# Report

## Organization: **Google**

https://www.oecd.org/en/about/news/press-releases/2024/07/oecd-launches-pilot-to-monitor-application-of-g7-code-of-conduct-on-advanced-ai-development.html

**Publication date:** Apr 22, 2025, 09:01 AM PDT

**Reporting period:** 2025

# Section 1 - Risk identification and evaluation

## a. How does your organization define and/or classify different types of risks related to AI, such as unreasonable 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.

We've codified our risk-mapping work into a taxonomy of potential risks associated with AI, building on industry guidelines such as the NIST AI Risk Management Framework, 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.

Our Frontier Safety Framework is a set of protocols that aims to address severe risks that may arise from powerful capabilities of foundation models. It is intended to complement Google's existing suite of AI responsibility and safety practices. The Framework is built around capability thresholds called "Critical Capability Levels (CCLs)." In the Framework, we describe two sets of CCLs: misuse CCLs that can indicate heightened risk of severe harm from misuse if not addressed, and deceptive alignment CCLs that can indicate heightened risk of deceptive alignment-related events if not addressed.

### b. What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks and misuse, throughout the AI lifecycle?

After identifying and understanding risks through mapping, we systematically assess our frontier AI models and systems. We evaluate how well our frontier 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.

We design our applications to promote user feedback on both quality and safety and our teams monitor user feedback via these and other channels closely. 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.

Our Frontier Safety Framework describes a set of evaluations called "early warning evaluations," with a specific "alert threshold" that flags when a CCL may be reached for a frontier model before the evaluations are run again. In our evaluations, we seek to equip the model with appropriate scaffolding and other augmentations to make it more likely that we are also assessing the capabilities of systems that will likely be produced with the model. We may run early warning evaluations more frequently or adjust the alert threshold of our evaluations if the rate of progress suggests our safety buffer is no longer adequate. Where necessary, early warning evaluations may be supplemented by other evaluations to better understand model capabilities relative to our CCLs. We may use additional external evaluators to test a model for relevant capabilities, if evaluators with relevant expertise are needed to provide an additional signal about a model's proximity to CCLs.

### c. Describe how your organization conducts testing (e.g., red-teaming) to evaluate the model's/system's fitness for moving beyond the development stage?

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

A subset of the mapped risks mentioned in our previous answer is relevant to test at the frontier model level. We evaluate the models for risks such as self-proliferation, offensive cybersecurity, child safety harms, and persuasion.

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. 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.

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.

Our Content Adversarial Red Team (CART) proactively identifies weaknesses in our AI systems, enabling us to mitigate risks before product launch. Our internal AI tools also assist human expert red teamers and increase the number of attacks they're able to test for.

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.

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.

Application 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.

### d. Does your organization use incident reports, including reports shared by other organizations, to help identify risks?

Yes

### e. Are quantitative and/or qualitative risk evaluation metrics used and if yes, with what caveats? Does your organization make vulnerability and incident reporting mechanisms accessible to a diverse set of stakeholders? Does your organization have incentive programs for the responsible disclosure of risks, incidents and vulnerabilities?

We evaluate how well our frontier 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. We evaluate the models for risks such as self-proliferation, offensive cybersecurity, child safety harms, and persuasion.

We conduct ongoing fundamental research into new evaluation methods for different kinds of risks from LLMs - such as FactsGrounding, a comprehensive benchmark for evaluating the ability of LLMs to generate responses that are not only factually accurate with respect to given inputs, but also sufficiently detailed to provide satisfactory answers to user queries.

We have a close relationship with the security research community. To honor all the cutting-edge external contributions that help us keep our users safe, we've maintained a Vulnerability Reward Program (mentioned in 1.c and 1.d) for Google-owned and Alphabet (Bet) subsidiary web properties, running continuously since November 2010. We recently updated this program to specifically clarify and encourage reporting of issues in our AI products. We released a 2024 year in review of our Rewards program that confirmed the ongoing value of engaging with the security research community to make Google and its models and products safer.

