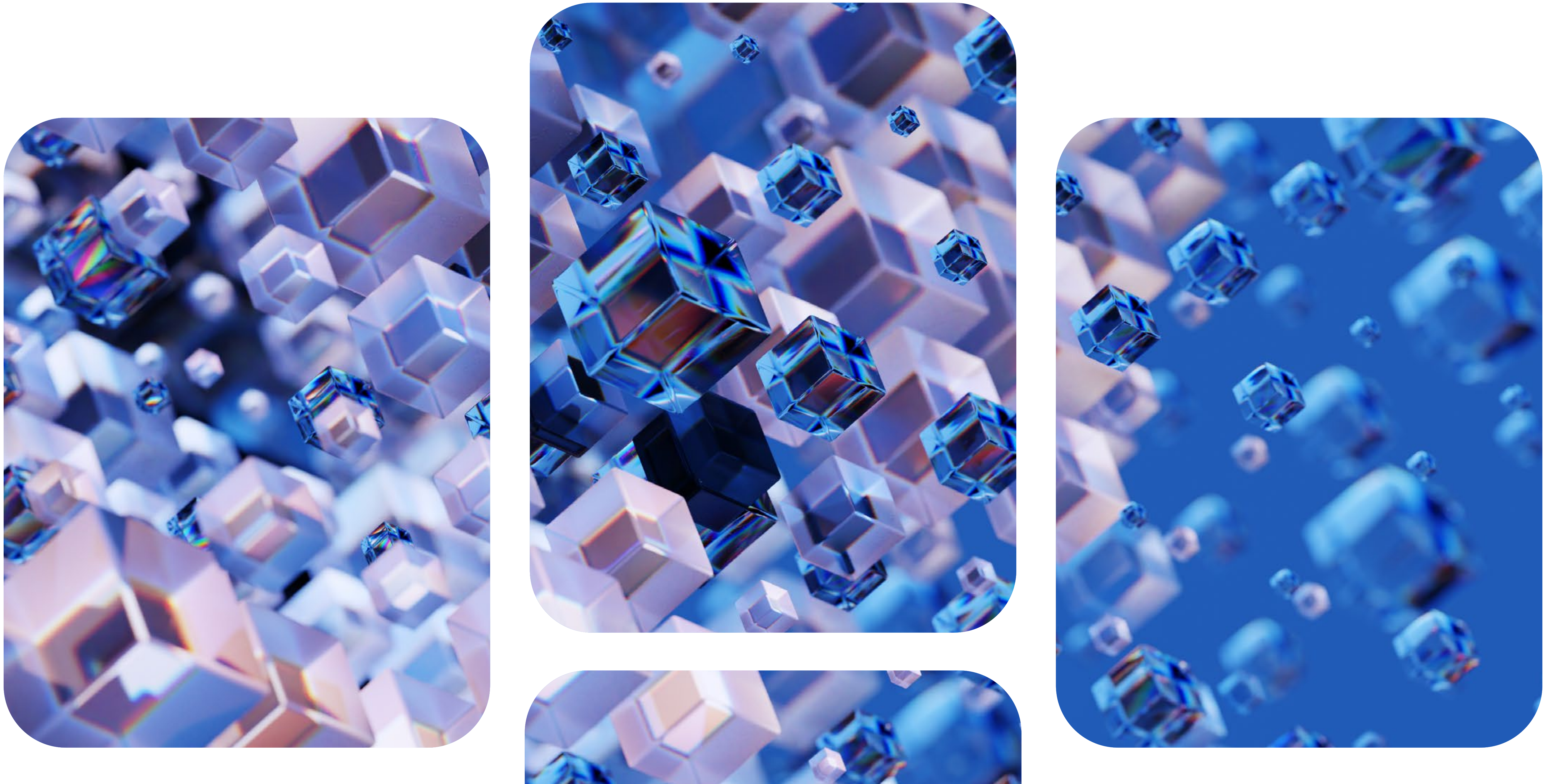
Google

Responsible AI Progress Report

Published in February 2026



Foreword: The opportunity of the AI era

If 2024 was defined by building out the foundations for an AI future, 2025 marked AI's shift into a helpful, proactive partner, capable of reasoning and navigating the world with users. As models grow even more sophisticated, we see users and businesses around the globe transitioning from exploration to integration, finding new ways to put these tools to work in their daily lives. From foundational advances in scientific discovery and clinical milestones in healthcare to the rise of agentic systems and new tools to support creativity such as 'vibe coding' and generative media, the transformational potential of these tools is coming more clearly into focus.

Since we started publishing these reports, our responsible AI development approach has continued to mature and is now fully embedded with our product development and research lifecycles. In 2025, as models became more capable, personalized, and multi-modal, we relied upon robust processes for testing and mitigating risks, and deepened the rigorous safeguards built into our products. To meet this challenge at the speed and scale of Google, we have paired 25 years of user trust insights with a comprehensive testing strategy that is driven by human expertise and supported by AI-enabled automation.

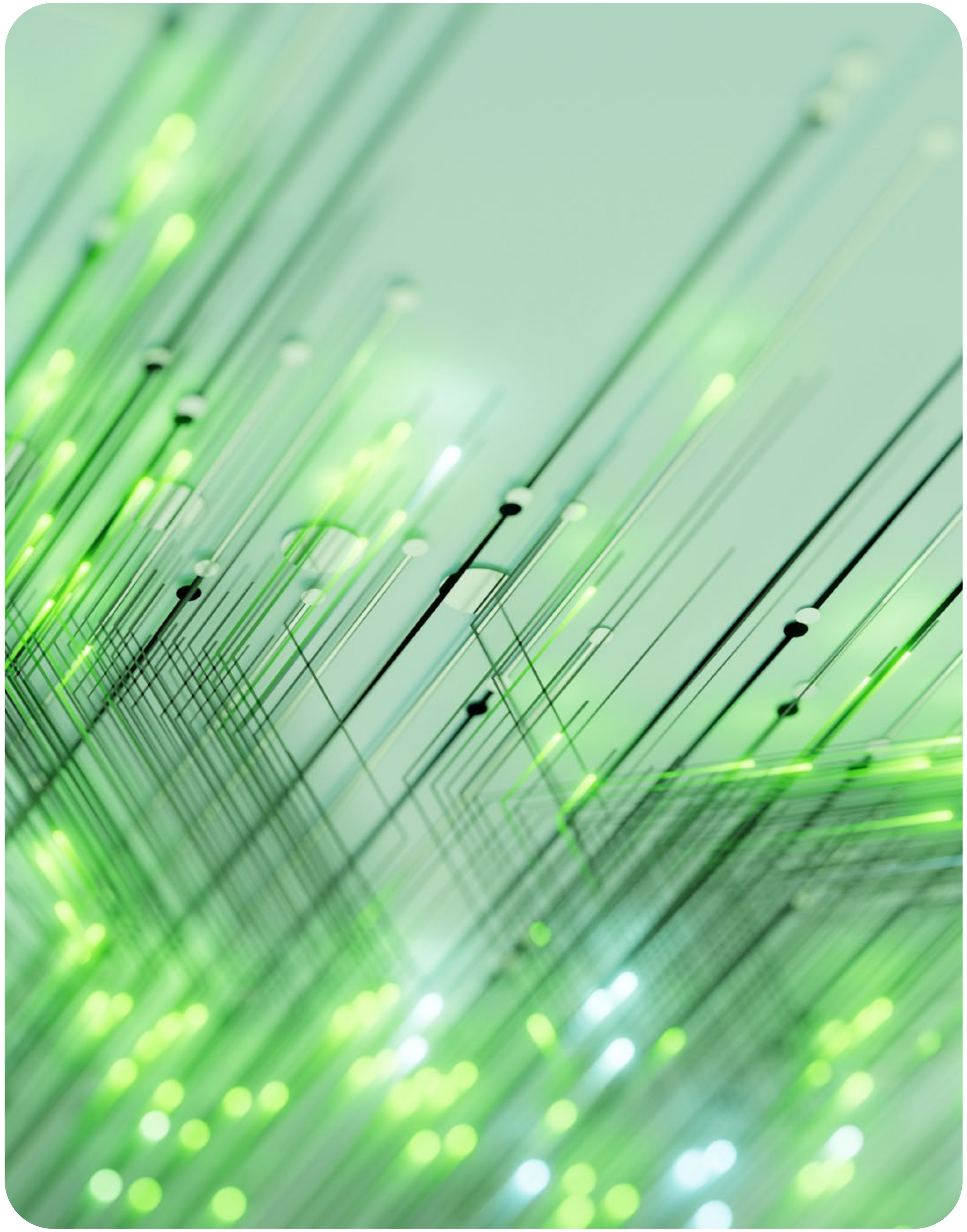
This work continues to be guided by our AI Principles, which we updated last year to reflect our latest understanding of the opportunities and risks presented by this platform shift. Today's report details our multi-layered approach to responsible AI governance, and focuses in particular on agentic and frontier risks from increasingly sophisticated models. In such a dynamic environment, it also shows how our systems are built to be able to detect and then adapt to emerging risks. Whether we are hardening agentic systems against adversarial manipulation or embedding provenance signals into every synthetic output, our goal remains clear: to ensure that we are "bold" and "responsible" in both our development and implementation.

Responsibility is not only about stopping bad outcomes. It is also about enabling broad access to these tools for the maximum benefit of people and society. By striking the right balance we can ensure that AI is used to tackle existential challenges that were previously insurmountable, from forecasting floods for 700 million people to decoding the human genome and helping prevent blindness.

