

OECD AI Transparency Report

Organization: Google (USA)

Reporting Period: Q2 2026

Published: October 2, 2026

Section 0 – Your role in the AI system lifecycle

Responding as (please select all that apply):

- Model developer (/provider)
- Application developer (/provider)
- Deployer

If any of the fields above are selected, please indicate your organization size:

Large enterprise: 250 or more employees

Section 1 – Risk Identification and Evaluation

1. How does your organization define and/or classify different types of risks related to AI?¹

- Maintain an internal taxonomy, classification system, and/or set of risk area definitions
- Define enhanced safety measures for specific dangerous capability thresholds, that, once reached, require safeguards (e.g. capability threshold for dangerous Chemical, Biological, Radiological and Nuclear (CBRN), cybersecurity attacks, dangerous advanced autonomy capabilities, or other capabilities).
- Align with specific external frameworks, standards, reports, and/or guidance documents such as taxonomies, classification systems, and/or sets of definitions for risk areas (e.g. the NIST AI Risk Management Framework, the ISO/IEC 42001, the OECD AI Principles, the EU General-Purpose AI Code of Practice, the International AI Safety Report). (please specify): NIST AI Risk Management Framework, the ISO/IEC 42001, The EU General-Purpose AI Code of Practice
- Consult with external stakeholders (e.g., from academia, industry groups, civil society, or standard development organisations) when developing or updating risk taxonomies

How does your organization define thresholds at which severe risks, unless adequately mitigated, may be considered "unreasonable"? Please describe the framework(s) most relevant to your organization's role and context. If your organisation does not use formal thresholds, you may describe instead the policies, decision criteria, or governance processes you apply when assessing severe AI risks. You may also describe:

- Specific examples of contexts in which your models or systems could, in your assessment, pose unreasonable or intolerable risk;
- How mitigations are graduated by capability or propensity level;
- How your organisation considers the specificities of different AI technologies, such as generative AI and agentic AI, including any specific risk assessment or governance mechanisms applied to agentic AI capabilities.

Our [Frontier Safety Framework](#) (FSF) is a set of protocols that aims to address severe risks that may arise from the high-impact capabilities of frontier AI models. It complements Google's suite of AI responsibility and safety practices, and enables Google's AI innovation and deployment consistent with our [AI Principles](#).

The FSF is built primarily around frontier model capability thresholds called "Critical Capability Levels (CCLs)." These are capability levels at which, absent mitigation measures, frontier AI models may pose heightened risk of severe harm. We identify CCLs for two kinds of risks: misuse risk and risks related to machine learning R&D and misalignment. For each type of risk, we define a set of CCLs and a mitigation approach for them. Where relevant, we also define Tracked Capability Levels (TCLs). TCLs are meant to capture significant risks that may manifest at a lower capability threshold than our CCLs and we apply our mitigation and risk acceptance determination process proportionately.

Specificities of Generative and Agentic AI Systems

Because we understand that advanced and novel AI systems require evolved testing paradigms, we implement specific governance frameworks, and forward-looking policy frameworks and recommendations, to address the unique capabilities of agentic and generative AI.

- Novel AI Testing: We formed a dedicated Novel AI Testing team to spearhead evaluations at scale for new AI systems—such as advanced agents and Personal Intelligence.
- The TRUST Framework: To govern the capabilities of agentic AI, we introduced the TRUST framework, which is built on five core pillars: Transparency, Reliability, Unified protocols and standards, Safety and security, and Task augmentation, to guide development in this space.
- Security Frameworks: 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.
- Regulatory Sandboxes: For high-stakes agent deployments, we recommend that policymakers leverage tools like regulatory sandboxes and pilot programs to collect real-world data on failure modes.

2. Both before and after deployment, what practices does your organization use to identify and evaluate vulnerabilities, incidents, emerging risks and misuse throughout the AI system lifecycle?

- Internal red teaming³
- Independent third-party red teaming
- Adversarial testing to detect technical vulnerabilities
- Publicly available benchmarks and measurement tools
- Internal benchmarks, standard metrics (e.g., frequency, severity scores)
- Structured techniques to explore the behavior or performance of a model and/or application, such as model elicitation or bottleneck assessments
- Qualitative assessments (e.g., expert judgment, narrative scoring)
- Specific risk identification practices for vulnerable groups
- Impact assessments such as personal data protection assessment and fundamental rights impact assessments
- Internal model, system, or application-independent research initiatives

- External model, system, or application-independent research initiatives
- Vulnerability reports
- Incident reports, including reports shared by other organizations
- Automated monitoring / remote status reporting (e.g., telemetry pipelines, automated alerts when risk thresholds are breached)
- Specific risk identification practices for agentic AI capabilities, e.g., autonomous action-taking, tool use, multi-step planning (please specify): Scaled evaluations by the Novel AI Testing team
- Assessment of risks arising from interactions among multiple AI agents
- Multilingual and cross-cultural evaluation, including collaboration with local ecosystems (e.g., national AI institutes, language-specific evaluation consortia) for under-represented languages and cultural contexts (please specify): Safety and fairness evaluations in multiple languages

Are any national AI safety / security institutes of the AISI network involved?

No answer provided

How does your organization monitor the interactions of AI agents under your control with other agents / assess risk from multiple agents interacting?

No answer provided

Please briefly describe your organization's decision-making process for deploying an AI model or system including: who has decision authority; what evidence and criteria are required (e.g., performance, safety/security, robustness, fairness, privacy, compliance, independent external testing, evaluation, validation and verification (TEVV) or assurance); contextual caveats or limitations required in interpreting risk evaluation metric results; what approvals/sign-offs are needed; and what safeguards (e.g., mitigations, monitoring, or phased release) must be in place to move into deployment.

Our governance process is grounded in a multi-layered approach to AI governance that ensures robust oversight throughout the AI lifecycle, including launch reviews for both models and applications developed using these models.

For model launches, we have in place a well-established and comprehensive internal governance structure designed to ensure the robust implementation of the processes outlined in the [Frontier Safety Framework](#). Responsibilities for assessing and mitigating risks are clearly defined and allocated across all levels of the organization. This includes legal, compliance, and safety reviews with escalation procedures to ensure appropriate oversight.

Before launching a model or system, 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](#).

The ultimate decision authority for deploying AI models or systems rests with specific governance forums and councils within the organization.

- Model launches are reviewed at Google DeepMind's Launch Review forum, which approves model releases.
- Application launches are reviewed systematically via launch infrastructure and centralized expert risk reviews, alongside application-focused launch review forums.

Following deployment, how do you use incident reports — including those shared by other organizations — to identify emerging issues, update risk assessments, and trigger corrective actions.

Incident detection and response are key aspects of our overall safety, security, and privacy program. We maintain rigorous processes for managing incidents involving frontier AI models and systems . These processes specify actions, escalations, mitigation, resolution, and notification of incidents.

Identifying Emerging Issues We use numerous internal and external signals including incident reports to identify new risks through continuous post-launch monitoring that leverages both automated systems and human review. We strive to learn from our post-market monitoring mechanisms, including as part of our detection, mitigation, response and/or reporting of incidents related to our AI risk domains. Actionable insights from these processes allow us to enhance our tools, training, processes, policies, and response efforts.

Updating Risk Assessments Information gathered from incident management and crisis response is directly fed back into our risk identification efforts. Our residual risk assessments, safety cases and mitigations may be updated as a result of post-market monitoring, including information about incidents and near misses relating to our frontier safety risk domains. Material updates to a safety case will be submitted to the appropriate governance function for review and might result in updates to the related residual risk assessment or safety case.

Triggering Corrective Action If an incident is detected, we utilize mature incident management and crisis response capabilities to rapidly mitigate and remediate where needed. We use a variety of means for taking corrective action tailored to the particular facts of an incident. For example, teams are explicitly enabled to have rapid-remediation mechanisms in place to immediately block content flagged as illegal.

3. Does your organization have mechanisms to receive and/or share reports of risks, incidents or vulnerabilities by/with third parties?

Yes, both receiving and sharing

If yes, please indicate which stakeholders reporting mechanisms may be accessible to, recognising specifics will depend on the nature of the report and risk level. For each, indicate whether access is on a standing basis (Always), activated based on context (Ad hoc/as relevant), or not currently accessible.