### f. Is external independent expertise leveraged for the identification, assessment, and evaluation of risks and if yes, how? Does your organization have mechanisms to receive reports of risks, incidents or vulnerabilities by third parties?

Yes. 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 evaluations, where appropriate, are conducted by independent external groups on our frontier models. The design of these evaluations is independent and results are reported periodically to the internal team and governance groups. Results are used to mitigate risks and improve evaluation approaches internally. For Gemini 1.5, for example, external groups, including domain experts and a government body, designed their own methodology to test topics within a particular domain area. The time dedicated to testing also varies per group, with some groups working full-time on executing testing processes, while others dedicated one to two days per week. Some groups pursue manual red-teaming and report on qualitative findings from their exploration of model behavior, while others develop bespoke automatic testing strategies and produce quantitative reports of their results.

We continue to invest in independent research on AI risks. For example, we co-created an AI Safety Fund, initially funded with $10 million to support independent researchers from around the world (academic institutions, research institutions, startups, etc.). The goal of the fund is to allow researchers to better evaluate and understand frontier systems, and – ultimately – develop new model evaluations and techniques for red-teaming AI models.

As described in our response to question 1.d., we signed a first-of-its-kind agreement with fellow members of the Frontier Model Forum (FMF), designed to facilitate information-sharing about threats, vulnerabilities, and capability advances unique to frontier AI.

### g. Does your organization contribute to the development of and/or use international technical standards or best practices for the identification, assessment, and evaluation of risks?

We support the wider ecosystem with AI safety practices and standards by participating in working groups within global organizations including MLCommons, the World Economic Forum's (WEF) AI Governance Alliance, the Coalition for Content Provenance and Authenticity (C2PA), Thorn, Partnership on AI (PAI), Frontier Model Forum, and the U.K. AI Safety Institute, among others. Examples of our contributions include:

- With PAI we jointly launched responsibility frameworks for safe deployment, synthetic media, and data enrichment sourcing, among other guidance for AI risk identification and mitigation, alongside industry peers, academics, governments and civil society organizations.
- We contributed to WEF's Industries in the Intelligent Age white paper series.
- We have 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.
- We are a founding member of MLCommons, an engineering consortium focused on AI benchmarks, including the  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 contributed to ISO 42001, an international standard that specifies requirements for establishing, implementing, maintaining, and continually improving an Artificial Intelligence Management System (AIMS) within organizations.
- 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 submittor's AI systems, as well as suggested mitigations, based on the responses they provided.
- We make public many of our best practices for identifying, assessing, and evaluating risk via the Responsible 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.

### h. How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks?

We promote industry collaboration on the development of standards and best practices for risk mitigation measures. We work with relevant stakeholders (as described in 1.g.), such as the Frontier Model Forum, Partnership on AI, and ML Commons, across sectors to assess and adopt risk mitigation measures. For example:

- The Frontier Model Forum releases regular publications that reference and support Google research, such as their recent publication on Thresholds for Frontier AI Safety Frameworks.
- In 2024 we launched the Coalition for Secure AI (CoSAI) with industry partners. This is the first major milestone and application of Google's Secure AI Framework (SAIF). The coalition will collectively invest in AI security research, share security expertise and best practices, and build technical open source solutions. CoSAI is an open source initiative designed to give all practitioners and developers the guidance and tools they need to create AI systems that are Secure-by-Design. The coalition will operate under the guidance of OASIS Open, the international standards and open source consortium. Founding members include: Amazon, Anthropic, Cisco, Cohere, GenLab, Google, IBM, Intel, Microsoft, Nvidia, Open AI, Paypal and Wiz.

### Any further comments and for implementation documentation

With respect to question 1.d (Does your organization use incident reports, including reports shared by other organizations, to help identify risks?), as referenced in 1.b. - 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, 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.

We updated our Vulnerability Reward Program to specifically clarify and encourage reporting of issues in our AI products.