Building trust in these tools requires deep partnership with governments, academics and civil society. We will continue to vigorously collaborate to set standards for this remarkable era. As AI advances, we'll continue to iterate and share research and tools with the broader ecosystem, with a goal to promote uses of AI that will improve lives everywhere.

Laurie Richardson
Vice President, Trust & Safety, Google

Helen King
Vice President, Responsibility, Google DeepMind



How we develop and deploy AI

The foundations for AI-driven innovation are systems that are developed and deployed responsibly from the start. We are bold in our ambition to deliver the economic and societal benefits of the AI era — benefits that can unlock opportunity for communities and accelerate scientific discovery. We achieve our goal of being bold and responsible through a comprehensive approach that spans the entire AI lifecycle — from model development and deployment to post-launch monitoring and remediation.

A multi-layered approach to responsible AI governance

We employ a multi-layered approach to AI governance that combines human expertise, user feedback, and automated systems that help scale our work to manage risk.

Research. We take a research-driven approach to AI risk and governance. This includes identifying current and emerging risks associated with our models and products across new modalities and form factors — such as robotics and agentic AI.

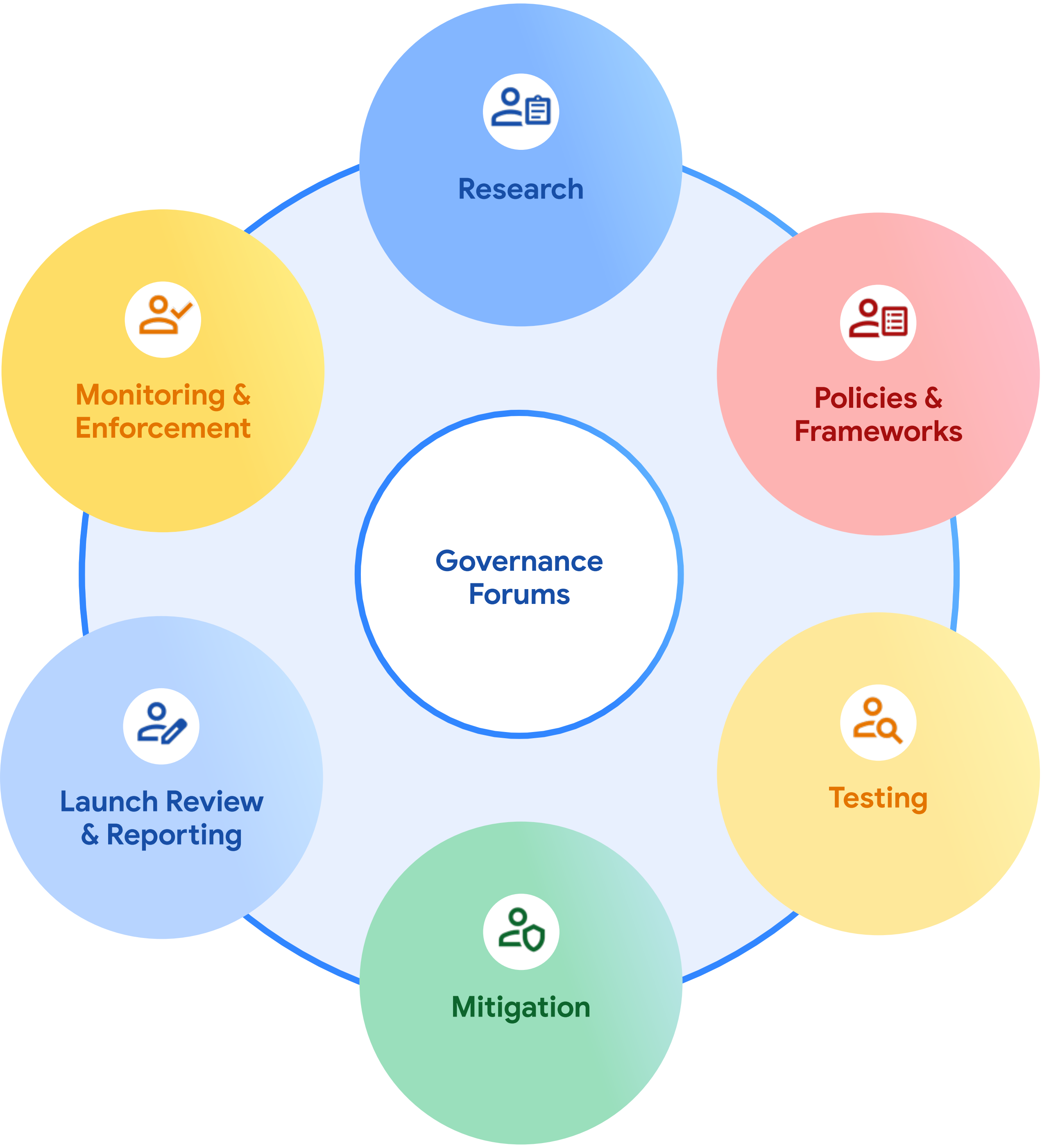
Policies and Frameworks. We develop rigorous AI policies and guidelines — such as our content safety policies and Prohibited Use Policy — that are designed to prevent potentially harmful outputs and misuse of our products. Developed with internal and external experts, these protections guide multi-modal outputs to mitigate risks in key areas including: child safety, dangerous content, sexual content, and medical information. We also develop frameworks for managing more nascent risks posed by frontier AI models, as illustrated in our latest Frontier Safety Framework and Secure AI Framework.

Testing. We take a comprehensive approach to stress test our systems against our policies and frameworks. Our testing includes both scaled evaluations and red teaming of our models and products, including our most advanced AI systems that leverage personal intelligence and agentic AI.

Mitigation. We proactively mitigate risks through both supervised fine-tuning and reinforcement learning to ensure models are aligned with our content safety policies. Additionally, we deploy out-of-model mitigations, such as safety filters and conditional system instructions, to provide additional layers of protection by identifying, filtering out, or steering model output away from harmful or inappropriate content. We also leverage our Search tools to factually ground responses that require fresh or authoritative information. To further minimize risk, we phase global expansion of models and products to allow sufficient time and safety considerations for different languages and regions. We implement added care for sensitive audiences, especially our under-18 users, for whom we enforce heightened protocols and mitigations.

Launch Review and Reporting. Before launching a model or product, we evaluate a wide array of risks to determine whether our safety guardrails appropriately mitigate those risks or if additional protections are needed. Our AI launches undergo expert reviews to confirm they meet rigorous responsibility standards, guided by our AI Principles. We also publish model cards and other reports to provide essential information regarding model creation, function, and intended use.

Monitoring and Enforcement. We use a combination of automated systems and human reviews to engage in continuous post-launch monitoring to improve our AI models and products, and detect activity and behavior that suggests misuse of our consumer products. This includes actively soliciting user feedback, evaluating logs data to identify known and emerging user adoption patterns, and monitoring third-party signals via social media and trusted partners. We collate these insights and extract opportunities to improve our models and products.



Governance Forums. Our multi-layered process includes launch reviews for both frontier models and applications developed using these models. Our model launches are reviewed at Google DeepMind's Launch Review forum, which approves model releases, and our many application launches are reviewed systematically via launch infrastructure and centralized expert risk reviews, as well as via various application-focused launch review forums. These launch-specific forums are complemented by our Artificial General Intelligence (AGI) Futures Council, which consists of members of Google's senior management and Alphabet's Board of Directors. Building on our AI Principles, the Council provides perspectives and recommendations to our Board and management team on long-term opportunities, risks, and impacts associated with the development of AGI. Council topics include promoting widespread benefits, addressing technical safety and security priorities, supporting scientific moonshots, and progressing alignment on national and international standards.

Case study

Responsible AI governance in action

Our multi-layered approach to responsible AI governance is designed to adapt to each unique innovation. Our most recent launches, including Gemini 3, our latest Frontier Safety Framework, and our progress in emerging AI fields such as agentic AI, personal assistance, and artificial general intelligence (AGI) demonstrate this responsibility in action.

Gemini 3: our most secure model yet

We conducted rigorous testing to assess model alignment with our policies and frameworks. We applied these insights to deploy targeted mitigations to further model alignment, while our ongoing monitoring helps inform continuous model improvement.

Gemini 3 represents our most secure model yet, having undergone the most comprehensive set of safety evaluations of any Google AI model to date. Developed in close partnership with internal safety and security teams, Gemini 3 was subjected to rigorous testing via red teaming and safety reviews aligned with our AI Principles and Gemini safety policies. Our evaluations showed that Gemini 3 achieved specific gains in reducing sycophancy, resisting prompt injections, and improving protection against cyber misuse.

Our updated Frontier Safety Framework, which incorporates lessons from previous versions and the latest industry best practices, was central to our approach for deploying Gemini 3. The framework contains a set of protocols designed to identify and mitigate severe risks from frontier AI models, such as cyberattacks, CBRN risks, and harmful manipulation.

The Framework is based around a set of "Critical Capability Levels" — thresholds where a model's capabilities, if unmitigated, could pose severe risks. This includes a new research Critical Capability Level (CCL) on harmful manipulation. This CCL is focused on a model's capability to systematically and substantially manipulate users in direct AI-human interactions and which may be misused to cause harm at a severe scale. This addition builds on and operationalizes research we've done to identify and evaluate mechanisms that drive manipulation from generative AI.