	Always	Ad hoc / as relevant	Not currently accessible
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities (e.g., public reporting offices for security incidents or data breaches)	☐		
Supervisory authorities (e.g., data protection authorities)	☐		
Industry partners or industry fora	☐		

	Always	Ad hoc / as relevant	Not currently accessible
Users (end users, enterprise customers, developers using APIs)	☐		
Business customers / deploying organizations (if different from “users”)	☐		
Supply-chain partners (cloud, data providers, etc.)			
Academia / researchers	☐		
Independent auditors	☐		
Standards bodies / multi-stakeholder initiatives (OECD.AI, ISO/IEC, IEEE, etc.)			
Specific civil society organizations	☐		
Media / general public (people or communities indirectly impacted)	☐		
Other			

Please describe how your organisation determines when and how reporting mechanisms are made accessible to different stakeholders (e.g. based on legal or regulatory requirements, internal policies, or risk level). Please also provide public URLs or details on how to submit reports. If none, write “Not applicable”

We make reporting mechanisms accessible on a standing basis to the public, researchers, and users via dedicated intake channels. To evaluate specific risks, we utilize dedicated intake channels tailored to the ecosystem, such as:

- User Reports and Feedback in Product: We provide specific reporting mechanisms tailored to how users are interacting with our products. For example, users can utilize in-product feedback mechanisms to directly flag unsafe content or file bugs. Details on how to report feedback can be found at <https://support.google.com/gemini/answer/13275746>
- Legal Troubleshooter: To report copyright or legal issues, users can submit concerns through the Google Legal Help Content Troubleshooter at support.google.com/legal/troubleshooter/1114905.
- Likeness Complaints: To report content where users believe their likeness has been generated or altered by someone else with Google's generative AI tools without consent, users can submit a formal complaint at <https://reportcontent.google.com/forms/ai-likeness-abuse>
- Commercial Users: Cloud Customers who believe they are experiencing a security incident, such as a breach, ransomware attack, or other compromise, and need immediate assistance can leverage Incident Response services at <https://cloud.google.com/security/report-incident> and https://support.google.com/code/contact/cloud_platform_report.
- Vulnerability Reward Program: The Vulnerability Reward Program (VRP) is a bug bounty initiative that compensates external security researchers for identifying and reporting security flaws in

Google's web properties, devices, and open-source code before they can be exploited by malicious actors. More info at <https://bughunters.google.com/>

How does your organization share relevant information as appropriate?

- Publicly
- Information-sharing partnerships (with confidentiality agreements as appropriate)
- Reporting to regulatory authorities as relevant

Please describe what type of information is shared.

We may share public information on incidents and mitigations via Google's public blogs. We may also consider disclosing information to other external organizations to promote shared learning and coordinated risk mitigation. If we assess that a model has reached a CCL that poses an unmitigated and material risk to overall public safety, we aim to share relevant information with appropriate government authorities where it will facilitate the safety of frontier AI. Where appropriate, and subject to adequate confidentiality and security measures and considerations around proprietary and sensitive information, this information may include:

- Model information: characteristics of the AI model relevant to the risk it may pose with its critical capabilities.
- Evaluation results: such as details about the evaluation design, the results, and any robustness tests.
- Mitigation plans: descriptions of our mitigation plans and how they are expected to reduce the risk.

We will continue to review and evolve our disclosure process over time.

4. If your organization has programs to incentivize risk(s) disclosure, incidents and systemic vulnerabilities, please share what type of programs you use?

- Paid rewards ("bug bounty")
- Vulnerability Disclosure Programs (VDPs) with safe harbor / liability protection
- Recognition and reputation incentives
- Early access and collaboration incentives
- Research grants, contracts, and hiring red teamers
- Internal incentives and culture

Please provide public URLs and/or details on your incentive program(s). If none, write "Not applicable."

Program Details: The dedicated AI VRP structures monetary incentives based on the specific category and severity of the reported vulnerability. The program offers rewards up to \$20,000 for "Rogue Actions" and up to \$15,000 for "Sensitive Data Exfiltration," up to \$5,000 for "Model Theft," "Context Manipulation," and "Phishing Enablement" in addition to rewards for other categories, when identified on flagship products.

Public URL: Full operational rules, scope guidelines, and program announcements can be accessed at: <https://bughunters.google.com/blog/6116887259840512/announcing-google-s-new-ai-vulnerability-reward-program>.

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

Contributes and uses

If contributes or uses, please select all that apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001, ISO/IEC 23894, ISO/IEC 42005)
- NIST AI Risk Management Framework (AI RMF)
- Industry-led best practices or codes of conduct
- General Purpose AI Code of Practice
- Other international or regional standards (please specify): Coalition for Content Provenance and Authenticity (C2PA) standard, OASIS Open standards, Appia Foundation

6. Does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks that could have broad and potentially cascading socio-economic impacts? Please indicate which stakeholders you collaborate with, and whether collaboration is on a standing basis (Always-on), activated based on context (Ad hoc/as relevant), or there is no current collaboration.

	Always-on	Ad hoc / as relevant	No current collaboration
Relevant employees (security, legal/compliance, etc.)	☐		
Public authorities / regulators / AI safety institutes / other public sector bodies and inter-governmental coordination structures	☐		
Industry partners or industry fora	☐		
Users (end users, enterprise customers, developers using APIs)	☐		
Business customers / deploying organizations (if different from “users”)	☐		
Supply-chain partners (cloud, data providers, etc.)			
Academia / researchers	☐		
Independent auditors	☐		
Standards bodies / multi-stakeholder initiatives (OECD.AI; ISO/IEC; IEEE; AIID etc.)	☐		
Independent civil society organizations (e.g., for evaluations, audits, or policy review)	☐		

	Always-on	Ad hoc / as relevant	No current collaboration
Media / general public (people or communities indirectly impacted)	☐		
Other(s) (please add any other relevant stakeholder groups):			
Not applicable			

Please describe examples of how your organization collaborates. Examples may include: participating in multi-stakeholder or industry initiatives; partnering with academia, civil society, or the public sector; establishing internal or joint governance structures with external experts or in partnership with other efforts to share knowledge and jointly address risks.

N/A

Any further comments and implementation documentation.

Due to the diverse nature of our product portfolio and the varying roles we play across the AI lifecycle, the selections marked and practices shared in this report may not apply uniformly to all products or deployment environments. Our responses reflect practices that are implemented where contextually appropriate and technically feasible.

Section 2 – Risk Management and Information Security

7. Across the AI system lifecycle, what steps does your organization take to manage risks and vulnerabilities? Please indicate which specific practices, tools or metrics your organization uses at each lifecycle stage. You may refer to tools or metrics present in the OECD.AI Catalogue of Tools and Metrics or suggest additions. Respondents may complete only the lifecycle stages relevant to their organisation's role. Stages that do not apply may be left blank.

Plan and design

Collect and process data

Build, integrate and interpret model

Verify and validate

Deploy

Operate and monitor

Please provide a consolidated description of your organisation's approach to risk management across the AI system lifecycle, including how the practices identified above operate together and any cross-cutting governance, review or escalation processes. Please also indicate how often these practices are

reviewed and what types of events trigger reviews and updates. Where relevant, please distinguish between different types of AI systems (e.g. generative or agentic AI).

Integration Across the AI System Lifecycle

Our risk management practices operate together in a continuous feedback loop:

- We take a research-driven, scientific approach to identify current and emerging risks across new modalities, codifying these inputs into a risk taxonomy.
- We proactively address risks through rigorous guidelines—such as content safety policies, the Prohibited Use Policy, the [Secure AI Framework](#), and the [Frontier Safety Framework](#)—intended to prevent harmful outputs and guide model behavior.
- We stress-test systems against these frameworks using scaled evaluations alongside internal and external red-teaming.
- We mitigate identified risks through supervised fine-tuning, reinforcement learning, and out-of-model mitigations like safety filters and conditional system instructions.
- Post-launch, we use automated systems and human reviews to monitor for misuse.
- 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.

Cross-Cutting Governance, Review, and Escalation Processes Our governance process is grounded in a system of cross-cutting governance bodies and review forums to manage escalations and approve launches.

- Our multi-layered process includes launch reviews for both frontier models and applications developed using these models.
- Model releases are specifically approved by Google DeepMind's Launch Review forum, while application launches are reviewed via centralized expert risk reviews and application-focused forums.
- For long-term and structural risks, our AGI Futures Council—consisting of senior management and Alphabet's Board of Directors—provides perspectives and recommendations to our Board and management team on long-term opportunities, risks, and impacts associated with the development of AGI.

Frequency of Reviews and Trigger Events

Our risk management practices are fundamentally iterative.

- Our approach evolves with developments in the underlying technology, new and emerging risks, and as new measurement techniques emerge, such as AI-assisted evaluations.
- To assess whether a model contributes to identified safety risks, we conduct a critical capability assessment against Threshold/Critical Capability Levels (T/CCLs) prior to its initial external deployment. For subsequent model iterations, we execute material capability change assessments during post-training—evaluating specific checkpoints based on provenance, size, modality, and benchmark performance—to identify meaningful capability increases that could alter our risk determination. This monitoring relies on "early warning evaluations" and defined "alert thresholds"—mechanisms designed to flag when a Critical Capability Level (CCL) may be reached before a full critical capability assessment is repeated—which can be dynamically adjusted or run more frequently if the observed rate of progress diminishes our established safety buffers.

