & bounties
- User feedback (thumbs up/thumbs down)
- User-experience studies
- Social media analyses
- Expert feedback
- LLMs to identify abusive content

Recent progress

Creating clearer accountability

In April, we [consolidated our Responsible AI efforts](#) across the company, moving Responsible AI teams in Research to Google DeepMind, to be closer to where the models are built and scaled. We’ve recently moved other responsibility teams into our central Trust & Safety team, where we are investing in AI testing. These shifts create clearer responsibility and accountability at every level as we build and deploy, and strengthen the feedback loop between models, products, and users.

Refining our risk taxonomy

As part of ongoing updates to our [AI responsibility protocols](#), we continue to develop our risk taxonomy by applying ongoing research on emerging risks, user feedback, internal and external red teaming testing results, and other insights. The evolving taxonomy helps inform product teams as they think about potential AI harms, ranging from content safety to privacy, and from child safety to well being.

Adapting risk assessments

We adapt risk assessments depending on the use case. This is why some of our product areas have developed their own specialized launch review processes. For example, teams at Google Cloud ensure that the [Vertex AI platform](#) and products align to Google's [AI Principles](#). Given the variety of use cases, and Cloud customers' own AI responsibility efforts, these partners play a vital role in product testing.

We also partner with external experts to help us in this process. In April, we announced a partnership with [Coalfire](#), a leading cybersecurity firm, to assess the readiness of the Cloud [AI Risk Management program](#) and the Google Cloud Vertex AI platform against two new AI frameworks: [NIST AI Risk Management Framework](#) and [ISO 42001](#) AI Management System standard.

Identifying common mitigations and standardized benchmarks

We apply standardized sets of protections integrated across our development and deployment platforms and tools. Our recent [technical report](#) for the Gemini family of models outlined several mitigations. We have also included standard academic benchmarks for the capabilities of models, such as [GSMK8](#), [MMLU](#),

[HumanEval](#), and [MATH](#), in our model cards (see [Gemini](#), [Gemini 1.5](#), and [Gemma](#)) so that the reporting of model capabilities, limitations, and testing results is consistent and auditable.

We're continuously improving how we measure AI safety as industry benchmark tools emerge, such as the UK AI Safety Institute's [open-source framework for LLM evaluations](#). This is why we are actively working with MLCommons on a proof of concept for an [AI Safety benchmark](#).

Private releases to test and iterate

Over the past year, Google Cloud has introduced many of our models through private releases. This release mechanism allows our product teams to gather valuable feedback to support better products before we make them generally available. Once we incorporate their feedback and prepare for customer use in production, we update our product documentation to account for any changes. In our product documentation, we provide known limitations of the model and we may issue [service-specific terms](#) to further advise customers on proper use of our products. Cloud also continues to invest in tools to support our customers including: Vertex's [Explainable AI](#), [Model Fairness](#), [Model Evaluation](#), [Model Monitoring](#), and [Model Registry](#) to support data and model governance.

Common product mitigations for large language models (LLMs):

- Disclosures in the Privacy Notice stating that people should not rely on a large language model's responses for medical, legal, financial or other professional advice.
- In-product disclosures reminding users that LLM responses should be double-checked for information accuracy.
- Feedback channels and operational support for user feedback to improve the model and address issues.
- Developer guides to support responsible use, so that developers building AI systems have a good understanding of the capabilities and limitations of the general purpose AI model.
- Feedback channels in the user interface to address issues and undesirable outputs (e.g., thumbs down).
- Age appropriate limitations on responses in risky content categories.

This is a non-exhaustive list

What's next?

Optimizing mitigations across the model and application layers

Whether mitigations are best deployed at the model or application layer is important as we think through how to allocate responsibilities between base model providers and deployers of AI. AI deployers may have more control over the application environment, so mitigation efforts will likely be needed at the application level. The nature of the application also significantly influences the overall risk profile — for example a creative tool to help artists or a scientific tool to aid natural disaster predictions.

Scaling protections and testing

We have been using deep learning and other AI techniques to build controls and protections for years. We're now on a journey to leverage generative AI natural language understanding capabilities to extend AI protection to many more use cases. We believe it's also necessary to innovate with methods for scaled automated testing. As such, we've been researching how to use LLM-based auto-raters to enable efficiency and scaling.

Continuous learning to enable efficiency

While monitoring how our AI products are performing at launch is key, it's also necessary to continuously adapt protections. This includes transferring learnings from live traffic and incidents to longer term development of infrastructure and model protections. Continuous learning is critical for improving accuracy, reducing latency, and increasing the efficiency of protections over time.

“Across the development and deployment lifecycle of our AI technologies, we use robust security and safety controls, which we adapt to risks for specific products and users. This is important as it enables early inclusion of prevention and detection controls, augmented by adversarial testing and red teaming.”