To accompany the launch of Gemini 3, we published a report documenting how we evaluated the model against these thresholds and why we ultimately deemed it safe to deploy. In addition to our own testing, we also partnered with world-leading subject-matter experts, provided early access to bodies such as the UK AI Security Institute, and obtained assessments from independent evaluators such as Apollo Research, Vaultis, Dreadnode, and more. Our responsible approach to Gemini 3 continues through our monitoring and enforcement, informed by our robust AI usage policies, our product-level policies, and feedback from user reporting.

Securing the next generation of browsing

As we begin to introduce agentic capabilities to Chrome — allowing Gemini to assist with complex, multi-step web tasks — we have designed a novel security framework to mitigate risks and protect the user experience.

User alignment
We deployed a specialized, high-trust AI model we call the User Alignment Critic that reviews proposed agent actions. The Alignment Critic acts as an independent reviewer, vetoing actions that do not align with the user's specific intent.

Strict boundaries
We introduced Agent Origin Sets, which restrict the agent's reach to interact only with data related to the task at hand.

Mitigation of social engineering
While the agent is active, it checks every page it sees for indirect prompt injection. In addition to Chrome's safety features and on-device AI that help detect traditional scams, this prompt-injection classifier helps prevent the agent from taking actions that are not aligned with the user's goal.

Mandatory human oversight
Sensitive actions — including payments and purchases, posting on social media, and credential use — require human confirmation before execution, giving users transparency and control over these types of interactions.

Ongoing testing, monitoring, and mitigation
In addition to other safeguards, we built automated red-teaming systems that try to derail the agent in Chrome. We start with a set of diverse attacks crafted by security researchers, and use LLMs to expand on them following a technique we adapted for browser agents, prioritizing testing against broad and high-impact attacks.

Launching personal assistance with controls built in

As part of the development of Personal Intelligence we identified the specific mitigations required to help keep users safe while pushing the boundaries of what AI can achieve.

User control
Users have a choice on whether or not to connect new data sources to the Gemini App or Search AI Mode, and they can also choose to engage in conversations without personalization, and set their activity to auto-delete.

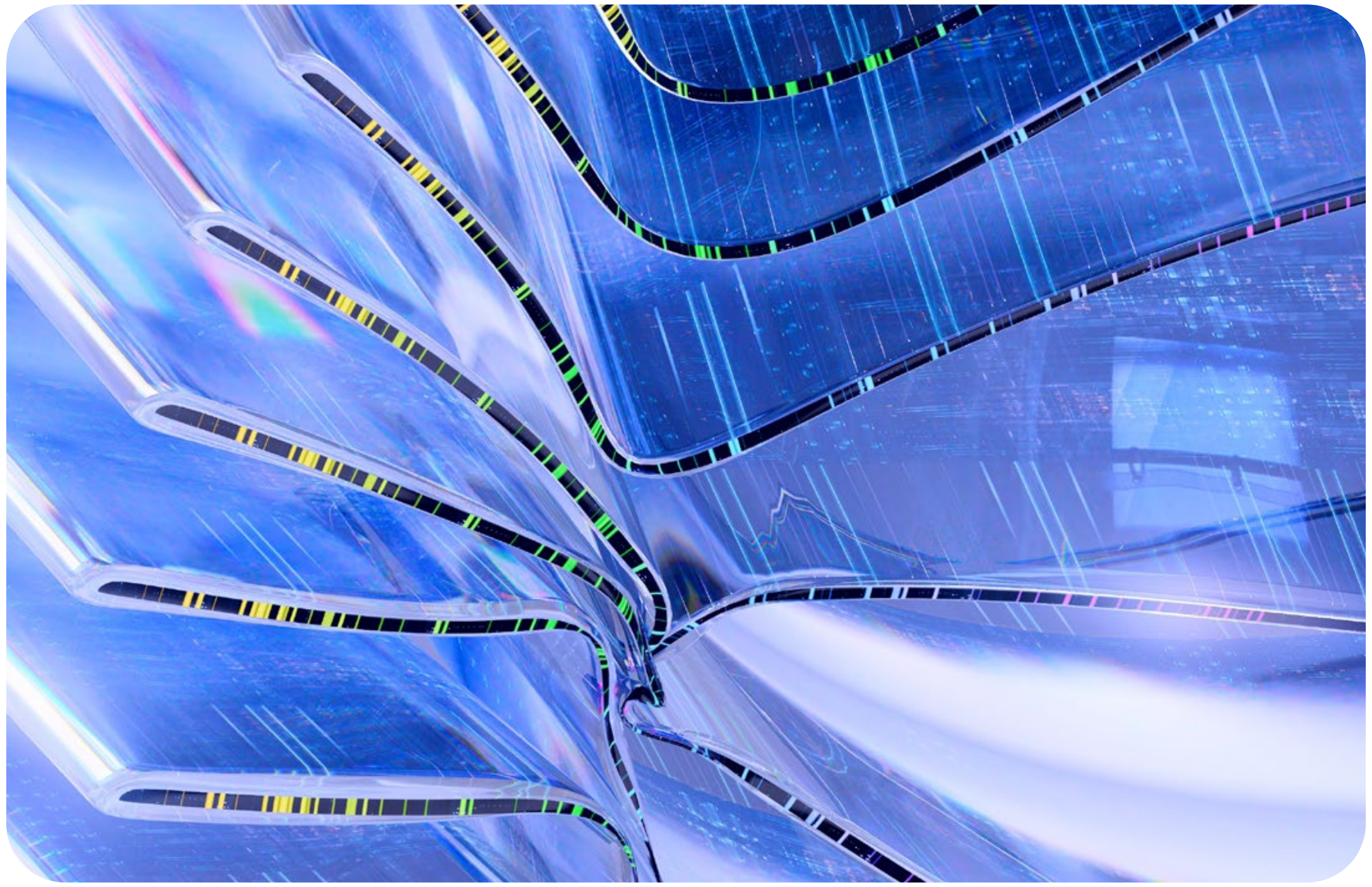
Data security
If users opt in, we use our best-in-class security infrastructure to ensure that users' data is securely connected to the Gemini App or Search AI Mode through Personal Intelligence, ensuring the data is protected even as it powers new, personal AI experiences.

Knowledge
We empower users with knowledge about Personal Intelligence, from acknowledging its limitations, to providing users with resources such as the Gemini app Help Center and the AI Mode Help Center to learn more about how Personal Intelligence interacts with their data.

Case study

Researching risks from advanced AI systems

As we push forward the frontiers of what AI is capable of, our research teams continue to study the potential risks that may emerge and how to best evaluate and mitigate them.



Preparing for AGI

In April 2025, our researchers published a proactive approach to building artificial general intelligence (AGI) safely and responsibly. The research assumes that highly capable AI could be developed by 2030 and analyzes the potential risks, from threat actors misusing AI capabilities to carry out cyberattacks against critical infrastructure, to AI systems becoming misaligned and deceiving human users. The research also considers various mitigations, such as blocking access to dangerous capabilities by using filters to prevent misuse, or using AI assistance to help maintain oversight.

New capabilities and form factors

The nature of AI risks depends on the capabilities of the underlying models, but also the form factors used to deploy these capabilities into the real world. In 2025, our team carried out research on different kinds of capabilities and form factors.

Robotics. Our Gemini robotics models are equipped with capabilities such as advanced spatial understanding, that will enable robots to perform a wider range of real-world tasks. To mitigate safety risks, we have developed an approach that combines multiple layers of safeguards, building on our ongoing safety research in this space. For example, in March 2025 we published a method for generating "constitutions," or rules of behavior, to guide robots' actions. We also partnered with Princeton University to demonstrate how to identify and predict robot failures in real-world scenarios without requiring physical hardware testing. Our industry-leading work on safety has helped make our Gemini robotics models best in class.

Agents. As new elements of AI models and systems, AI agents can act autonomously on behalf of the user — performing tasks such as researching, planning, and using tools.

In May 2025, we published a paper outlining security principles for Secure AI Agents.

In September 2025, we published research examining the impacts that may occur as AI agents become more capable and interconnected, and begin to transact with each other, in the economy at scale and speeds beyond direct human oversight. The authors propose a range of potential interventions, from identifiers for agents to sandbox environments.

In December 2025, our researchers mapped potential risks of a hypothetical future in which AGI may not emerge as a single powerful model, but rather as a distributed network of specialized, sub-AGI agents that can collectively perform complex tasks that no individual agent could do alone. In response, they recommend that safety interventions move beyond individual model alignment toward a "defense-in-depth" framework that governs the entire ecosystem through controlled agentic markets, systemic circuit breakers, and robust oversight of collective behaviors.

Focus areas

Across Google, our experts undertake and support research on a range of priority topics, from relationships and how to protect the mental wellbeing of AI users, to chemical, biological, radiological, and nuclear risks. Some recent examples include:

Cybersecurity. In March 2025, we published a framework for evaluating the offensive cyber capabilities of AI systems. This evaluation covers every phase of the cyberattack chain, addresses a wide range of threat types, and is grounded in real-world data.

Information Quality. In November 2025, we published the FACTS Leaderboard, a suite of methods to evaluate the accuracy of LLMs. It evaluates models on their ability to accurately answer different kinds of questions, including questions about images, questions that rely on using search tools, "closed-book" questions that models must answer without external tools, and questions about long-form documents.

Mental health. In July 2025, we announced our partnership with Wellcome Trust, one of the largest charities in the world, on a multi-year investment in AI research for treating anxiety, depression, and psychosis. We also worked with Grand Challenges Canada and McKinsey Health Institute to create a practical field guide for mental health organizations on how to use AI for scaling evidence-based mental health interventions.