Our [Frontier Safety Framework](#) will be reviewed at least once a year—more frequently if we have reasonable grounds to believe the adequacy of the Framework or our adherence to it has been materially undermined. The process will involve (i) an assessment of the Framework’s appropriateness for the management of significant and severe risk, drawing on information sources such as record of adherence to the framework, relevant high-quality research, information shared through industry forums, and evaluation results, as necessary, and (ii) an assessment of our adherence to the Framework.

Distinctions Between Types of AI Systems

We proactively assess and tailor our risk management approach based on the specific type and capability of the AI system:

- Frontier AI Models: Governed by the [Frontier Safety Framework](#), which is built around capability thresholds called Critical Capability Levels (CCLs) intended to address severe risks like misuse and deceptive alignment.
- Agentic AI & Personal Intelligence: In this agentic era—where AI systems autonomously interact with services and users—we require a new testing paradigm; our Novel AI Testing team was formed to spearhead evaluations at scale for new AI systems, such as advanced agents and Personal Intelligence.
- Artificial General Intelligence (AGI): Managed with a focus on long-term, structural risks, exploring four main risk areas—misuse, misalignment, accidents, and structural risks—under the guidance of the AGI Futures Council and the approach detailed in Google Deepmind’s [AGI Safety paper](#).

Please describe any measures used to ensure data quality and fairness across the data collection and curation stage (e.g., dataset audits, filtering training data to remove harmful content or personal information, fine-tuning on curated data, use of moderation tools, assessment of limitations to model generalisability).

We ensure data quality and fairness across the AI lifecycle through robust dataset audits, rigorous pre-training safety filtering, and supervised fine-tuning. We deploy a combination of automated and human-led moderation tools to steer outputs away from harm, and we actively document known limitations regarding generalisability to mitigate unfair bias.

Dataset Audits, 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.

To ensure this, governance requirements for models focus on filtering training data for quality, model performance, and adherence to policies, as well as documenting training techniques in technical reports and model cards. Furthermore, we also prioritize neutral and inclusive design principles, with a goal of minimizing unfair bias.

Filtering Training Data 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. Additionally, as part of making pre-trained models safe and reliable, we used automated techniques to filter out certain personal information and other sensitive data from training sets.

Fine-Tuning on Curated Data Following initial data collection and filtering, we proactively mitigate risks through both supervised fine-tuning and reinforcement learning to ensure models are aligned with our content safety policies. For example, we used extensive fine-tuning and reinforcement learning from human feedback to align our instruction-tuned models with responsible behaviors.

Use of Moderation Tools To enforce our policies at scale, 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. 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.

Assessments of Limitations to Model Generalisability 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. To offer transparency, [external model cards](#) and technical reports are published regularly to provide essential information on models, including known limitations, mitigation approaches, and safety performance.

8. What security measures are used in testing AI systems?

- Isolated and secure test environments
- Locally hosted testing environments
- Restrictions on outbound Internet or API access, including a list of permitted domains or APIs
- Isolation of staging environments (from production / operational environments)
- Controlled environments for red-teaming and adversarial testing
- Containment measures for prompt-injection and jailbreak attempts
- Controlled security vulnerability testing
- Human-in-the-loop controls for higher-risk test cases
- Access controls, including roles-based permissions
- Disabling, restricting or simulating high-impact capabilities during testing
- Audit logging and monitoring
- Use of “AI-safe” test-data, as appropriate
- Use of synthetic, anonymized, or dummy data, as appropriate

8.A. Please indicate, if relevant, the team or role responsible for overseeing secure testing and whether this responsibility is shared across multiple teams and/or functions.

Our governance process is grounded in a multi-layered oversight structure to govern secure testing throughout the AI system lifecycle. Responsibility is distributed across specialized central functions, creative red-teaming units, and senior governance councils:

- Central Security and Reliability Teams: Our frontier models are developed, trained, tested, and stored within our own infrastructure, where we apply our [Secure AI Framework](#) (SAIF) throughout the AI lifecycle—supported by central security teams and a security, safety, and reliability organization consisting of engineers and researchers with world-class expertise.
- Frontier Model Testing & External Partnerships: For model-level frontier testing, the Google DeepMind Responsibility Team operates directly within model development. Their core mission is to proactively track emerging risks, conduct structural safety evaluations (such as red-teaming), and align future models with ethical frameworks before deployment. Additionally, we work with external

expert organizations through our External Safety Testing Program. These experts independently test and provide feedback on frontier safety domains. The team also actively collaborates with independent groups like the U.K. AI Security Institute and is a founding member of the Frontier Model Forum.

- **Application Testing:** At the application/product level, our Content Adversarial Red Team (CART) proactively identifies unknown risks in our AI systems, enabling us to mitigate before launch. 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.
- **Security-focused red teaming:** Our AI Red Team combines security and AI expertise to simulate attackers who might target AI systems; based on threat intelligence from teams like the Google Threat Intelligence Group, this unit explores how AI features can cause security issues, recommends improvements, and helps ensure that real-world attackers are detected and thwarted before they cause damage.
- **Governance Councils:** 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 senior management and our Board of Directors.

Technical Description of Security Measures in Testing

Because we understand that AI, as a still-emerging transformative technology, involves evolving complexities and risks, we pursue AI responsibly by grounding our testing processes in the systematic application of technical controls to isolate risks and rigorously validate frontier models prior to deployment.

Testing Environments and Target Isolation Testing is performed in appropriately secured environments. Our frontier models are developed, trained, tested, and stored within our infrastructure, where we can apply our [Secure AI Framework](#) (SAIF) throughout the AI lifecycle, ensuring that any external trusted testing groups also receive secure access to test models. To address the safety and scalability challenges inherent in live internet testing of agentic products, where appropriate, we either restrict internet access during evals or require human-in-the-loop approval. This allows us to proactively identify and mitigate high harm risks without exposing the public web to potential harm. Additionally, we mandate that the processing of untrusted inputs occurs within sandboxed environments. A gradual approach to deployment serves as a critical risk mitigation—starting with testing internally, then releasing to trusted testers externally, before opening up to a small portion of our user base.

Containment, Capability Restrictions, and Controls To protect frontier models against adversarial inputs, we deploy robust cybersecurity controls and repeatable mitigations. For prompt injections and jailbreaks, we apply robust filtering and processing of inputs and outputs, while thorough training, tuning, and evaluation processes help fortify the model against these attacks. For data poisoning, we implement data sanitization, secure AI systems, enable access controls, and deploy mechanisms to ensure data and model integrity.

Agentic systems require a multi-layer defense strategy. We employ a hybrid defense-in-depth strategy that combines deterministic measures with dynamic, reasoning-based defenses to achieve a layered safety and security strategy. For example, in Chrome, we use a combination of rule-based and LLM-

based approaches to limit agents' permitted actions. This mechanism vets agents' actions against the user's stated goal, rejecting misaligned steps. Ultimately, this comprehensive strategy combines human-in-the-loop expertise with AI-assisted scale to enable data-driven safety and security assessments.

Audit Logging, Monitoring, and Test-Data Curation We maintain oversight during evaluations by implementing automated monitoring agents that log interactions and assess the compliance of the agent being tested in real-time in certain contexts. Automatic testing for security vulnerabilities helps us assess code risks dynamically. To optimize test-data quality by design, we use automated techniques to filter out certain personal information and other sensitive data from training sets. Furthermore, we invest in "few shot" learning methodologies where an AI creates a synthetic testing set based on inputs from experts—significantly accelerating the testing scale when compared with the manual human creation of testing sets.

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

9.A. Safeguards that promote lawful data use and transparency in AI training:

- Maintain a copyright policy that addresses data sourcing, content provenance, safeguards, and organizational roles and responsibilities
- Publishing "model cards" or contextual information on training data and model limitations
- Apply provenance controls to guard against training data being sourced from websites or repositories that have been determined to engage in large-scale piracy
- Refrain from the unlawful circumvention of technological protection measures in data sourcing
- Maintain policies or practices to respect copyright holders' preferences through, where appropriate/applicable, mechanisms like Robots.txt "do not train" tags or other machine-readable protocols developed through voluntary, industry-led standards processes
- Conduct data cleaning, qualification, and/or tokenization activities as part of a transformative data preparation process
- Engage in AI training that comprises a non-consumptive use of any underlying data
- Consult and/or retain an intellectual property attorney to protect, enforce, and comply with, intellectual property rights.
- Other(s) (please specify): Note: With respect to "provenance" in relation to AI training data within the 9A checkbox question, we interpret this term as pertaining strictly to data lineage for training purposes. For clarity, this differs from our habitual use of the term "provenance" as referring to content authenticity tools and standards (like the SynthID and C2PA standards), such as those discussed in question 26. These tools are built to offer transparency on how a piece of content was created rather than managing copyright tracking, which the framing of the question might otherwise suggest.