Through our Frontier Model Forum membership, along with other member firms, we have signed a first-of-its-kind agreement designed to facilitate information-sharing about threats, vulnerabilities, and capability advances unique to frontier AI. Our Detection & Response team provides 24/7/365 monitoring of Google products, services and infrastructure – with a dedicated team for insider threat and abuse.

# Section 2 - Risk management and information security

## a. What steps does your organization take to address risks and vulnerabilities across the AI lifecycle?

Because we understand that AI, as a still-emerging transformative technology, involves 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. 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.

We employ guardrails in our frontier models and products to reduce the risk of generating harmful content. For example we steer our frontier models to produce content that aligns with our safety guidelines by using system instructions – prompts that tell the frontier model how to behave when it responds to user inputs; and we fine-tune our frontier models to produce helpful, high-quality answers that also align to our safety guidelines.

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. 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.

Our Frontier Safety Framework discusses how our approach to risk mitigations with frontier models informs an appropriate response plan. For example, with misuse risks, we have security mitigations intended to prevent the exfiltration of model weights and deployment mitigations (such as safety fine-tuning and misuse filtering, detection, and response) intended to counter the misuse of critical capabilities in deployments. For deceptive alignment risk, automated monitoring may be applied to detect and respond to any deceptive behaviors for models in this category. Overall, the mitigations described in our Framework are aimed toward addressing severe risks from powerful capabilities. Other risk management and security considerations may result in more stringent mitigations applied to a particular model than the mitigations specified by our Framework if such considerations are intended to address a different scope of risks. Our approach is continually evolving, incorporating new measurement techniques as they become available.

### b. How do testing measures inform actions to address identified risks?

We rigorously test our frontier models and infrastructure at every layer of the stack, combining the best of AI with our world class teams of safety experts. This end-to-end approach enables advanced AI experiences that put safety first.

Red-teaming exercises, conducted both internally and externally, proactively assess advanced 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. More details on our approach to testing can be found in our response to question 1.c.

### c. When does testing take place in secure environments, if at all, and if it does, how?

Testing is performed in appropriately secured environments. Our frontier models are developed, trained, tested and stored within Google's infrastructure, where we can apply our Secure AI Framework (SAIF) throughout the AI lifecycle, which is supported by central security teams and by a security, safety and reliability organization consisting of engineers and researchers with world-class expertise. Any external trusted testing groups also receive secure access to test models.

### d. How does your organization promote data quality and mitigate risks of harmful bias, including in training and data collection processes?

As outlined in our AI Principles, we are committed to employing rigorous design, testing, monitoring, and safeguards to mitigate unintended or harmful outcomes and avoid unfair bias.

Training data, including from publicly available sources, reflects a diversity of perspectives and opinions. We continue to research how to use this data in a way that ensures that an LLM's response incorporates a wide range of viewpoints, while minimizing inaccurate overgeneralizations and biases.

We apply safety filtering to our pre-training data for frontier models for our strictest policies. For a subset of our training data, we add control tags, e.g., based on classifier-annotated labels of the text's toxicity, similar to (Anil et al., 2023b). These tags can help structure the learned representation to make post-training for safety easier.

### e. How does your organization protect intellectual property, including copyright-protected content?

As outlined in our AI Principles, we are committed to respecting intellectual property rights. It's important to acquire content responsibly and lawfully, such as by giving websites the ability to opt out of having their sites used for AI training. Existing industry standards governing web crawling are an important way to accomplish this. These standards are simple and scalable, and build on long-established machine-readable robot.txt protocols widely used across the web to control how content is accessed by web crawlers. And now thousands of web publishers are using the Google-Extended protocol and similar AI-specific protocols offered by other companies.

### f. How does your organization protect privacy? How does your organization guard against systems divulging confidential or sensitive data?