Kids and Families. In October 2025, we announced the winners of the Google Academic Research Awards, through which we have supported research exploring critical topics, including the impact of AI on teenagers and early childhood development. In addition to the funds attached to these awards, awardees are matched to a Google research sponsor, providing direct connection to our own research community.

Case study

Stress testing our systems

As AI capabilities continue to develop, we are evolving our rigorous testing frameworks and specialized teams to address new risk profiles. By integrating human expertise with AI-assisted automation, we are ensuring these advanced systems scale safely while remaining helpful for everyone.

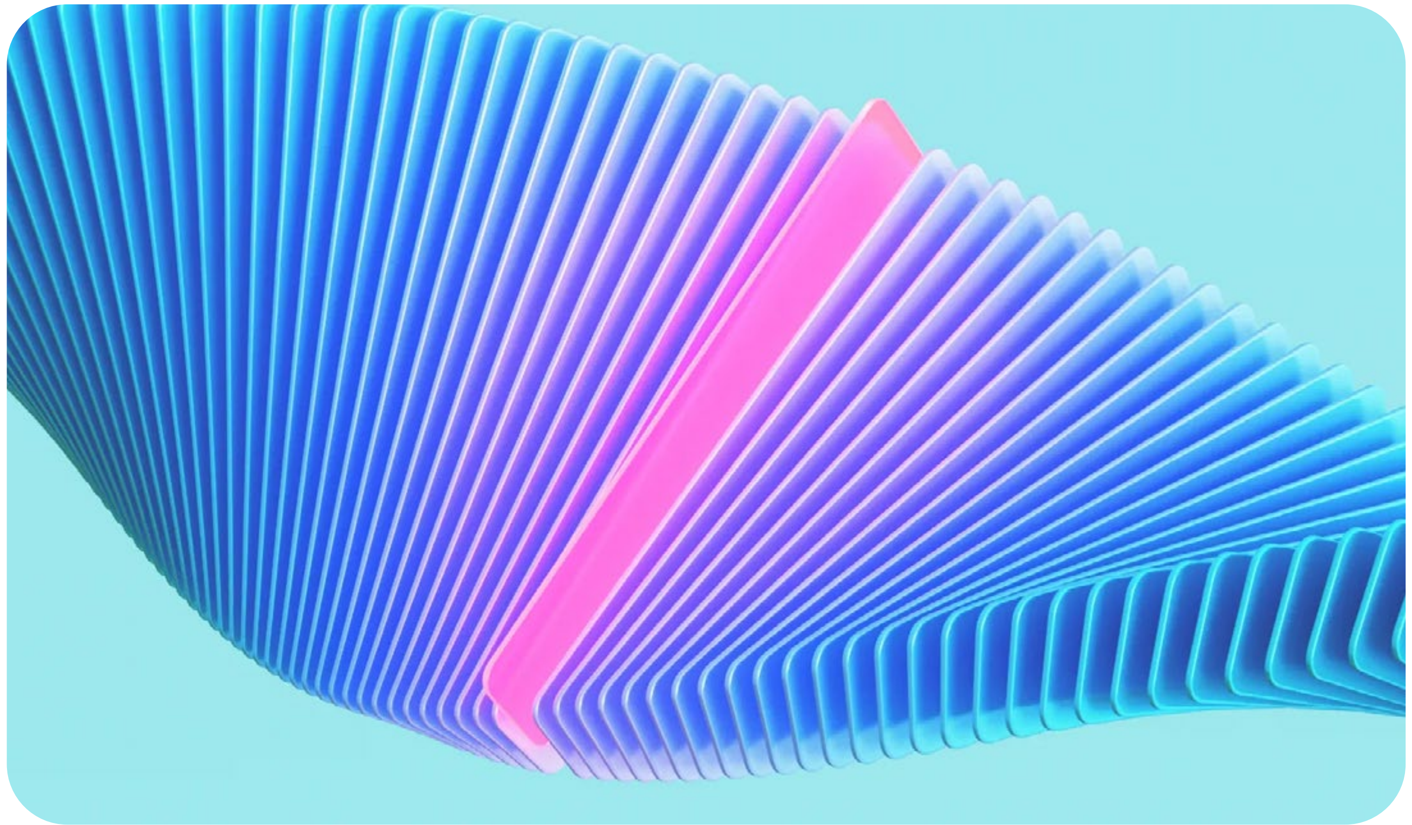
Mapping unexpected risks through adversarial red teaming
A core aspect of our testing strategy is red teaming — unstructured, adversarial testing designed to uncover unexpected risk vectors that standard evaluations might miss. Relying on lateral thinking and methodical exploration, our teams simulate how malicious actors might attempt to misuse our systems. These specialists cover a broad range of key risk areas, including child safety and content safety. In 2025 alone, our Content Adversarial Red Team (CART) completed over 350 exercises. This work spans all major modalities — including text, audio, images, and video — as well as complex capabilities like agentic AI, allowing us to map risks to stay ahead of a rapidly shifting threat landscape.

Our CART teams are experts in conducting human-driven unstructured tests at scale. To support this, we additionally deploy automated red teaming techniques to systematically explore adversarial attacks to enable a broad assessment of model vulnerabilities.

Addressing novel and emerging risks
Novel AI systems can mean there is potential for novel risks. To evaluate our most advanced frontier systems, our Novel AI Testing team was formed to spearhead evaluations at scale for new AI systems, such as advanced agents and Personal Intelligence. Within personalization testing, the team engineered a scaled approach for dynamic, context-aware evaluations.

Managing safety through collaborative scrutiny
Our internal rigor is complemented by external validation to ensure objective assessments. We partner with independent evaluators including Apollo, Vaultis, and Dreadnode, and provide early access to our models to bodies such as the UK AI Security Institute. This external scrutiny validates that our models adhere to the safety practices outlined in our updated Frontier Safety Framework, helping us to stress test our models in different risk areas, from cyber to harmful manipulation.

Ultimately, this comprehensive strategy — combining human-in-the-loop expertise with AI-assisted scale — enables data-driven safety and security assessments, and ensures that we are able to address new and emerging risks while enabling the next generation of bold AI experiences.



Approach to agentic testing

In this agentic era — where AI systems autonomously interact with services and users — we require a new testing paradigm designed specifically for these interaction-based risks. To ensure our testing keeps pace with the speed of product innovation, we are evolving our capabilities to be more authentic, automated, and actionable.

The sandbox
We developed an authentic, interactive sandbox environment that replicates complex, multi-turn digital user experiences and state-of-the-art attacks. This platform looks to address critical safety, legal, and scalability challenges inherent in live internet testing of agentic products, and allows us to proactively identify and mitigate high harm risks without exposing the public web to potential harm.

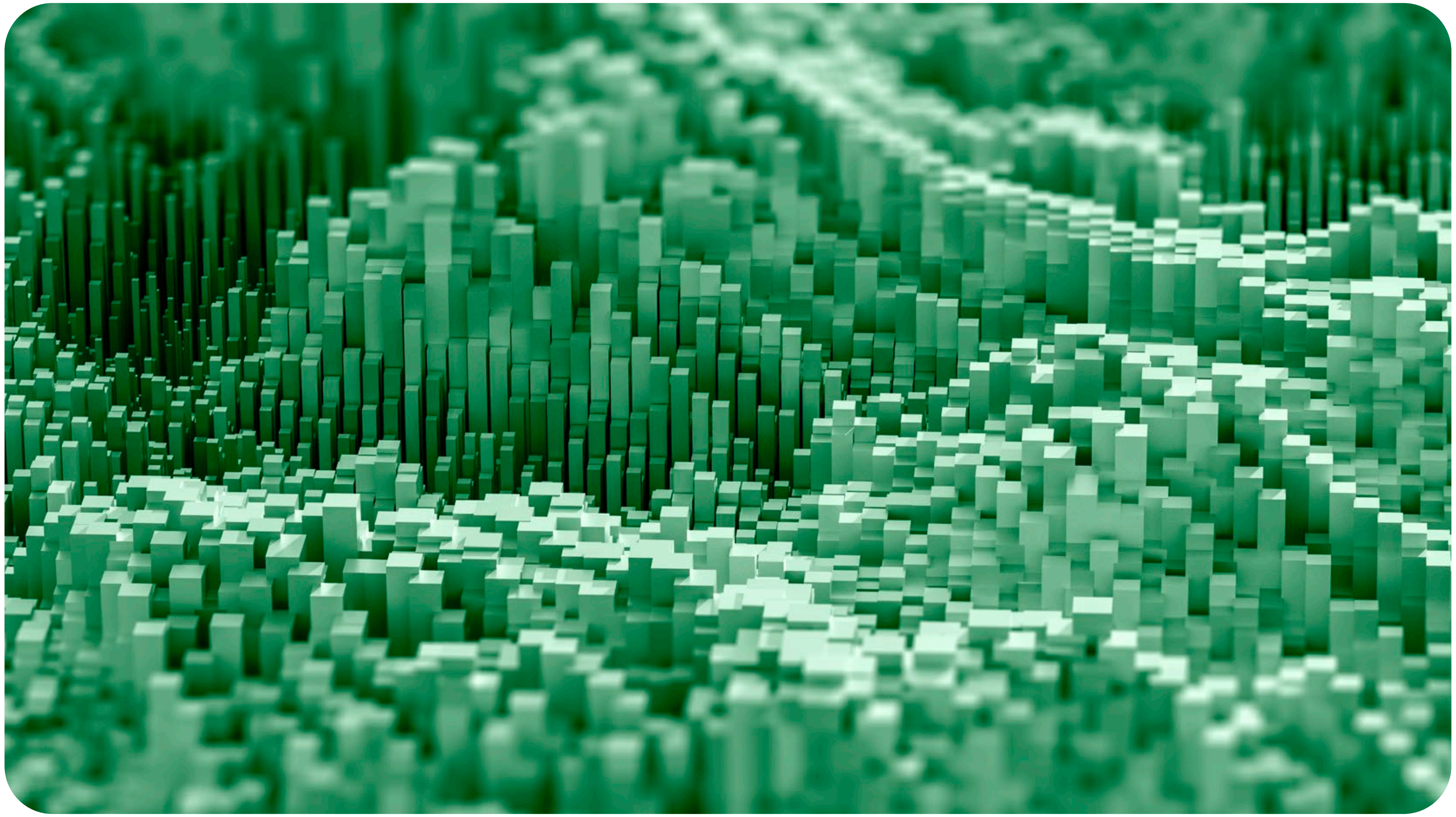
“Buddy Agents”
We are currently implementing automated monitoring agents that log interactions and assess compliance in real-time of the agent being tested.