04 Share

Context

Working with others

Transparency is an essential part of a lifecycle approach to AI responsibility. The goal of AI transparency is both to inform how a model is built, and also to identify new ways to understand and mitigate emerging risks. As we and others continue to make technical advances, we know it's more important than ever to keep sharing what we learn with partners in research, industry and government.

Sharing with external researchers and civil society

To help researchers understand how a model is trained and tested, we publish [technical reports](#) with details on how we evaluate safety. We share [model cards](#) that summarize essential facts from these reports in a structured way, so that it's easier to find and understand this information. We also put out [research papers](#) for different types of academic audiences, [hands-on guides](#), and [tools for developers](#). We provide funding through programs like the [Digital Futures Project](#), which distributes grants to leading global think tanks and academic institutions to research AI and society.

Sharing best practices with the industry

We came together with other companies and civil society groups to foster responsible practices in the development, creation, and sharing of AI. Together with other AI companies like OpenAI and Anthropic, we launched the [Frontier Model Forum](#) to help advance AI safety research and support efforts to develop AI applications to meet society's most-pressing needs. For specialist industry areas like security, we've hosted [workshops](#) with practitioners, and published AI security [best practices](#).

Sharing with governments

Working with governments around the world is important to further AI's potential to help address society's greatest challenges, and shape it responsibly together. We're actively working with governments by contributing our cutting-edge tools and compute and data resources to projects like the National Science Foundation's [National AI Research Resource pilot](#), which aims to democratize AI research across the U.S. We have given bodies like the [UK AI Safety Institute](#) access to some of our most capable models for research and safety purposes to build expertise and capability for the long term. And, we have committed to advancing responsible practices in the development and use of AI through forums like the [Voluntary White House AI Commitments](#) and the [G7 Code of Conduct](#).

Transparency

- Structured model and data information (cards)
- Technical reports and research papers
- Performance metrics and benchmarks
- Standards
- Explainability and user context (user notices in product, policy pages, FAQs, etc.)
- Tools and guides for developers

This is not an exhaustive list

Recent progress

Resources for researchers and civil society

In May, we updated [our model card hub](#) making it easier to find information on some of our recent model cards, including [API models](#), [open models](#) and [large language foundation models](#). These model cards help researchers, developers and civil society groups understand a model's strengths and limitations.

More broadly, we are committed to providing funding to researchers and civil society groups active in AI safety research. In April, we launched a series of AI Opportunity Funds in the [US](#), [Europe](#) and [developing nations](#) to train millions of people worldwide on critical AI skills. In May, we also announced new funding to [support AI advancement in Central and Eastern Europe](#).

Resources for the industry

In February, [we launched Gemma](#), a new generation of open models to assist developers and researchers in building AI responsibly. At the same time, teams at Google released a [Responsible Generative AI Toolkit](#) to provide guidance and tools to create safer AI applications with these new open models. The toolkit includes guidance on setting safety policies and methodologies for building robust safety classifiers.

Resources for governments

Like any emerging technology, AI presents new opportunities as well as challenges. For example, generative AI makes it easier than ever to create new content, but it can also raise questions about trustworthiness of information, like we see with “deepfakes.” As more than 1 billion people around the world head to the voting booth in 2024, our teams are particularly focused on maintaining the integrity of information related to elections on our platforms. Earlier this year, we [published updates](#) on our holistic approach to election integrity, including how we are handling the labeling of synthetic content related to elections.

We also continue to actively engage with governments through different national and international bodies looking at AI safety. In February, the Frontier Model Forum became a founding member of the [U.S. Artificial Intelligence Safety Institute Consortium \(AISIC\)](#), a new consortium which brings together more than 200 organizations to develop science-based and empirically backed guidelines and standards for AI measurement and policy. We have also continued our close collaboration with bodies like the UK's new AI Safety Institute, the European Commission and others.

What's next?

Refining responsibility processes as AI advances

Our teams are working on a number of initiatives which will help us provide even more transparency on our work building bold and responsible AI. Moving forward, we will continue integrating advances like watermarking and other emerging techniques to secure our latest generations of Gemini, Imagen, Lyria, and Veo models. We are also continuing to advance work with the [Frontier Model Forum](#),

including improving how we share information across the industry and developing charters on safety standards and evaluations for the most advanced AI models.

Our teams also continue to work on furthering standards around generative AI detection, expanding external testing, and building developer safety tools.

Conclusion

As we continue to expand AI use cases and make technical advances, collaboration across industry, governments, researchers and civil society is crucial. This involves sharing knowledge, identifying ways to mitigate emerging risks, preventing abuse, and furthering the development of tools to increase transparency. As we continue learning more through our lifecycle approach to AI responsibility, we commit to sharing knowledge with the whole ecosystem. You can find updates [here](#).