Our AI Principles include our commitment to promoting privacy and security. We apply lessons from our longstanding approach to privacy for our users. Our safety policies align with Google's standard framework for the types of harmful content that we do not permit our frontier models to generate. These are designed to help prevent frontier models from generating harmful content, including revealing personal identifiable information that can lead to harm (e.g., Social Security Numbers). In addition, our Secure AI Framework (SAIF) focuses on the security and privacy risks and dimensions of AI.

### g. How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security?

- **i. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems?**
- **ii. How does your organization protect against security risks the most valuable IP and trade secrets, for example by limiting access to proprietary and unreleased model weights? What measures are in place to ensure the storage of and work with model weights, algorithms, servers, datasets, or other relevant elements are managed in an appropriately secure environment, with limited access controls in place?**
- **iii. What is your organization's vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?**
- **iv. How often are security measures reviewed?**
- **v. Does your organization have an insider threat detection program?**

We use the Secure AI Framework (SAIF) to mitigate known and novel AI security risks in our frontier AI models and applications. This 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.

- **g.i.** We embrace an open and collaborative approach to cybersecurity – working with partners to combine cyber threat intelligence, contribute to AI research, and provide developer tools such as through our AI Cyber Defense Initiative.
- **g.ii.** Our frontier models are developed, trained, and stored within Google's infrastructure, supported by central security teams and by a security, safety and reliability organisation consisting of engineers and researchers with world-class expertise. We were the first to introduce zero-trust architecture and software security best practices such as fuzzing at scale, and we have built global processes, controls, and systems to ensure that all development (including AI/ML) has the strongest security and privacy guarantees. As mentioned in earlier responses, our Detection & Response team provides 24/7/365 monitoring of Google products, services and infrastructure – with a dedicated team for insider threat and abuse. We also have several red teams that conduct assessments of our products, services, and infrastructure for safety, security, and privacy failures.
- **g.iii**. As mentioned in previous responses, we maintain a Vulnerability Reward Program.
- **g.iv.** Members of our security teams review security plans for our networks and services and provide project-specific consulting services to our product and engineering teams. They monitor for suspicious activity on our networks and address information security threats as needed. The teams also perform routine security evaluations and audits, which can involve engaging outside experts to conduct regular security assessments.
- **g.v.** Yes, we have an incident response process for incidents at Google, including pertaining to artificial intelligence and a dedicated team for insider threat and abuse. As mentioned in an

earlier response, we updated our vulnerability rewards program to specifically clarify and encourage reporting of issues in our AI products.

### h. How does your organization address vulnerabilities, incidents, emerging risks?

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.

We have developed forms of automated red-teaming to understand how vulnerable our systems may be to indirect prompt injection attacks, and to inform how we mitigate the risk.

We released Version 2.0 of our Frontier Safety Framework, a set of protocols to help us stay ahead of possible severe risks from powerful frontier AI models. The Framework includes a section that describes our mitigation approach for models that pose risks of severe harm through misuse, and then details our set of misuse Critical Capability Levels, as well as the mitigations that we provisionally assess as appropriate for them. Since Version 1.0, we've collaborated with experts in industry, academia, and government to deepen our understanding of risks, empirical evaluations to test for them, and mitigations we can apply.

We also recently released the Google Deepmind AGI Safety paper, which details how we’re taking a systematic and comprehensive approach to AGI safety, exploring four main risk areas: misuse, misalignment, accidents, and structural risks.

# Section 3 - Transparency reporting on advanced AI systems

a. Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?

- i. How often are such reports usually updated?
- ii. How are new significant releases reflected in such reports?
- iii. Which of the following information is included in your organization's publicly available documentation: details and results of the evaluations conducted for potential safety, security, and societal risks including risks to the enjoyment of human rights; assessments of the model's or system's effects and risks to safety and society (such as those related to harmful bias, discrimination, threats to protection of privacy or personal data, fairness); results of red-teaming or other testing conducted to evaluate the model's/system's fitness for moving beyond the development stage; capacities of a model/system and significant limitations in performance with implications for appropriate use domains; other technical documentation and instructions for use if relevant.