Multi-turn interactions
We are developing the capability to provide insights into how agents perform in complex, multi-turn interactions using personalized data – allowing us to accurately evaluate the intersection of multiple novel capabilities as they converge.

How we apply AI to benefit society

We believe that bold and innovative AI requires a responsible foundation that safely minimizes potential risks, while maximizing the extraordinary opportunities AI presents — from unlocking new growth across the global economy, to improving the treatment of diseases, to accelerating scientific timelines from years to months. True responsibility in the AI era goes beyond safeguards. It demands that we use our scale and infrastructure to help AI address society's most pressing challenges.

Working alongside our partners, we take a comprehensive and responsible approach to AI development that enables our most advanced models to be applied where they can have the greatest positive impact.



Accelerating scientific progress
We are fostering a new golden age of discovery by applying AI to fundamental sciences. This includes advancing nuclear fusion research and utilizing quantum computing to solve problems that were previously intractable. Alongside this, we are creating tools like AI co-scientist that help scientists generate novel hypotheses to accelerate the speed of scientific discoveries.

Improving global health
We are driving progress in genomics and disease detection, automating administrative burden for clinicians, and partnering with institutions like Yale University to discover new potential cancer therapy pathways. Through AI tools like AlphaFold, which predicts protein structures, we are accelerating our understanding of disease — enabling drug discovery and opening new frontiers in diagnostics and treatment.

Strengthening resilience
We are strengthening global resilience by providing responsible agencies with experimental tools that give earlier warning for floods, cyclones, and earthquakes. When used, these tools can help communities prepare for and respond to disasters more effectively. Beyond crisis resilience, we are applying AI to high-accuracy weather forecasting and identifying techniques for sustainable agriculture, helping society adapt to the realities of a changing climate.

Supporting education
We are committed to empowering every learner. AI is transforming education by supporting teachers and personalizing learning experiences. Through initiatives focused on AI literacy and learning, and technologies like LearnLM, we are helping to unlock human potential and make knowledge universally accessible.

Case study

Accelerating scientific progress through specialized AI agents

Our commitment to accelerating scientific discovery is operationalized through specialized AI systems that empower researchers to navigate vast datasets and overcome historical barriers. Practical breakthroughs such as AlphaGenome and AlphaEvolve demonstrate the tangible results of applying scientific methodology to bold AI innovation.

AlphaGenome

AlphaGenome is an AI model specifically designed to further decode the human genome. By analyzing up to 1 million DNA letters at once, it provides a unified system for predicting how genetic mutations interfere with gene regulation — the complex process that dictates when, where, and how much individual genes are activated. While science has historically focused on the 2% of the genome that codes for proteins, AlphaGenome unlocks the remaining 98% of non-coding regions where many disease-linked variants reside. These capabilities are already helping researchers at University College London and Memorial Sloan Kettering Cancer Center better identify the genetic drivers of rare diseases and cancers.

We believe AlphaGenome will be an important tool for better understanding the genome function and disease biology, and ultimately, drive new biological discoveries and the development of new treatments. We're committed to working alongside external experts across academia, industry, and government organizations to ensure AlphaGenome benefits as many people as possible.

> *"This tool will provide a crucial piece of the puzzle, allowing us to make better connections to understand diseases like cancer."*
>
> **Professor Marc Mansour**
> University College London

AlphaEvolve

AlphaEvolve is an evolutionary coding agent specifically designed for general-purpose algorithm discovery and optimization. AlphaEvolve has the potential to create broad impact across a range of sectors, with real-world applications already being implemented at Google and in the research community.

At Google, we were able to use AlphaEvolve to design algorithms that enhanced the efficiency of our data centers, improved the design of our Tensor Processing Units and even accelerated AI training processes — including the processes used to train the Gemini models powering AlphaEvolve. Across the scientific community, AlphaEvolve has already demonstrated its potential as a research partner, pushing the boundaries of our understanding and helping advance the fields of mathematics, computer science, quantum, and nuclear fusion.

Given the potential for new algorithms to continue solving key societal challenges, AlphaEvolve has the capability to be transformative in fields as diverse as material science, biotech and pharma, energy, financial services, and logistics.

Case study

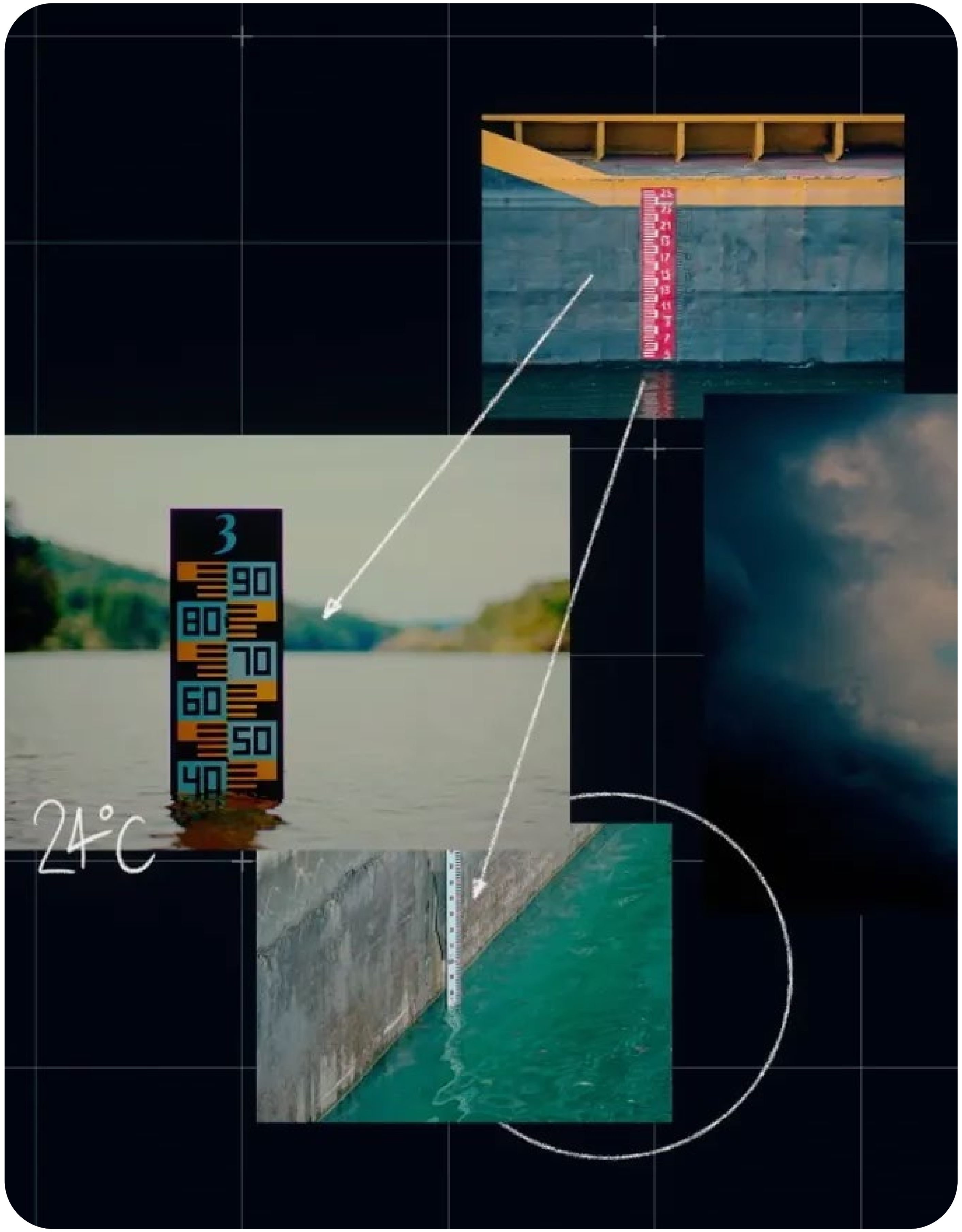
Using AI for reliable flood forecasting at a global scale

We are committed to using AI to enhance global disaster preparedness and response. Our development of specialized flood forecasting systems is just one example of how AI can help handle the growing complexities of a changing climate and strengthen the resilience of impacted communities.