9.B. Safeguards that reduce the risk of dissemination of outputs that may infringe copyright:

- Restricting outputs that raise ethical, human rights, or legal concerns—such as non-consensual intimate imagery, unauthorized deepfakes, or outputs that are substantially similar to copyrighted

works.

- Filtering results using keyword or image-based restrictions to reduce risks of misuse
- Designating clear points of contact so creators or rights holders can raise concerns about infringing outputs

10. How do you protect privacy and prevent your systems from disclosing confidential or sensitive data across the AI lifecycle?

- Make information about data privacy, practices, and controls accessible to a diverse set of stakeholders
- Use security controls, including encryption at rest and in transit
- Provide organizational customers with controls to configure and implement privacy choices (e.g., data retention, logging, access, or processing settings)
- Apply Privacy-Enhancing Technologies (PETs) where applicable (e.g. differential privacy, federated learning, synthetic data generation, homomorphic encryption)

Describe key privacy measures and data-protection standards applied.

As outlined in our [AI Principles](#), we are committed to promoting privacy and security. We employ a comprehensive approach to protect privacy and sensitive data across the AI lifecycle.

- **Data Governance and Lineage:** We utilize technical controls and governance processes to ensure data is used safely and in compliance with policies. Protecting data requires both technical controls to enable safe data use at scale, and governance processes to ensure that companies have visibility and control over how their data is used. This fundamentally requires understanding data and ensuring it has sufficient metadata, controlling the data and implementing policies to allow or disallow usage based on attributes, transforming data to enable policy-compliant usage, and maintaining human oversight.
- **Encryption and Infrastructure Security:** To prevent the disclosure of confidential or sensitive data, robust encryption standards are applied to user data both at rest and in transit. We start by building on our best-in-class security infrastructure and implement additional industry-leading safeguards to ensure that this data remains protected even as it powers new AI experiences. We've also done work to increase resistance to prompt injections and have improved protection against misuse via cyberattacks.
- **Privacy-Enhancing Technologies (PETs):** Because privacy risks are highly context-dependent, our deployment decisions are guided by a risk-based framework rather than a single, one-size-fits-all approach. Furthermore, we remain committed to advancing research in the application of Privacy-Enhancing Technologies (PETs) to the AI lifecycle, such as applying differential privacy to LLMs in the context of trade-offs such as reduction in training stability, significantly increasing batch size, and computation costs. Our PETs research facilitated the release of [VaultGemma](#)—the largest (1B-parameters) open, DP-trained language model as of the date of release. This was a significant step forward in the journey toward building AI that is both powerful and private by design.
- **User Controls and Training Limitations:** When deploying advanced AI systems, we build in strict user controls. 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. If users opt in, we use our security infrastructure to securely connect data through Personal Intelligence, ensuring it is protected.

Please note: Due to the diverse nature of our product portfolio and the varying roles we play across the AI lifecycle, the selections marked and practices shared in this report may not apply uniformly to all products or deployment environments. Our responses reflect practices that are implemented where contextually appropriate and technically feasible.

11. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems across the AI system lifecycle? Respondents may complete the checkbox subsections in 11.A–11.D or describe their overall approach under 11.E.

11.A. Governance, policy and assessments

- Conduct regular cybersecurity risk and vulnerability assessments tailored to AI models or systems
- Maintain or update security policies aligned with recognized standards (e.g., NIST CSF, ISO/IEC 27001, SOC 2, CIS Critical Security Controls), and review them regularly in response to emerging risks or regulatory guidance
- Maintain vulnerability management processes (scanning, channels for external reporting, patching, remediation)
- Implement incident response, disaster recovery, or business continuity plans, including escalation procedures for high-severity incidents
- Train staff on cybersecurity best practices and AI-specific threats, including how to recognize and respond to misuse or unintended outcomes of advanced AI systems
- Integrate cybersecurity AI risks into broader cybersecurity risk management activities
- Regularly review security measures and update them in response to emerging risks, incidents, or major releases
- Other(s) (specify): Implement a full-stack approach to AI governance from responsible model development to post-launch monitoring and remediation
- Perform threat modeling that includes AI-specific risks (e.g., data poisoning, prompt injection)

11.B. Technical and operational controls

- Apply multi-factor authentication (MFA) and role-based access controls for advanced AI systems
- Encrypt AI assets (datasets, models, configurations) in storage and in transit as appropriate
- Restrict physical access to servers, GPUs, or devices hosting AI models, systems, and/or applications
- Monitor system activity and log access to detect anomalies or misuse
- Conduct secure AI development and/or deployment practices
- Conduct security testing focused on AI-specific risks (e.g., red-teaming, adversarial testing, stress tests)
- Other(s) (please specify): Utilize digital watermarking (SynthID) and supply chain integrity tools (SLSA and Sigstore) to protect the overall integrity of AI supply chains

11.C. Protection of sensitive or high value AI assets

- Limit access to proprietary AI assets based on role or need-to-know
- Store model weights, datasets, and algorithms in secure, access-controlled environments
- Track and log access to sensitive AI resources
- Require NDAs or confidentiality agreements for staff and contractors with access

11.D. Insider risk protection detections

- Conduct pre-employment screening or role-sensitive vetting
- Monitor for unauthorized internal access or suspicious activity
- Provide staff training on cybersecurity best practices and AI-specific threats
- Implement insider-risk escalation or response procedures

11.E. Please describe the main security measures your organization applies and indicate which are specific to particularly sensitive AI assets, where relevant.

Our governance process is grounded in a comprehensive, full-lifecycle approach to AI security, heavily guided by our [Secure AI Framework](#) (SAIF) and our [AI Principles](#). Because we understand that advanced AI involves evolving complexities, we proactively integrate AI risks into our broader cybersecurity risk management, utilizing a secure-by-default methodology across the ecosystem. For particularly sensitive AI assets—such as proprietary and unreleased model weights—we implement specific, heightened measures.

- **Asset Protection Commitments:** We maintain a formal commitment to invest in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights.
- **Secure Infrastructure:** To ensure the robust protection of these high-value assets, we mandate that our frontier models are developed, trained, tested, and stored exclusively within our own secure infrastructure.
- **Continuous Monitoring:** Within this protected infrastructure, [SAIF](#) is applied throughout the AI lifecycle and supported by central security teams—including our Detection & Response team—which provides 24/7/365 monitoring of products, services, and infrastructure, featuring a dedicated team for insider threat and abuse.

12. 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?

- Defined escalation criteria and follow-on actions for identified vulnerabilities (e.g., for external reporting or information sharing)
- Recognition of external finders of confirmed risks or vulnerabilities
- Reporting to regulators / authorities when relevant
- Vulnerability disclosure program that is aligned with internationally recognized standards, e.g. ISO/IEC 30111 and 29147.
- Collaborative vulnerability sharing (e.g., with partners, industry groups)
- Scanning for and mitigation of similar risks or vulnerabilities
- Testing of patches or other mitigations
- Development of patches or other mitigations for confirmed risks or vulnerabilities

- Bug bounty or incentive programs for external reporters
- Investigation of reported or identified risks and vulnerabilities
- Internal vulnerability disclosure receipt and tracking processes

13. What post-deployment practices does your organization use to monitor and respond to vulnerabilities, incidents, emerging risks and misuse of advanced AI systems?

- Conduct post-deployment monitoring
- Perform root-cause analysis and apply lessons learned to improve future development and/or deployment practices
- Monitoring and emergency controls specific to agentic AI systems (e.g., ability to interrupt autonomous action sequences, override tool-use permissions, or revoke delegated authority)
- Other(s) (please specify): Safety case updates and collaborative industry information-sharing.

14. Do you develop or deploy agentic AI systems (systems that can autonomously plan/act or use tools)?⁴

Yes

14.A. If yes, please indicate which categories of controls your organization applies. Please select all.

- Constraints on the agent's action space (e.g., tool-use permissions, environment sandboxing, allowlists/blocklists of resources, scope limits per task)
- Human oversight or interruptibility mechanisms (approval requirements for sensitive actions, escalation paths, interruption interfaces)
- Monitoring (anomaly detection, multi-agent interaction monitoring)
- Logging of agent actions

Please provide any further comments and implementation documentation.