Yes, we have published Responsible AI Progress reports every year since 2019. Our latest Responsible AI Progress Report details how we govern, map, measure, and manage AI risk throughout the AI development lifecycle.

We also release model cards, which are intended to provide essential information on models, including known limitations, mitigation approaches, and safety performance. Additional 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.

- **a.i.** – We release our Responsible AI Progress Report annually. External model cards and technical reports (such as for Gemini 1.5) are published regularly.
- **a.ii** – Our annual Responsible AI Progress Reports describe progress we've made on risk mitigation techniques across different frontier model and product releases, including safety tuning, security and privacy controls, the use of provenance technology in our products, and broad AI literacy education.
- **a.iii** – Details and results of the evaluations conducted for potential safety, security, and societal risks including risks to the enjoyment of human rights, results of red-teaming, and other testing conducted to evaluate frontier models' fitness for moving beyond the development stage can be found in technical reports and model cards.

### b. How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?

We are 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. As mentioned in our response to 1.d, we signed a first-of-its-kind agreement with fellow members of the Frontier Model Forum (FMF), designed to facilitate information-sharing about threats, vulnerabilities, and capability advances unique to frontier AI. We regularly participate in working groups to share information with multi-stakeholder organizations such as Partnership on AI (PAI), and contribute to external whitepapers such as PAI's Documenting the Impact of Foundation Models.

We helped launch The Adversarial Nibbler Challenge, a data-centric AI competition to engage the research community in jointly identifying current blind spots in harmful image production. This competition is the result of collaboration between six different organizations to jointly produce a shared resource for use and reuse by the wider research and development community.

### c. Does your organization disclose privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

We clearly disclose our Privacy Policy and Terms of Service. We have established product guidelines for advanced AI outputs. Our Generative AI Prohibited Use Policy makes clear restrictions that apply to users' interactions with generative AI in the Google products and services that refer to this policy.

### d. Does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

Yes, information on training data and methods can be found in our model cards and technical reports.

### e. Does your organization demonstrate transparency related to advanced AI systems through any other methods?

Our deepmind.google and ai.google websites offer additional transparency on how we're advancing frontier models and products safely and responsibly.

Our Explainability resources provide key information and guidance on how to provide transparency to users about how and when they are interacting with AI systems, including an Explainability Rubric which contains guidance on the key information that can be used to explain AI technology to users. We created this rubric based on real and fictional examples of AI systems, and tested it with AI practitioners.

Our Responsible AI Toolkit offers transparency into best practices. The toolkit provides developers with new and enhanced tools for evaluating model safety and filtering harmful content and presents product principles that cover end-to-end safety and practical guidance to help user experience, product, engineering, and AI teams to build responsible generative AI products.

# Section 4 - Organizational governance, incident management and transparency

### a. How has AI risk management been embedded in your organization governance framework? When and under what circumstances are policies updated?

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.

Our governance process is grounded in our principles and frameworks, including:

- Our AI Principles, which we established and evolved to guide our approach to developing and deploying AI models and applications, focus on pursuing AI efforts where the likely overall benefits substantially outweigh the foreseeable risks.
- The Frontier Safety Framework, which we recently updated, helps us to proactively prepare for potential risks posed by more powerful frontier AI models. The Framework follows the emerging approach of Responsible Capability Scaling proposed by the U.K.'s AI Safety Institute.
- 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.
- Our Secure AI Framework focuses on the security and privacy dimensions of AI.
- Additional policies address design, safety, and prohibited uses.

We operationalize our principles, frameworks, and policies through a system of launch requirements, leadership reviews, and post-launch requirements designed to support continuous improvement. For example, Google DeepMind Responsibility and Safety Council (RSC), Google DeepMind's governance body, reviews the initial ethics and safety assessments on novel model capabilities in order to provide feedback and guidance during model development. The RSC also reviews data on the model's performance via assurance evaluations and RSC decisions inform releases.