Google

AI Principles

Our approach to building beneficial AI

Our approach to developing and harnessing the potential of AI is grounded in our founding mission — to organize the world’s information and make it universally accessible and useful. We believe [our approach](#) to AI must be both bold and responsible. Bold in rapidly innovating and deploying AI in groundbreaking products used by and benefiting people everywhere, contributing to scientific advances that deepen our understanding of the world, and helping humanity address its most pressing challenges and opportunities. And responsible in developing and deploying AI that addresses both user needs and broader responsibilities, while safeguarding user safety, security, and privacy.

We approach this work together, by collaborating with a broad range of partners to make breakthroughs and maximize the broad benefits of AI, while empowering others to build their own solutions.

Our approach to AI is grounded in these three principles:

1 **Bold** innovation

We develop AI that assists, empowers, and inspires people in almost every field of human endeavor; drives economic progress; and improves lives, enables scientific breakthroughs, and helps address humanity’s biggest challenges. This means:

- Developing and deploying models and applications where the likely overall benefits substantially outweigh the foreseeable risks.
- Advancing the frontier of AI research and innovation through rigorous application of the scientific method, rapid iteration, and open inquiry.
- Using AI to accelerate scientific discovery and breakthroughs in areas like biology, medicine, chemistry, physics, and mathematics.
- Focusing on solving real world problems, measuring the tangible outcomes of our work, and making breakthroughs broadly available, enabling humanity to achieve its most ambitious and beneficial goals.

2 Responsible development and deployment

Because we understand that AI, as a still-emerging transformative technology, poses evolving complexities and risks, we pursue AI responsibly throughout the AI development and deployment lifecycle, from design to testing to deployment to iteration, learning as AI advances and uses evolve. This means:

- Implementing appropriate human oversight, due diligence, and feedback mechanisms to align with user goals, social responsibility, and widely accepted principles of international law and human rights.
- Investing in industry-leading approaches to advance safety and security research and benchmarks, pioneering technical solutions to address risks, and sharing our learnings with the ecosystem.
- Employing rigorous design, testing, monitoring, and safeguards to mitigate unintended or harmful outcomes and avoid unfair bias.
- Promoting privacy and security, and respecting intellectual property rights.

3 Collaborative progress, together

We make tools that empower others to harness AI for individual and collective benefit. This means:

- Developing AI as a foundational technology capable of driving creativity, productivity, and innovation across a wide array of fields, and also as a tool that enables others to innovate boldly.
- Collaborating with researchers across industry and academia to make breakthroughs in AI, while engaging with governments and civil society to address challenges that can't be solved by any single stakeholder.
- Fostering and enabling a vibrant ecosystem that empowers others to build innovative tools and solutions and contribute to progress in the field.

Responsible AI

AI is transforming industries and solving important, real-world challenges at scale. This vast opportunity carries with it a deep responsibility to build AI that works for everyone.



Staying ahead of the curve: The business case for responsible AI

[Read the report](#)

Benefits

How values-based AI is good for your business

Safer and more accountable products

Advanced technologies are most successful when everyone can benefit from them. Evaluating your AI systems, both when they perform as intended and when they don't, is crucial to building accountable products.

Earn and keep your customers' trust

Lack of trust in AI systems is a growing barrier to adoption in enterprise with more organizations selecting enterprise products based on AI commitments and practices. A responsible AI approach earns trust.

A culture of responsible innovation

Empowering AI decision-makers and developers to take ethical considerations into account enables them to find new, innovative ways to drive your mission forward.

Learn more about our [perspectives on issues and AI governance](#) at Google and how we work to [build responsible AI for everyone](#).

Key features

Google Cloud's approach to responsible AI

AI Principles

Since 2018, Google's [AI Principles](#) have served as a living constitution, keeping us motivated by a common purpose. Our center of excellence, the Responsible Innovation team, guides how we put these principles to work company-wide, and informs Google Cloud's approach to building advanced technologies, conducting research, and drafting our policies.

Putting principles into practice

Rigorous evaluations are a critical component of building successful AI. To drive alignment with our AI Principles at Google Cloud, two diverse review bodies undertake deep ethical analyses and risk and opportunity assessments for any technology product we build and early-stage deals involving custom work. [Learn more](#).

Tools and education

Responsible AI tools are an increasingly effective way to inspect and understand AI models. We're building resources like [Explainable AI](#), [Model Cards](#), and the [TensorFlow open-source toolkit](#) to provide model transparency in a structured, accessible way. We share what we're learning through [Responsible AI practices](#), [fairness best practices](#), [technical references](#) and [tech ethics](#) materials.