Post-Deployment Monitoring and Governance Our governance continues post-launch with assessments for issues that might arise across products. Post-launch governance helps us to identify risks, alongside opportunities to improve our models, applications, and our governance processes.

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 continuous monitoring ecosystem includes, where relevant, 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. Furthermore, our Detection & Response team provides 24/7/365 monitoring of Google products, services, and infrastructure—featuring a dedicated team for insider threat and abuse.

To proactively assess and identify external vulnerabilities, we also have a Vulnerability Reward Program to specifically encourage the reporting of issues in our AI products.

Root-Cause Analysis and Incident Remediation We employ mature incident management and crisis response capabilities to rapidly mitigate and remediate where needed, and feed this intelligence back into our risk

identification efforts. As a core mechanism of our lifecycle governance, safety cases and mitigations may be updated as a result of post-market monitoring, including information about incidents relating to our frontier safety risk domains. Material updates to a safety case will be submitted to the appropriate governance function for review and might result in updates to the related residual risk assessment or safety case.

Ecosystem Collaboration 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. For example, through our Frontier Model Forum membership alongside other member firms, we signed an agreement designed to facilitate information-sharing about threats, vulnerabilities, and capability advances unique to frontier AI.

Disclaimer: Due to the diverse nature of our product portfolio and the varying roles we play across the AI lifecycle, the selections marked and practices shared in this report may not apply uniformly to all products or deployment environments. Our responses reflect practices that are implemented where contextually appropriate and technically feasible.

Section 3 – Transparency Reporting on Advanced AI Systems

15. Does your organization publish public reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems? Are these clear and understandable by non-technical audiences?

Yes

15.A. If yes, which of the following information is included in your organization's publicly available documentation for any of your AI products?

- Risk management policies
- Summary of results from evaluations of safety, security, and societal risks including potential adverse impacts on individuals or groups (e.g. impacts on human rights, privacy or differential outcomes across user groups).
- Summary of results from red-teaming and other testing (e.g., adversarial evaluations, stress tests, and other pre-deployment safety assessments).
- Capabilities, behaviours, and limitations, including domains of appropriate and inappropriate use or indicative thresholds associated with systemic risks.
- Technical documentation and instructions for use if relevant.
- Plain-language or non-technical summaries

Please provide links or references showing where this information is made public.

Our governance process is grounded in a baseline commitment to transparency and systematic disclosure throughout the advanced 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, ensuring that our strategic stance scales safely alongside capability advancements. To maintain an objective, scientifically grounded

understanding of these systems, our mapping process is fundamentally iterative, evolving alongside the technology.

Updates to transparency disclosures occur on a multi-tiered cadence to inform both downstream actors and the broader ecosystem. We release our Responsible AI Progress Report annually. Concurrently, external model cards and technical reports are published regularly.

References and URLs

The information identified above is systematically externalized across the following official repositories:

- [Responsible AI Progress Reports](#)
- [Transparency Center](#)
- [The Keyword blog](#)
- [Ai.google](#)
- [publicpolicy.google/](#)
- [Google DeepMind model cards repository](#)
- [Google DeepMind recent updates](#) — including Model documentation and papers on responsible AI topics.

15.B. If yes, how often are such reports usually updated?

Other (e.g., triggered by material changes, incidents, or new risks)

16. 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?

- Public documentation
- Protected information sharing
- Regulatory reporting as relevant
- Direct engagement with AI safety institutes, civil society or academic partners

Describe how evaluation results are shared and with which stakeholders (e.g. other organizations, governments, civil society and academia, etc). You may also indicate whether non-technical or plain-language summaries are provided for broader audiences.

We are investing in industry-leading approaches to advance safety and security research, pioneer technical solutions, and share our learnings with the broader ecosystem. We actively share information regarding the outcomes of evaluations of risks and impacts related to advanced AI systems through a multi-layered communication strategy tailored to different stakeholder groups.

- **Public Documentation:** We regularly publish technical reports and model cards to share evaluation results publicly. 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 pre-deployment safety assessments—are published within these public technical reports and model cards to ensure transparency. In addition, information on our progress and approach to Responsible AI development can be found in our annual Responsible AI Progress Reports.
- **Direct Engagement with Civil Society, Academia, and Industry Partners:** We regularly participate in working groups to share information with multi-stakeholder organizations—such as Partnership on AI

(PAI)—and actively contribute to external whitepapers, including PAI's [Transparency Report on Foundation Model Impacts Report](#). We also maintain an 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. Furthermore, where appropriate, external evaluations are conducted by independent external groups on our frontier models, and these results are reported periodically to internal teams and governance groups to improve safety.

- Protected Information Sharing: Under the disclosure protocols established in our [Frontier Safety Framework](#), if we assess that a model has reached a CCL that poses an unmitigated and material risk to overall public safety, we aim to share relevant information with appropriate government authorities where it will facilitate safety of frontier AI. Where appropriate, and subject to adequate confidentiality and security measures and considerations around proprietary and sensitive information, this information may include model information (characteristics of the AI model relevant to the risk it may pose with its critical capabilities), evaluation results (such as details about the evaluation design, the results, and any robustness tests), and mitigation plans (descriptions of our mitigation plans and how they are expected to reduce the risk). We may also consider disclosing information to other external organizations to promote shared learning and coordinated risk mitigation. We will continue to review and evolve our disclosure process over time.

17. Does your organization publish public privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

Yes

If yes, provide links to privacy policies or notices covering personal data, user inputs, and AI-generated outputs.

We clearly disclose our [Privacy Policy](#) and [Terms of Service](#), and maintain established product guidelines for advanced AI outputs to govern user interactions. Our [Generative AI Prohibited Use Policy](#) outlines explicit restrictions that apply to users' interactions with generative AI across the specific products and services that reference this policy.

18. How 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?

- Public documentation
- Provide information or tools to downstream actors on data quality
- Regulatory reporting as relevant
- Direct engagement with civil society or academic partners

Please provide links to any publicly available information.

Our governance process is grounded in a commitment to transparency regarding the data assets that power our advanced AI systems. We provide information about the sources of training data — including annotation and enrichment — through the following public channels:

- Model Documentation: Detailed information on our training data and methods is available in the Google DeepMind repository: <https://deepmind.google/models/model-cards/>.
- Transparency Reporting: Our Transparency Reports shed light on our global operations, including regulatory reporting, content moderation, and government requests: <https://transparencyreport.google.com/>.

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

Yes

If yes, describe additional method(s), including where relevant transparency measures related to AI agents

Ecosystem Transparency and Developer Resources

To demonstrate transparency related to advanced AI systems, we leverage multiple public platforms and specialized toolkits:

- Dedicated Public Web Platforms: Our deepmind.google and ai.google websites offer additional transparency on how we are advancing frontier models and products safely and responsibly.
- Operational Best Practices: 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.

Transparency Measures Related to AI Agents

We continuously invest in research regarding 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. To govern agentic AI capabilities, we introduced the TRUST framework, composed of 5 pillars, the transparency pillar ensures users understand agent actions through notices, controls, and context-appropriate oversight. Data access must align with user needs, and the framework advocates for adaptable oversight, allowing intervention or course-correction. While shifting toward "human-on-the-loop" as reliability improves, explicit intervention remains required for high-stakes actions. Therefore, for sensitive actions — including payments and purchases, posting on social media, and credential use — we require human confirmation before execution, giving users transparency and control over these types of interactions.

20. Does your organization publish statistical insights derived from anonymised, aggregated usage data to advance understanding of AI adoption and its economic impacts in areas including occupations and tasks by region, sector and country? To this end, do you contribute to international cooperation efforts?

Yes

If yes, please provide links or describe the scope of insights published, including where relevant contributions to understanding AI adoption in specific sectors (e.g. public administration, education

systems).

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 the socio-economic and environmental benefits derived from AI systems.

To advance a scientifically grounded understanding of how AI affects the labor market and real-world tasks, we actively collaborate with international partners to simulate downstream economic impacts. Through our formal partnership with the UK AI Security Institute (AISi), we pursue collaborative research to evaluate AI's potential impact on the economy by simulating real-world tasks across different environments—helping policymakers and researchers understand potential long-term effects on occupations and specific economic areas. We are active participants in the [Global AI Adoption Index](#) initiative. The index is a pioneering joint initiative led by the World Bank Group's Development Data Partnership (DDP). The program brings together multilateral development banks, international organizations, and AI laboratories to build a standardized framework and arrangement for sharing aggregated data for tracking global AI adoption.

We also engage directly with government authorities to catalyze progress and support tangible ventures focused on the strategic integration of AI within national education systems and scientific discovery pipelines.

Please estimate the total staff time and annual expenses associated with conducting compliance activities and include any further comments and implementation documentation (optional).