### b. Are relevant staff trained on your organization's governance policies and risk management practices? If so, how?

Our teams receive training on general and machine-learning-specific safety and security practices, as well as on our AI Principles and Responsible AI Practices. We have a set of internal policies and guidelines establishing best practices for safety, security, and privacy across the development and deployment of responsible AI, such as those mentioned in our previous answers.

### c. Does your organization communicate its risk management policies and practices with users and/or the public? If so, how?

Risk management policies and practices can be found on our Transparency Center, in our Responsible AI Progress Report and in posts on The Keyword, ai.google, and publicpolicy.withgoogle.com.

The SAIF Risk Assessment is an interactive tool for AI developers and organizations to take stock of their security posture, assess risks and implement stronger security practices. We have published a full list of AI security risks and controls.

### d. Are steps taken to address reported incidents documented and maintained internally? If so, how?

Yes, 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 governance processes. 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. We also invest in tooling and central repositories for model and data lineage to promote transparency and accountability.

### e. How does your organization share relevant information about vulnerabilities, incidents, emerging risks, and misuse with others?

As mentioned in previous answers, we signed a first-of-its-kind agreement with fellow members of the The Frontier Model Forum (FMF), designed to facilitate information-sharing about threats, vulnerabilities, and capability advances unique to frontier AI.

And we participate in additional information sharing with stakeholders, as appropriate, including the Stanford Foundation Model Transparency Index where we submitted scoring for the Gemini AI model used for the Cloud API.

### f. Does your organization share information, as appropriate, with relevant other stakeholders regarding advanced AI system incidents? If so, how? Does your organization share and report incident-related information publicly?

Yes, we are committed to addressing AI-related risks so that we can maximize AI's benefits for people and society. We have shared public information on incidents and mitigations via Google Keyword blog, such as this example. We have also published information regarding incidents on Google Cloud Service Health.

### g. How does your organization share research and best practices on addressing or managing risk?

We have published more than 300 papers on responsible AI topics, many of which are relevant to addressing or managing risk. (These are available to the public on Google Research's publications repository and Google DeepMind's publications archive.) And we collaborate with research institutions around the world. Papers focused on best practices for addressing or managing risks include publications on holistic safety evaluations, a study of misuses of generative AI, ethical concerns regarding advanced AI systems, and a context-based framework for comprehensively evaluating the social and ethical risks of AI systems.

In addition to the papers published on our research sites, we share updates for a broader audience on sites such as publicpolicy.google or the Keyword blog. More information about our comprehensive AI risk management processes and research in the space can be found in the Responsible AI Progress Report.

We also share best practices for managing risk via the Responsible AI Toolkit (mentioned in various responses), which features tools such as ShieldGemma, a series of state-of-the-art safety classifiers that application 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.

### h. Does your organization use international technical standards or best practices for AI risk management and governance policies?

Yes. Google Cloud, Gemini App, and Google Workspace are ISO/IEC 42001-2003 certified. ISO/IEC 42001:2023 outlines and provides the requirements for an Artificial Intelligence Management System (AIMS), specifies a set of best practices, and details AI management controls that can help manage AI risks.

As mentioned in various responses above, we participate in working groups on industry standards for AI risk management and AI governance policies within global multi-stakeholder organizations including MLCommons, the World Economic Forum's AI Governance Alliance, the Coalition for Content Provenance and Authenticity (C2PA), Thorn, Partnership on AI, Frontier Model Forum, and government partners. 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. And we follow the NIST AI Risk Management Framework, as documented in the most recent Responsible AI Progress Report.

# Section 5 - Content authentication & provenance mechanisms

### a. What mechanisms, if any, does your organization put in place to allow users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?

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

Mechanisms we employ in this space include interstitials and FAQs, such as the FAQ for Gemini Apps, that provide users with the knowledge and answers that they are engaging with an advanced system.

Our Explainability Rubric provides guidance to help users understand how AI operates and makes decisions. Where appropriate, our generative-AI 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.