Forecasting floods with AI
Floods are the world's most common natural disaster. Our flood forecasting initiative leverages advanced AI to provide free, real-time riverine flood warnings up to seven days in advance. By training a global AI model on extensive public streamflow data, we have extended reliable forecasting to data-scarce regions — particularly in low- and middle-income countries across Africa and Asia — that previously lacked traditional gauge networks. This dual-AI system combines hydrologic and inundation models to provide coverage for significant flood events impacting over 2 billion people across 150 countries.

Central to this initiative is our deep integration with local governments, intergovernmental organizations, and NGOs to ensure technical breakthroughs translate into life-saving action. By collaborating with these partners, we are able to integrate our forecasts directly into existing early warning infrastructure through the Google Flood Hub and our flood forecasting API. A key example of this collaborative model is our partnership with the nonprofit GiveDirectly in Nigeria, where Google's forecasts were used to trigger anticipatory cash transfers. This collaboration allowed over 3,250 households to evacuate and secure assets before flooding hit, resulting in a 90% drop in food insecurity and demonstrating how public-private-NGO partnerships can create measurable climate resilience in the world's most vulnerable communities.

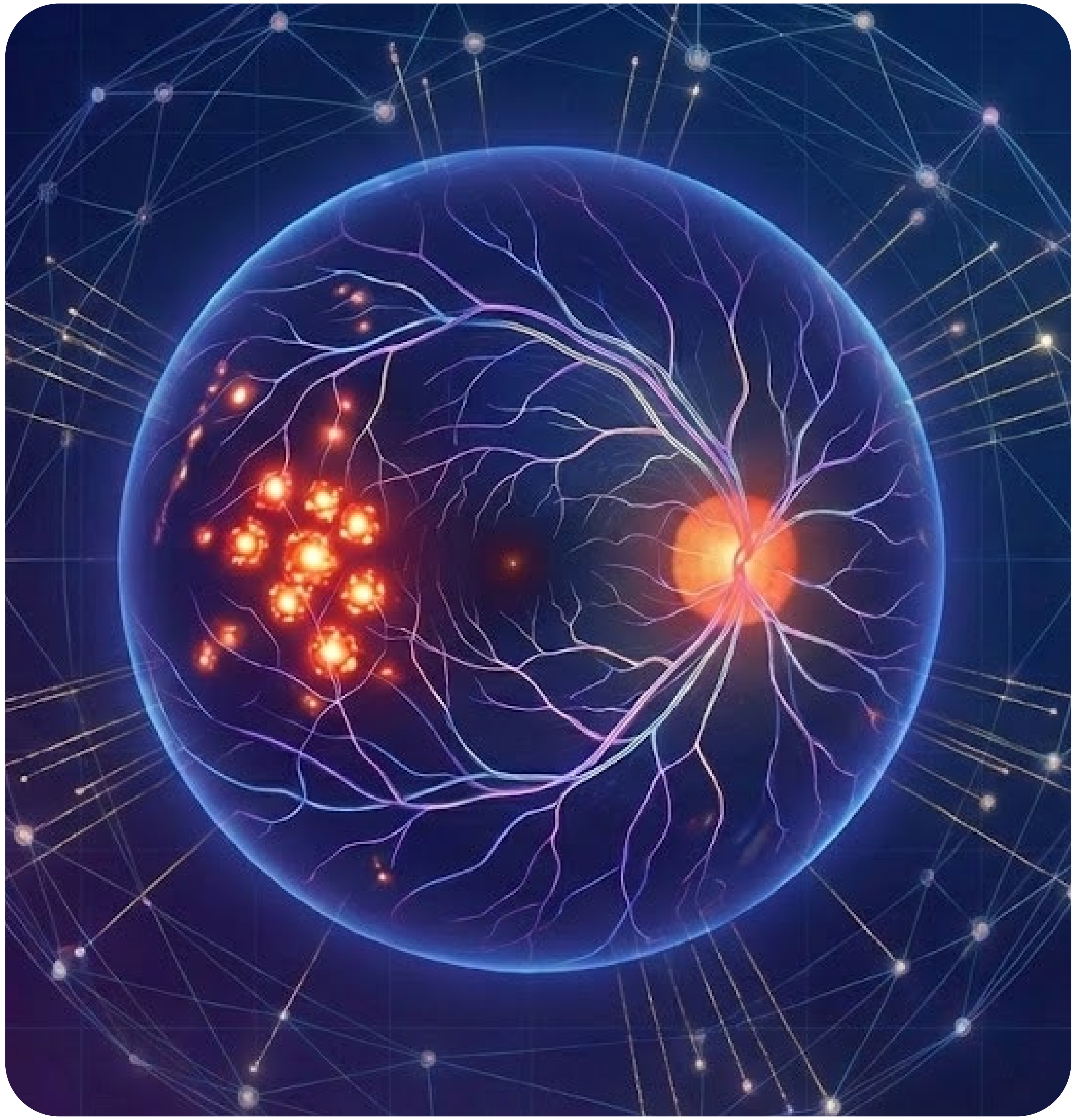
A transformative partnership for anticipatory action
In a major step forward for disaster resilience, Nigeria launched its first AI-driven large-scale Floods Anticipatory Action Program. Led by the UN Country Team and the UN Office for the Coordination of Humanitarian Affairs, in collaboration with the government of Nigeria, this $7 million initiative protected vulnerable populations by delivering aid before disaster struck. The program utilized a sophisticated multi-source trigger system — integrating Google Flood Forecasts and national data. When a flood is predicted, the system automatically activates a suite of pre-emptive measures, such as aid distribution, shelter preparation, and the pre-positioning of livestock feed. Following an initial activation in 2025 based on national data, the program is slated for further expansion in 2026.



Case study

Preventing blindness through specialized AI screening

AI-based solutions can help improve healthcare for millions of people around the globe. Our work to help address preventable blindness by empowering local experts and communities provides a sustainable framework for leveraging innovative AI to improve health outcomes.



From research to global impact
Diabetes affects over 500 million adults globally, and for nearly half of them, the disease carries a hidden threat: diabetic retinopathy. This condition can lead to total blindness if left untreated, making early detection a matter of life-changing urgency. Yet, for millions of people — particularly in regions with few medical specialists — regular screenings remain out of reach. Since 2015, we have focused on bridging this gap by applying AI to the challenge of preventable blindness. Following a landmark 2016 study published in JAMA that confirmed our model's high diagnostic accuracy, the model was deployed globally. To date, this specialized AI has supported nearly 1 million screenings, acting as a vital assistant in the global effort to preserve vision.

The path of responsible innovation
Because this initiative impacts human health, prioritizing a safe and responsible rollout was essential. Our process began with retrospective validation alongside medical professionals to confirm model accuracy, followed by a large-scale prospective study to verify performance within target populations in India and Thailand. To uphold rigorous medical device standards, we secured a CE marking and conducted a health economic study demonstrating the approach's financial viability.

Our final phase pre-rollout focused on ensuring local scalability by empowering health-tech firms to expand their services through AI. We externalized our model to healthcare providers and local partners, including Forus Health, AuroLab, and Perceptra, who obtained the necessary regulatory clearances for integrating the model into clinical care systems in India and Thailand.

Empowering local communities
A key pillar of our responsible approach is ensuring that AI serves as a collaborative partner to existing healthcare infrastructure. By empowering local health-tech firms to expand their own services through our AI, we ensure the technology is both scalable and culturally relevant.

This commitment to equity is also seen in our recent work with the Lions Eye Institute in Australia. Here, the technology is being used to bolster screening for Aboriginal communities in rural areas, where medical access is often limited. By bringing "expert eyes" to remote regions, we are ensuring that the benefits of AI are shared by those who need them most.

By transforming specialized AI into a collaborative tool for local experts, we can build a sustainable, global framework for managing preventable disease. This initiative embodies our commitment to bold, innovative AI — ensuring that technical breakthroughs in the digital world create tangible, life-saving action for communities worldwide.

How we drive responsible progress together

We understand that the scale of the AI opportunity — and the complexity of its challenges — demand deep cooperation. We believe that innovating with AI to deliver widely accessible benefits to people and society, while mitigating its risks, must be a collective effort involving us and others — including researchers, developers, users (individuals, businesses, and other organizations), governments, regulators, and citizens. This involves actively engaging with these partners, and sharing our technical solutions, knowledge, and research with the broader ecosystem.

As part of our commitment to improving AI safety and security across the ecosystem, we develop relevant and robust industry partnerships and programs. Whether through training millions of people in digital skills, providing developers with state-of-the-art infrastructure, releasing open-weight AI models, or through open collaboration on safety benchmarks, our goal is to ensure that the benefits of the AI era are accessible to everyone.