No answer provided

Section 4 – Organizational Governance, Incident Management, and Transparency

21. How has AI risk management been embedded in your organization's governance framework?

- Board of Directors oversight
- Executive oversight
- Senior management oversight
- Governance council or committee

Please describe how AI risk management is integrated into governance structures, including: which roles or bodies have advisory versus decision-making authority; which roles or bodies have final accountability for AI risk-related decisions; and how often policies are reviewed or updated (e.g., following major incidents, regulatory changes, or material updates to systems).

- 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.;

- 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](#).

22. Are relevant staff trained on your organization's governance policies and risk management practices?

Yes

If yes, describe training methods (e.g., onboarding modules, regular sessions, and confidential channels to report concerns).

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.

23. Are steps taken to address reported incidents documented and maintained internally?

Yes

If yes, describe how incident records are kept and used for follow-up (e.g., incident logs, root-cause analysis reports, internal tracking systems). You may indicate whether and how you define escalation thresholds for when incidents must be surfaced to senior leadership or external parties.

Our incident response programs are managed by teams of expert incident responders across many specialized functions to ensure each response is well-tailored to the challenges presented by each incident.

At the resolution stage, the focus is on investigating the root cause, limiting the impact of the incident, resolving immediate safety, security, or other risks (if any), implementing necessary fixes as part of remediation, and recovering affected systems, data, and services.

Following the successful remediation and resolution of a data incident, the incident response team evaluates the lessons learned from the incident. When the incident raises critical issues, the incident commander might initiate a post-mortem analysis. We strive to learn from every incident and implement preventative measures to avoid future incidents. Actionable insights from incident analysis enable us to enhance our tools, training, processes, overall security and privacy data protection program, security policies, and response efforts.

24. Does your organization share research and best practices on addressing or managing risk?

Yes

If yes, describe how research or good practices are shared (e.g., publications, conferences, collaborations with other sectors).

We are committed to transparently sharing research and best practices to benefit the broader ecosystem. Because we understand that AI, as a still-emerging transformative technology, involves evolving complexities and risks, we work with stakeholders across the technology sector, government, and civil society regarding AI policies, remain abreast of emerging challenges, and establish, share, and learn from industry best practices and research.

Publications and Public Updates

We actively publish our findings, risk management methodologies, and research-backed guidelines across multiple public platforms to ensure accessible transparency:

- **Research Papers:** We have continuously published research 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](#). Our publications address specific, critical risk domains - including holistic safety evaluations, the study of misuses of generative AI, and broader ethical concerns regarding advanced AI systems.
- **Blogs and Reports:** To reach audiences beyond the specialized research community, we share updates on dedicated platforms—such as [research.google/blog](#), [deepmind.google/blog/](#), [publicpolicy.google](#) and [The Keyword blog](#), as well as in our [annual Responsible AI Progress Report](#).
- **External and Contextual Frameworks:** We contribute to external whitepapers in partnership with multi-stakeholder consortia—such as the Advertising Council Research Institute's "[Calibrating Trustworthiness in GenAI](#)". Additionally, we publish policy papers and frameworks for addressing and managing risk, such as "[The TRUST Framework: A Policy Approach to Agentic AI](#)," designed to provide guiding principles for the governance of AI agents

Tools and Actionable Resources

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. The toolkit also 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 targeting hate speech, harassment, sexually explicit content, and dangerous content.

Cross-Sector Partnerships and Collaborations

We heavily emphasize structured, cross-sector partnerships to develop technical standards, build joint solutions, and securely share threat intelligence:

- **Multi-Stakeholder Working Groups:** 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), the Frontier Model Forum (FMF), and government partners, such as the U.K. AI Security Institute. As of June, we are also founding members of the Appia Foundation – an international group that builds on standards from bodies like ISO/IEC to create clear criteria, allowing organizations across the AI value chain to prove they meet their obligations, including those related to managing AI risk.
- **Information Sharing Agreements:** To stay ahead of emerging vulnerabilities, we signed a first-of-its-kind agreement with fellow members of the FMF—designed to facilitate information-sharing about threats, vulnerabilities, and capability advances unique to frontier AI.
- **Joint Governance Initiatives:** We launched the Coalition for Secure AI (CoSAI) with industry partners to share security expertise under the guidance of OASIS Open. Additionally, we have collaborated with PAI to jointly launch responsibility frameworks for safe deployment, synthetic media, and data enrichment sourcing.

Ecosystem Funding and Community Engagement

To foster independent validation and incentivize the global research community to proactively uncover potential system limitations, we invest in external safety initiatives:

- **Data-Centric Challenges:** We partnered with universities across India, Nigeria, Ghana, to conduct community-centered harms analysis of image generation models. Our methodology and findings are shared [publicly](#) for advancing global AI safety standards. We continue to closely partner with MLCommons and Korea AI Safety Institute, among others to build culturally-specific and geographically-relevant AI safety datasets.
- **Multi-agent AI Safety Research:** In a partnership between Google DeepMind, Google.org, [Schmidt Sciences](#), the [Cooperative AI Foundation](#), the [Advanced Research and Invention Agency](#), we launched a new \$10M technical research fund to help understand how multi-agent systems behave, how to mitigate new potential risks, and how these independent systems interact safely and predictably across various networks.

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

Yes

If yes, please indicate which standards or frameworks apply:

- ISO/IEC standards (e.g., ISO/IEC 27001, ISO/IEC 42001)
- NIST AI Risk Management Framework (AI RMF)
- OECD AI Principles
- Industry-led best practices or codes of conduct

- Other international or regional standards (please specify): U.K. AI Security Institute's approach to Responsible Capability Scaling and open-source initiatives managed under OASIS Open.

If yes, describe which international technical standards or best practices your organization uses.

We are committed to developing safe, secure, and trustworthy advanced AI systems by aligning our internal frameworks with recognized international standards and multi-stakeholder guidelines.

ISO/IEC Standards and NIST AI RMF Alignment As outlined in our [AI Principles](#), we participate in the development and adoption of international technical standards to manage AI risks across the product lifecycle. Google Cloud, Gemini app (as a Workspace Core Service), and Google Workspace are ISO/IEC 42001-2003 certified. This international standard outlines the fundamental requirements for establishing and continually improving an Artificial Intelligence Management System (AIMS) — specifying repeatable management controls to systematically mitigate AI risks. We also share our technical expertise to advance AI measurement and benchmarks by participating in NIST's AI Consortium and the Center for AI Standards and Innovation (CAISI). Additionally, we engage in open comment periods of foundational security publications like NIST 800-218 (SSDF) and NIST IR 8596 (Cyber AI Profile) to ensure Google's real-world experience deploying safe AI systems informs best practices shared across the industry.

Industry-Led Best Practices and Voluntary Commitments We heavily emphasize collaborative, ecosystem-wide responsibility by engaging with multi-stakeholder groups and signing voluntary codes of conduct to establish shared industry benchmarks:

- We participate in working groups on industry standards and 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 (PAI), the Frontier Model Forum and the Appia Foundation.
- We jointly launched comprehensive responsibility frameworks with PAI to provide concrete guidance for safe deployment, synthetic media, and data enrichment sourcing alongside industry peers, academics, and civil society organizations.
- We have signed key voluntary commitments to promote safety-by-design — 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 actively support the G7's voluntary Code of Conduct to facilitate safe, trustworthy, and secure AI worldwide.

Other International and Regional Frameworks Acknowledging the evolving nature of the technology as context for our risk mitigations, we align our internal frameworks with relevant regional standards and open-source consortia.

Our Frontier Safety Framework 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 Security Institute. Additionally, we launched the Coalition for Secure AI (CoSAI) with industry partners as a major milestone and application of Google's [Secure AI Framework](#) (SAIF). Operating under the guidance of OASIS Open — an international standards and open-source consortium — CoSAI functions as an open initiative to equip practitioners and developers with the technical guidance and tools they need to create AI systems that are Secure-by-Design. We also support the OECD's AI Principles and the G7's voluntary Code of Conduct, which aims to promote safe, trustworthy and secure AI worldwide.

Any further comments and implementation documentation.

Due to the diverse nature of our product portfolio and the varying roles we play across the AI lifecycle, the selections marked and practices shared in this report may not apply uniformly to all products or deployment environments. Our responses reflect practices that are implemented where contextually appropriate and technically feasible.

Section 5 – Content Authentication & Provenance Mechanisms

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

- User disclosure (e.g., text notice, verbal statement, or system label)
- Persistent visual or audio indicators (e.g., watermark, icon, chime, banner) during AI interaction
- Distinct branding or labeling for AI-generated content (e.g., “AI-generated” tag, metadata label)
- Documentation or FAQs explaining when and how AI is used in the system
- Access to provenance information (e.g., model version, content authenticity signals)
- Disclosure to end-users when an AI system is acting autonomously or executing actions on their behalf (e.g., making purchases, sending communications, modifying files)

Provide any additional details or context on how or when users are informed they are interacting with advanced AI systems (e.g. if such mechanisms are planned, in testing, or applied only to certain products).