We also published a case study with Partnership on AI regarding how we offer disclosures to users on AI-generated content.

### b. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems? If yes, how? Does your organization use international technical standards or best practices when developing or implementing content provenance?

We improved SynthID, our industry-leading image and audio watermarking tool, to also be able to identify text and video generated by the Gemini app and web experience. 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. Our solution, “About this image”, gives users multiple ways to quickly get context on images, such as how other sites use and describe the image, image metadata, and digital watermarking provided it contains SynthID embedded within its pixels. We’ve also open sourced SynthID-Text 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.

We joined the C2PA (Coalition for Content Provenance and Authenticity) as a steering committee member. The coalition is a cross-industry effort to provide more transparency and context for people on digital content. Google is helping to develop its technical standard and further adoption of Content Credentials, tamper-evident metadata which shows how content was made and edited over time. Google Search, Ads, and YouTube are 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.

## Section 6 - Research & investment to advance AI safety & mitigate societal risks

### a. How does your organization advance research and investment related to the following: security, safety, bias and disinformation, fairness, explainability and interpretability, transparency, robustness, and/or trustworthiness of advanced AI systems?

We continuously invest in responsible AI, and have published more than 300 papers on responsible AI topics (available on Google's Research publications repository and Google DeepMind's publications archive, as mentioned in various answers above), 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 cybersecurity, dangerous capabilities, misinformation, adversarial use of generative AI, socio-technical approaches to red-teaming, health equity and bias, 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.

We enable the ecosystem through research funding such as with our Frontier Model Forum partners, with whom we co-founded the AI Safety Fund (AISF), mentioned in answer 1.h, which provides grants to researchers to help identify, evaluate, and mitigate risks and improve the safe deployment of AI for the benefit of society. The goal of the fund is to allow researchers to better evaluate and understand frontier systems, and – ultimately – develop new model evaluations and techniques for red- teaming AI models, to help develop and test evaluation techniques for potentially dangerous capabilities of frontier systems.

### b. How does your organization collaborate on and invest in research to advance the state of content authentication and provenance?

As outlined in our response to question 5.b, we joined the Coalition for Content Provenance and Authenticity (C2PA) as a steering committee member. The coalition is a cross-industry effort to provide more transparency and context for people on digital content. Google is helping to develop this technical standard and promote wider adoption of Content Credentials, tamper-evident metadata that shows how content was made and edited over time.

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.

### c. Does your organization participate in projects, collaborations, and investments in research that support the advancement of AI safety, security, and trustworthiness, as well as risk evaluation and mitigation tools?

We actively participate in, collaborate on, and invest in projects advancing AI safety, security, and trustworthiness, including the development of risk evaluation and mitigation tools. We pursue collaborative research into different types of impacts that AI may have, such as impacts on children and vulnerable groups, misinformation, equity, human rights, and democratic values.

We support external research to increase safe and responsible AI development across the ecosystem. In October 2024, we announced the winners of Google's 2024 Academic Research Awards (GARA) program. These research projects span a range of topics including:

- Creating ML benchmarks for climate problems
- Making education equitable, accessible and effective using AI
- Using Gemini and Google's open model family to solve systems and infrastructure problems
- Society-centered AI
- Trust & Safety, including using AI to improve digital safety across online ecosystems

The GARA program supports groundbreaking research in computing and technology, addressing global challenges such as climate change, education, quantum computing, the societal impact of AI, digital safety and infrastructure optimization.

### d. What research or investment is your organization pursuing to minimize socio-economic and/or environmental risks from AI?

Research reports and updates are posted to research.google under Science, AI & Society with focuses on areas such as, but not limited to, Climate & Sustainability and Health & Bioscience. A recent example is research on environmental risks and TPU emissions efficiencies gained over generations, in Life-Cycle Emissions of AI Hardware: A Cradle-To-Grave Approach and Generational Trends.

## Section 7 - Advancing human and global interests

### a. What research or investment is your organization pursuing to maximize socio-economic and environmental benefits from AI? Please provide examples.