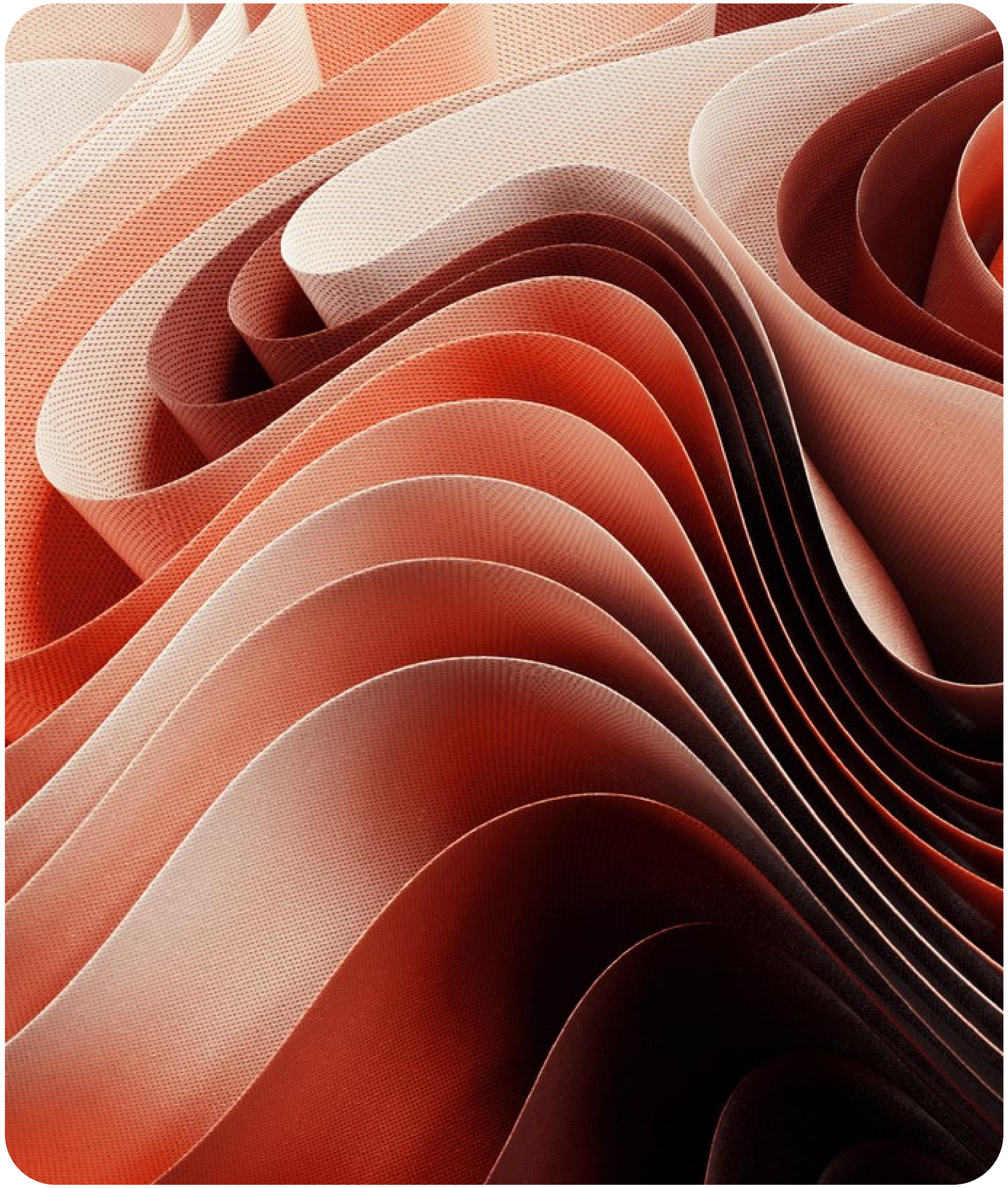
We focus our efforts in core areas that support the creation of a responsible ecosystem

Advancing innovation through ground-breaking research
We believe that collaborations between academia and industry are critical to pushing the boundaries of innovation, including in research on emerging AI risks. We actively collaborate with researchers to support the development of bold and responsible AI. By partnering with leading labs and institutions, we aim to accelerate scientific breakthroughs in areas such as biology and medicine, while simultaneously developing the technical safeguards needed to manage frontier risks.

Shaping the future with governments and civil society
We are committed to engaging with governments and civil society to address AI-related risks and advocate for international standards. We believe that companies and governments should work together to utilize AI to advance innovation, scientific discovery, and security for all.

Empowering the global community
We look to foster a vibrant and safe ecosystem that empowers our users to build and utilize innovative tools and solutions. With AI transforming industries from healthcare to education, we are dedicated to preparing the workforce and ensuring wide accessibility to these tools.

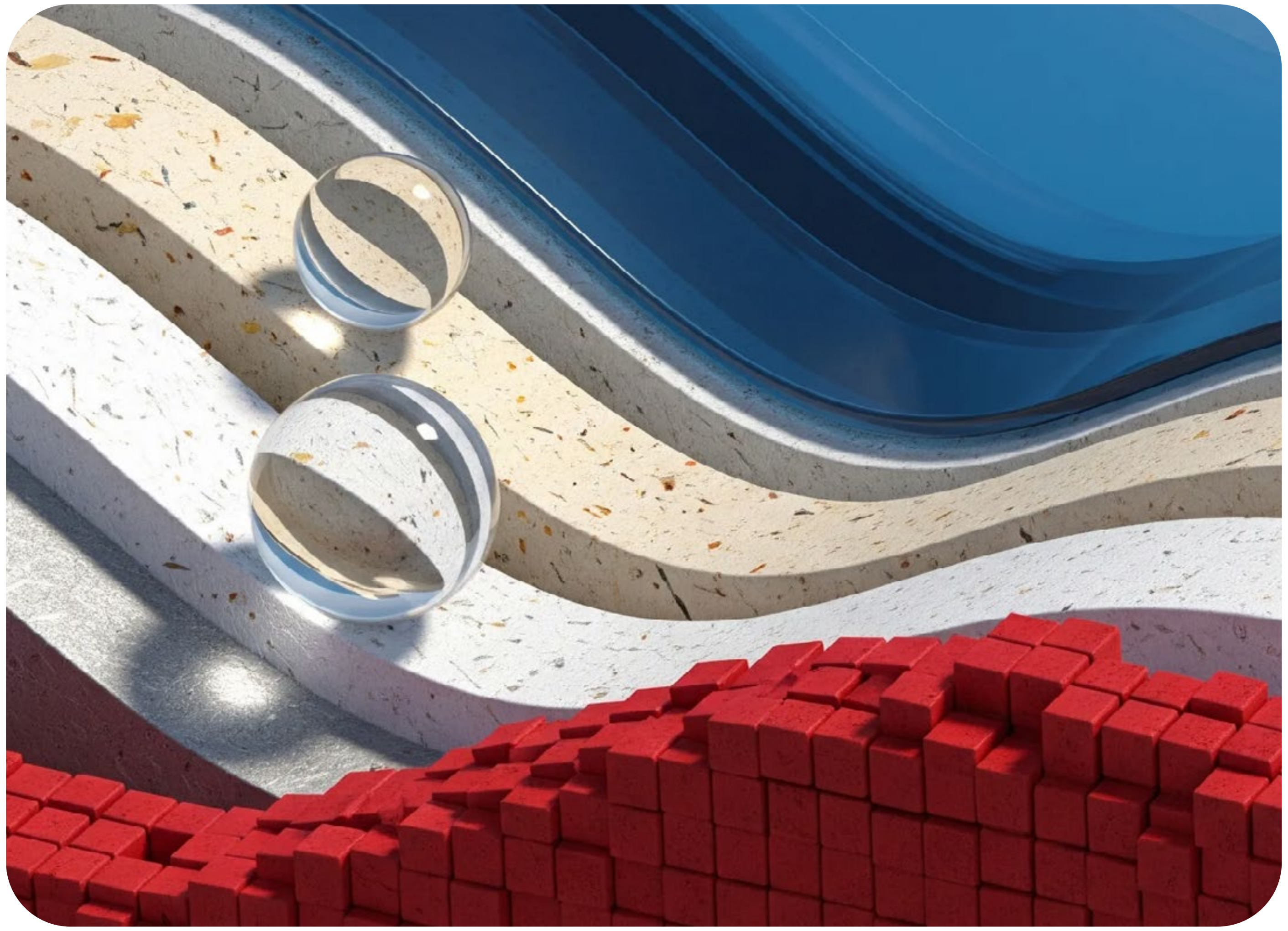
We understand the importance of strong partnerships with others to create a responsible ecosystem where innovation can thrive, and we remain steadfast in our commitment to fostering a digital environment that is safe, secure, and beneficial for all.



Case study

Accelerating progress in science, security, and education in the UK

Partnership across the AI ecosystem can take many forms, and we strongly believe that industry working closely with governments can lead to meaningful progress. One example of this in action is our partnership with the UK government. This partnership is centered on providing access to frontier AI in two key areas: scientific discovery and education.



Partnering on research, testing, and education

Our collaboration with the UK government is manifested in tangible ventures that look to catalyze progress in these fields. Some of our areas of collaboration include:

Frontier AI access
We are providing priority access to our most powerful "AI for Science" models to UK scientists. These include models such as AlphaEvolve, AlphaGenome, AI co-scientist, and WeatherNext.

Automated materials science laboratory
We will establish our first automated laboratory in the UK in 2026. The lab will focus on materials science, and use a full integration with Gemini to direct world-class robotics in materials science research, significantly shortening traditional research timelines.

Enhancing education
We are supporting research to understand how AI tools impact teaching and learning through a rigorous scientific approach, and exploring how to tailor our Gemini model to complement England's national curriculum.

This collaboration serves as a blueprint for how public and private sector organizations can unite to unlock the transformative potential of AI, with an aim to deliver tangible advancements in science, education, and safety.

Partnering with the AI Security Institute

Our partnership with the UK AI Security Institute (AISI) includes a Memorandum of Understanding focused on foundational security and safety research. This formal agreement facilitates a collaborative approach to creating a scientific understanding of advanced AI risks, and includes access to proprietary models, joint reports and publications, technical cooperation, and collaborative research.