As outlined in our [AI Principles](#), we pursue AI responsibly throughout the AI development and deployment lifecycle, and we work to advance user understanding of AI through innovative developments in provenance technology, our research-backed explainability guidelines, and AI literacy education.

User Disclosure, Documentation, and FAQs Our governance process is grounded in setting clear, consistent expectations for end-users across the digital ecosystem.

- 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.
- 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 proactively empower users with robust 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 the system interacts with their data.

Persistent Indicators, Labeling, and Provenance Information Because we understand that AI involves evolving complexities, we implement by design specific technical mitigations to verify authenticity and surface provenance signals.

- We developed SynthID to help address issues around provenance by embedding digital watermarks directly into AI-generated content — including text, audio, images, and video.
- 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.
- 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.
- In May 2026, we added C2PA content credential detection to the Gemini App to share more signals with users about the origin and authenticity of online content. We also announced that we were expanding C2PA and SynthID to Search and Chrome so even more people can get helpful context relevant to the origin and authenticity of a given piece of content.
- These technical solutions are complemented by Backstory — 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, investigating how it has been used on the internet and providing other helpful metadata.

27. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems?

Yes

27.A. If yes, how?

- Latent disclosure (e.g., hidden provenance information)
- Manifest disclosure (e.g., visible label)
- Detection tool(s) for hidden provenance signals (e.g., watermark or metadata checks)
- Use of international or industry standards (e.g., C2PA)

27.B. If yes, does your organization use international technical standards or best practices when developing or implementing content provenance?

Yes

27.C. Does your organization collaborate on or invest in research to advance content authentication and provenance?

Yes

Describe how provenance mechanisms are implemented and which standards or tools are used, and any collaborations or investments to improve content authentication (e.g., C2PA credentials, joint R&D projects, open-source initiatives, standard-setting groups). You may also indicate whether these mechanisms are tested for robustness against removal, tampering, or circumvention.

Implementation and Technical Mechanisms Because we understand that AI, as a still-emerging transformative technology, involves evolving complexities and risks, we pursue AI responsibly by implementing a multi-faceted approach to provenance using both invisible watermarking and manifest disclosures. Our technical implementation operates sequentially across different media types:

- **Latent Disclosure via 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. These watermarks are imperceptible to humans but remain detectable by specialized software. We have improved SynthID to identify text and video generated by the Gemini app and web experience.
- **Detection Tools:** To enable verification, we announced SynthID Detector — a verification portal that allows users to quickly and efficiently upload content and scan for SynthID watermarks to identify if content was generated or modified by our AI tools. Furthermore, to help more organizations identify AI-generated media, we are launching a new AI Content Detection API on Google Cloud's Gemini Enterprise Agent Platform. In May 2026, we added C2PA content credential detection to the Gemini App to share more signals with users about the origin and authenticity of online content. We also announced that we were expanding C2PA and SynthID to Search and Chrome so even more people can get helpful context relevant to the origin and authenticity of a given piece of content.
- **Manifest Disclosure and Context:** We deploy visible notices to ensure audience awareness. For example, YouTube requires creators to disclose when they have created realistic altered or synthetic content, displaying a visible label to indicate when the content being watched is synthetic.

International Technical Standards and Robustness We are 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.

Collaborations and Investments We actively participate in, collaborate on, and invest in projects advancing AI safety, security, and trustworthiness. To advance the state of content authentication, we pursue multiple cross-cutting initiatives:

- We joined the C2PA as a steering committee member to support this cross-industry effort to provide more transparency and context on digital content.

- We joined the Partnership on AI (PAI) Synthetic Media Framework to jointly develop responsibility frameworks alongside industry peers.
- We are executing joint R&D projects — partnering with NVIDIA to watermark AI-generated video from their Cosmos world foundation models. We have expanded this partnership on SynthID to other industry leaders such as OpenAI, Kakao and ElevenLabs.
- We support open-source initiatives by making our in-house SynthID text watermarking tool available on Hugging Face — making it easier for any developer to apply watermarking for their generative text models and detect outputs from their own large language models.

Any further comments and implementation documentation.

Due to the diverse nature of our product portfolio and the varying roles we play across the AI lifecycle, the selections marked and practices shared in this report may not apply uniformly to all products or deployment environments. Our responses reflect practices that are implemented where contextually appropriate and technically feasible.

Section 6 – Research & Investment to Advance AI Safety & Mitigate Societal Risks

28. Does your organization advance research and investment in the following areas?

- Security
- Safety
- Data quality and representativeness
- Information integrity
- System performance and outcomes (e.g. accuracy, consistency)
- Explainability or interpretability
- Transparency
- Robustness or reliability

If yes, please describe key research or investment initiatives in the areas selected above (e.g., dedicated safety teams, funding of research programs, contributions to open-source safety tools, publishing research, partnerships and collaborative research initiatives). You may also indicate whether related outputs are publicly shared, shared with partners or within consortia or not publicly disclosed.

We continuously invest in and publish papers on responsible AI topics—made publicly available via Google’s Research publications repository and Google DeepMind’s publications archive—to advance safety, security, and transparency across the ecosystem. We pursue a wide range of internal research, funding programs, and open-source contributions to ensure systems are robust by design.

Internal Research Focus Areas Our internal scientific research focus is iteratively updated to map potential vulnerabilities and establish proactive mitigations across several core fields:

- **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.

- **Emerging Lifecycle Risks:** We invest in exploring emerging risks from AI in areas like cybersecurity, dangerous capabilities, misinformation, adversarial use of generative AI, socio-technical approaches to red-teaming, health equity, bias, and privacy.
- **AI Misuse:** Mapping the potential misuse of generative AI is a core area of research that directly shapes how we evaluate our own models and calibrate deployment guardrails.

Ecosystem Funding and Grants Our governance process is grounded in enabling the broader ecosystem with financial backing to improve digital safety and support safe AI development across the industry:

- **Google Academic Research Awards (GARA):** This program supports external research across major global and technical domains—including quantum neuroscience, AI for security, and Trust & Safety across online ecosystems.
- **Multi-agent AI Safety Research:** In a partnership between Google DeepMind, Google.org, [Schmidt Sciences](#), the [Cooperative AI Foundation](#), the [Advanced Research and Invention Agency](#), we launched a new \$10M technical research fund to help understand how multi-agent systems behave, how to mitigate new potential risks, and how these independent systems interact safely and predictably across various networks.

Partnerships and Collaborative Initiatives We actively partner with industry peers, research institutes, and civil society to push the technical boundaries of AI safety and share our findings with the broader community:

- **UK AI Security Institute (AISi) Collaboration:** Our joint research under a formal Memorandum of Understanding focuses on foundational safety research—including developing techniques to monitor an AI system's reasoning or chain-of-thought (CoT) processes, assessing how model misalignment impacts human wellbeing, and evaluating economic impact by simulating real-world tasks across environments to understand downstream effects on the labor market.
- **Multi-Stakeholder Engagement:** We regularly participate in working groups with organizations like the Partnership on AI (PAI)—contributing to external whitepapers such as PAI's [Documenting the Impact of Foundation Models](#). We partnered with universities across India, Nigeria, Ghana, to conduct community-centered harms analysis of image generation models. Our methodology and findings are shared [publicly](#) for advancing global AI safety standards.

Open-Source Security Tools To enhance industry-wide robustness, we actively invest in open-source initiatives to ensure advanced AI systems are secure by design:

- **Coalition for Secure AI (CoSAI):** Launched in 2024 with industry partners, CoSAI marks the first major milestone and application of Google's [Secure AI Framework](#) (SAIF), allowing practitioners to collectively invest in AI security research, share safety expertise, and build technical open-source solutions under the guidance of OASIS Open.

Information Integrity and Sector-Specific Research Dedicated units within our organization research threats to information integrity to improve data quality and enhance user transparency:

- **Context and Provenance Research:** We publish research on determining trustworthiness through context and provenance—demonstrating how better assessment tools empower users to make informed decisions about what they encounter on the internet.

- Jigsaw-led Research: Our Jigsaw unit explores threats to open societies and develops scalable solutions on platforms like YouTube. This includes the deployment of Accuracy Prompts (APs) that serve users bite-sized digital literacy tips to encourage critical evaluation.

29. Does your organization pursue research or investment related to socio-economic or environmental considerations in the development and deployment of AI?