As per our AI Principles, we are committed to 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. We invest in research covering topics on socio-economic and environmental benefits from AI. For example, 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. AlphaFold was awarded the Nobel Prize for Chemistry in 2024. Scientists can access the majority of Alphafold 3's capabilities, for free, through our AlphaFold Server, an easy-to-use research tool, or via open code and weights.

In addition, we publish papers and host initiatives on a variety of topics on beneficial AI, including:

- Environmental benefits from Google Sustainability report
- AI for Science Research, such as:
- Weather Forecasting
- Fusion
- Materials Science

### b. Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities, limitations and impacts of advanced AI systems? Please provide examples.

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 and have launched AI training for businesses, developers, and younger learners. Additional initiatives include:

- Digital literacy and training to improve user awareness, via educational programs such as Google AI Essentials and the Applied Digital Skills Program.
- Google.org has made significant investments in AI literacy and education, including through the Experience AI program with Raspberry PI; University initiatives across Europe, the Middle East and Africa; and new education tools like a streamlined class management system for teachers.
- Various forms of digital literacy and educational materials, for example through scholarships to community resources offered by Google Deepmind.

### c. Does your organization prioritize AI projects for responsible stewardship of trustworthy and human-centric AI in support of the UN Sustainable Development Goals? Please provide examples.

Yes. As governments, the private sector, and civil society work to achieve the UN Sustainable Development Goals (SDGs) by their 2030 target date, our CEO has publicly discussed AI opportunities aligned with the SDGs, including:

- Helping people access the world's information and knowledge in their own language. Using AI, in just the last year, we have added new languages to Google Translate and we're working toward 1,000 of the world's most spoken languages.
- Accelerating scientific discovery to benefit humanity. We opened up AlphaFold AI technology (also mentioned in answer 7.a), which predicts some of the building blocks of life, including proteins and DNA, to the scientific community free of charge, and it's been accessed by more than two million researchers from over one hundred and ninety countries, with thirty percent in the developing world. Globally, AlphaFold is being used in research that could help make crops more resistant to disease, discover new drugs in areas like malaria vaccines and cancer treatments, and much more.
- Helping people in the path of climate-related disaster, building on the UN's initiative for "Early Warnings for All." Our Flood Hub system provides early warnings up to seven days in advance, helping protect over 460 million people in over 80 countries. And for millions in the paths of wildfires, our boundary tracking systems are available in multiple countries on Google Maps. We also launched FireSat technology, which will use satellites to detect and track early-stage wildfires, with imagery updated every 20 minutes globally, so firefighters can respond. AI gives a boost in accuracy, speed and scale.
- Meaningful contributions to economic progress by enabling entrepreneurs and small businesses, empowering governments to provide public services and boosting productivity across sectors.

### d. Does your organization collaborate with civil society and community groups to identify and develop AI solutions in support of the UN Sustainable Development Goals and to address the world's greatest challenges? Please provide examples.

Yes. Our collaborations to support the UN Sustainable Development Goals include:

- Google Developer Student Clubs (GDSC): Since 2020, GDSC members from around the world come together to create innovative solutions to tackle some of the world's most pressing challenges.
- AI in Action: Accelerating Progress Towards the Sustainable Development Goals: This research highlights AI capabilities (including computer vision, generative AI, natural language processing, and multimodal AI) and showcases how AI is altering how we approach problem-solving across all 17 SDGs through specific use cases, with a spotlight on AI-powered innovation in health, education, and climate.
- AI for the Global Goals: Building on 5 years of support for AI-powered initiatives, we launched a $25M open call for organizations using AI to accelerate progress on the Sustainable Development Goals.
- Google Research's AI for Social Good awards, in partnership with Google.org, aim to help academics and nonprofits develop machine learning-based technologies that can improve people's lives. Through this program, we aim to bring nonprofits and academics together to collaborate on projects that tackle social, humanitarian and environmental challenges.