Our joint research will help push the boundaries of AI safety and security in critical areas such as:

Monitoring reasoning processes
We will continue our research and collaboration with AISI on developing techniques to monitor an AI system's "thinking," or its chain-of-thought (CoT), to understand how it produces answers.

Assessing social and emotional impact
We will build on our existing research on how model misalignment can negatively impact human wellbeing.

Evaluating economic impact
We will research AI's potential impact on the economy by simulating real-world tasks across different environments, helping understand potential downstream impacts of AI on the labor market and in other economic areas.

Case study

Enabling the ecosystem with provenance tools and standards

We are building a holistic approach that empowers users with the tools and information to make informed decisions about the content they encounter online — from the hardware where an image is captured, to the context around an image or video's creation and use. Addressing these challenges requires robust, scalable, cross-industry solutions that provide transparency without hindering innovation.

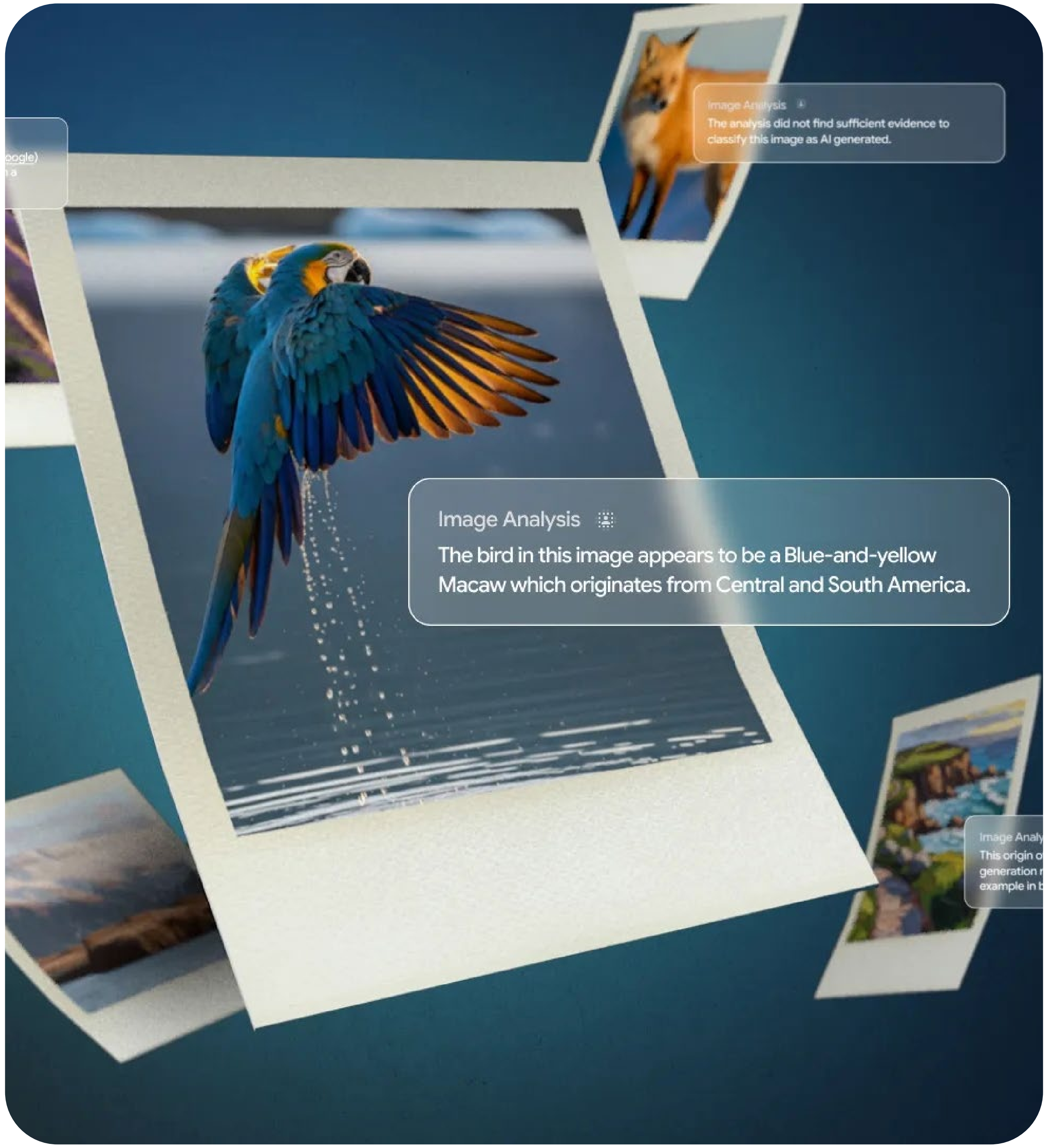


Image Analysis
The analysis did not find sufficient evidence to classify this image as AI generated.
Image Analysis
The bird in this image appears to be a Blue-and-yellow Macaw which originates from Central and South America.

SynthID
We developed SynthID to help address issues around provenance by embedding digital watermarks directly into AI-generated content — including text, audio, images, and video. We've also open-sourced the text watermarking SynthID technology to make it easier for any developer to apply watermarking for their own generative text models. These watermarks are imperceptible to humans but are detectable by specialized software. In 2025, we announced SynthID Detector, a verification portal that allows users to quickly and efficiently upload content and scan for SynthID watermarks, and identify if content was generated or modified by Google's AI tools. We also made SynthID verification available directly in the Gemini app, meeting users where they are and enabling them to directly upload images and videos in the Gemini app to check if they were generated by AI.

Backstory
While watermarking can verify AI-generated content, understanding the history of an image offers a different layer of transparency. Backstory is an experimental AI tool that helps identify whether an image was AI-generated, even without watermarks, or whether it is authentic but presented in a misleading context. It then investigates how it has been used on the internet and provides other helpful metadata. The result is a holistic assessment of the image that can help determine its integrity, based on the context surrounding the image's use and presentation.

C2PA
Google is actively working with industry partners to develop and implement the Content Credentials standard developed by the Coalition for Content Provenance and Authenticity (C2PA). We provided substantial contributions to the latest version (2.1) of the C2PA standard to ensure it meets the complex needs of modern hardware and software. Furthermore, we continue to integrate these standards directly into our hardware and services, such as Pixel 10, which is the first phone to implement content credentials within its native camera app, and the inclusion of C2PA metadata into images generated by our generative AI image model, Nano Banana Pro.

Case study

Enhancing AI security across the ecosystem

As the cybersecurity landscape evolves with the rise of agentic AI, we are advancing a comprehensive strategy that combines autonomous tooling with updated governance frameworks. We are operationalizing this approach by deploying AI agents to proactively fix vulnerabilities, and expanding our Secure AI Framework (SAIF) to address the unique risks of autonomous systems. Crucially, we augment these internal defenses by incentivizing the global research community to stress-test our models against emerging threats.

Building AI tools for autonomous cyber defense
Our AI-based efforts such as Big Sleep and OSS-Fuzz have demonstrated AI's ability to find new vulnerabilities in well-tested, widely used software before it's even released. As we achieve further breakthroughs in speed and scale with AI-powered vulnerability discovery, it will become increasingly difficult for humans alone to keep up. We developed CodeMender to help tackle this. CodeMender is an AI agent using the advanced reasoning capabilities of our Gemini models to automatically fix critical code vulnerabilities. CodeMender scales security, reducing the time it takes to patch vulnerabilities across the open-source landscape, representing a major leap in proactive AI-powered defense.

Helping secure AI agents across the ecosystem
We expanded our Secure AI Framework to address the rapidly emerging risks posed by autonomous AI agents. SAIF 2.0 extends our proven AI security framework with new guidance on agent security risks and controls to mitigate them. It is supported by three new elements:

1. An agent risk map to help practitioners map agentic threats across the full-stack view of AI risks.
2. Security capabilities rolling out across Google agents to ensure they are secure by design and apply our three core principles.
3. Our donation of SAIF's risk map data to the Coalition for Secure AI Risk Map initiative, which provides a structured map of the AI security landscape and a common language for addressing vulnerabilities. As a founding member of the Coalition for Secure AI, we are actively engaged in promoting cross-industry security principles for agentic systems.

Incentivizing critical security research
As AI systems become more complex, the global security research community remains an indispensable partner in our defense-in-depth strategy. While internal red teaming and other safety and security evaluations are critical, we believe that a diverse, global network of researchers incentivized to find novel flaws provides another layer of stress testing that helps make our products safer. In 2025 we launched a dedicated AI Vulnerability Reward Program (AI VRP). Building on the success of our existing Vulnerability Rewards Program, this dedicated AI VRP recognizes that securing generative AI requires specialized research distinct from traditional vulnerabilities and provides updated rules and scope to offer clearer incentives for research into high-impact issues such as rogue actions, data exfiltration, and context manipulation.

Conclusion

The AI era is no longer a distant promise; it is a present reality that is beginning to unlock extraordinary opportunities for society. From faster scientific discoveries to the transformation of daily business, we are witnessing a once-in-a-generation moment to reimagine what is possible.

As we continue to advance our AI models, our governance must remain as dynamic as the technology itself. We are committed to being bold in our innovation, responsible in our development, and collaborative in our progress.

There is no finish line in responsible AI. By sharing our lessons, empowering the ecosystem, and adhering to our core AI Principles, we will work to make AI a force that meaningfully improves the lives of people everywhere.