- Environmental considerations (e.g., energy use, computational efficiency, resource use)
- Socio-economic considerations (e.g., AI adoption and productivity, workforce transitions, skills development, access to AI tools)

If yes, please provide examples of relevant research or investment initiatives for the areas selected above.

As outlined in our [AI Principles](#), we are committed to focusing on solving real-world problems and measuring the tangible outcomes of our work—focusing our efforts where the likely overall benefits substantially outweigh foreseeable risks. Because we understand that AI, as a still-emerging transformative technology, involves evolving complexities, we continuously invest in research covering topics on the socio-economic and environmental dimensions of advanced AI systems.

Socio-Economic Considerations Our governance process is grounded in maximizing the beneficial impacts of technology while actively supporting workforce readiness and equitable access across the ecosystem:

- **Workforce Transitions and Skills Development:** We have committed \$120 million for AI education and training around the world—launching targeted training programs for businesses, developers, and younger learners. To address specific industrial adjustments, we provided a \$10 million grant to the Manufacturing Institute to integrate AI into vocational training—an initiative designed to train 40,000 US manufacturing workers. Furthermore, we launched the AI & Executive Economy Research Program to offer data-driven insights for navigating the evolving nature of work.
- **Access to AI Tools and Literacy:** Google.org has made significant investments in AI literacy and education—including the Experience AI program with Raspberry Pi, university initiatives across Europe, the Middle East, and Africa, and new education tools like a seasonal class management system for teachers.

Environmental Considerations We take a scientifically grounded approach to minimizing environmental risks while developing AI-powered solutions to support global sustainability:

- **Energy Use and Computational Efficiency:** We conduct ongoing research into the lifecycle impacts of computing infrastructure—including tracking environmental risks and Tensor Processing Unit (TPU) emissions efficiencies gained over generations, as documented in [Life-Cycle Emissions of AI Hardware: A Cradle-To-Grave Approach and Generational Trends](#).
- **Climate and Sustainability Research:** We invest in AI for Science Research covering critical fields—including weather forecasting, fusion, and materials science. This includes exploring how advanced AI can accelerate the development of renewable energy—specifically by designing new materials for batteries or solar panels, and optimizing how the electrical grid operates and integrates renewable sources.

- **Disaster Preparedness and Resource Management:** We prioritize environmental resilience tools to assist communities in the path of climate-related disasters. Our Flood Hub system provides early flood warnings up to seven days in advance, helping protect over 460 million people across more than 80 countries. Additionally, we invested in FireSat technology—which leverages specialized satellites to detect and track early-stage wildfires, with imagery intended to be updated every 20 minutes globally to scale the speed and accuracy of firefighting responses.

Any further comments and implementation documentation.

Due to the diverse nature of our product portfolio and the varying roles we play across the AI lifecycle, the selections marked and practices shared in this report may not apply uniformly to all products or deployment environments. Our responses reflect practices that are implemented where contextually appropriate and technically feasible.

Section 7 – Advancing Human and Global Interests

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

Yes

If yes, describe any digital literacy or training initiatives your organization supports (e.g., user education, public AI-education programs, school or university partnerships, community workshops). You may also indicate whether these efforts include or prioritise underserved or underrepresented communities.

We are committed to focusing on broad educational accessibility and user awareness across the advanced AI lifecycle—ensuring that downstream actors and end-users understand both the capabilities and limitations of advanced AI technologies. To complement the structural mitigations implemented at the model level, we support a wide array of digital literacy, education, and training initiatives globally to reinforce alignment with safety standards.

Financial Commitments and Public Education Programs

- We have committed \$120 million for AI education and training around the world and have launched AI training for businesses, developers, and younger learners.
- We scale digital literacy and training to improve user awareness via specific educational programs—including Google AI Essentials and the Applied Digital Skills Program.

School and University Partnerships

- Google.org has made significant financial investments in AI literacy and education 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.
- We co-developed Experience AI with the Raspberry Pi Foundation—an educational program that offers cutting-edge resources on AI for teachers and students aged 11–14.

- We deploy scalable AI literacy training that provides, free of charge, comprehensive Gemini training to all 6 million K-12 and higher education faculty in the U.S..

User Education on Capabilities and Limitations

- We work to advance user understanding of AI through innovative developments in provenance technology, our research-backed explainability guidelines, and AI literacy education to protect the integrity of the digital ecosystem.
- We employ in-product mechanisms—including interstitials and FAQs, such as the FAQ for Gemini Apps—that provide users with the knowledge and answers that confirm they are engaging with an advanced system.
- Our Explainability Rubric provides guidance to help users understand how AI operates and makes decisions.
- Where possible and 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.
- The Super Searchers content was refreshed in November 2025 to reflect innovations in Search—such as AI Mode and AI Overviews—incorporating these tools into the curriculum to help users identify and evaluate the accuracy of information found online.

Prioritization of Underserved and Underrepresented Communities

- To ensure equitable access across the global ecosystem, we provide various forms of digital literacy and educational materials—for example, through scholarships to community resources offered by Google DeepMind.

31. Does your organization pursue AI projects that aim to deliver societal or public benefits, including through partnerships with civil society and community groups?

Yes

If yes, please provide examples, including any collaborations with civil society, community groups, or efforts to address broader societal challenges.

We believe that maximizing the positive impact of technology requires ecosystem-wide responsibility. We actively pursue and fund AI projects aimed at delivering tangible public benefits through direct collaborations with civil society, academic institutions, and community groups.

Support for Social Good

- We support the Google Developer Student Clubs (GDSC), where members from around the world have come together since 2020 to create innovative solutions to tackle the world's most pressing challenges.
- Our *AI in Action* research highlights capabilities across computer vision, generative AI, natural language processing, and multimodal AI to showcase how AI is altering problem-solving across several development areas — with a dedicated spotlight on AI-powered innovation in health, education, and climate.
- Through Google Research's AI for Social Good awards — in partnership with Google.org — we bring nonprofits and academics together to collaborate on machine learning-based technologies that can improve people's lives.

- Through the [Google.org Impact Challenge: AI for Government Innovation](#), we provide significant funding, ranging from \$1M to \$3M USD per organization, and dedicated pro bono technical support to nonprofits, social enterprises, and academic institutions that are partnering with governments. These projects leverage generative and agentic AI to transform public services for social good across critical areas such as health, resilience, and economy.
- [Our AI Collaboratives](#) represent a new funding approach that actively unites public, private, and nonprofit organizations, and researchers to create AI-powered solutions for pressing global challenges. For instance, the AI Collaborative: Wildfires is a strategic initiative that has seen Google.org contribute tens of millions of dollars and convene more than 15 organizations to leverage AI for identifying, tracking, and mitigating wildfires. Similarly, the AI Collaborative: Food Security brings together 10 organizations to strengthen global food systems and improve food security for vulnerable populations through AI technologies, collaborative research, data-sharing, and coordinated action. These collaborations underscore our belief that AI's potential lies not just in the technology itself, but also in the partnerships that make its responsible application possible.

Healthcare and Scientific Discovery

- We maintain a partnership with Wellcome Trust on a multi-year investment in AI research for treating anxiety, depression, and psychosis.
- We worked collaboratively with Grand Challenges Canada and the 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.
- Through [Google.org Impact Challenge: AI for Science](#) is a \$30M global open-call designed to empower researchers and organizations with funding, from \$500K to \$3M USD, and technical expertise to accelerate scientific breakthroughs. This initiative specifically focuses on projects in AI for Health & Life Sciences and AI for Climate Resilience & Environmental Science.
- Through a five-year partnership with the Wellcome Sanger Institute, Google DeepMind and Google.org are committing \$5M USD annually to launch a new global AI consortium for genomics. This strategic initiative focuses on generating large-scale, AI-ready biological datasets to train and power the next generation of foundational machine learning models, ultimately accelerating breakthroughs and decoding the complexities of biomedical science for the global research community.

Climate Resilience and Disaster Relief

- In a collaborative partnership with the nonprofit GiveDirectly in Nigeria, Google's forecasts were used to trigger anticipatory cash transfers — allowing over 3,250 households to evacuate and secure assets before flooding hit, resulting in a 90% drop in food insecurity.
- To scale disaster resilience, we collaborated with the government of Nigeria, the UN Country Team, and the UN Office for the Coordination of Humanitarian Affairs to support a \$7 million initiative that delivered aid before disaster struck via Nigeria's first AI-driven large-scale Floods Anticipatory Action Program.

Any further comments and implementation documentation.

Due to the diverse nature of our product portfolio and the varying roles we play across the AI lifecycle, the selections marked and practices shared in this report may not apply uniformly to all products or deployment environments. Our responses reflect practices that are implemented where contextually appropriate and technically feasible.