Google

Fulfilling the Voluntary Industry Commitments on AI

In July, we were proud to be part of a milestone moment at the White House that brought industry together to commit to advancing responsible practices in the development of artificial intelligence. Building on work that we started back in 2014, we committed to specific practices for the safe, secure, and trustworthy development and use of AI. We've made significant progress on these commitments, which will help ensure AI is developed boldly and responsibly, for the benefit of everyone.

Here's a look at our work on the commitments so far.

Safety

For years, we have been building AI responsibly by design with guardrails to protect users and society, guided by our AI Principles. We put our models and products through adversarial testing to mitigate risks, and our teams are advancing the state of the art on topics like helping AI communicate in safer ways, preventing advanced models from being misused, and designing systems to be more ethical and fair. For Generative AI product launches, we're leveraging trust and safety policy and enforcement expertise, which includes setting policies and standards, conducting safety evaluations and red teaming, proactively mitigating risks related to harmful bias, discrimination, privacy, security, and more - and analyzing user feedback at scale to improve models on an ongoing basis.

Throughout the ecosystem, we're supporting work to continue rigorously testing internal and external models. For example, participating in the White House-sponsored red teaming event at DEFCON and participating in the launch of the Frontier Model Forum to develop standards and benchmarks for emerging safety and security issues of frontier models.

Here's a look at our progress.

Commitment 1: Commit to internal and external red-teaming of models or systems in areas including misuse, societal risks, and national security concerns, such as bio, cyber, and other safety areas.

- Participated in the White House-sponsored red teaming event at DEFCON, which drew over 2,000 people to test industry-leading large language models (LLMs) in an effort to better understand risks and limitations of these advanced technologies.
- Hosted LLM Hackathon for 30 expert Bug Hunters alongside DEFCON to test Google's public AI offerings in more technical depth.
- Running ongoing and systematic adversarial testing, conducting safety and fairness evaluations in multiple languages, and monitoring and measuring performance for major Generative AI product launches to ensure alignment with our content safety policies. Conducted ongoing in-product and user experience analysis of major Generative AI launches.
- Hosted an internal, company-wide LLM red teaming "Hack-AI-thon" with hundreds of security, trust & safety, and responsible AI experts. There were over 2,600 safety-focused conversations with the model, which helped us further enhance the safety and security of the technology.
- Established a dedicated Google AI Red Team focused on different risks, including security, abuse, bias and other societal risks.

Commitment 2: Work toward information sharing among companies and governments regarding trust and safety risks, dangerous or emergent capabilities, and attempts to circumvent safeguards.

- With other leading AI companies, we established a cross-industry forum – the Frontier Model Forum – to develop standards and benchmarks for emerging safety and security issues of frontier models.
- Joined the Partnership on AI (PAI) Synthetic Media Framework to help develop and foster best practices across the industry for the development, creation, and sharing of media created with Generative AI.
- Participated in a number of information sharing sessions about Generative AI, including at the National Conference of State Legislatures Summit where we hosted a Bard x Safety kiosk, presented on Generative AI safety, and sat for a fireside chat. We're looking forward to participating in more summits in the near future.

Security

We design our products to be secure-by-default — and our approach to AI is no different. We recently introduced our Secure AI Framework (SAIF) to help organizations secure AI systems, and we expanded our bug hunters programs (including our Vulnerability Rewards Program) to incentivize research around AI safety and security.

To continue to fulfill these commitments, we're prioritizing cybersecurity safeguards to protect proprietary and unreleased models and we're participating in industry-wide events to support broader protections for governments, companies, and civil society, like the Defense Advanced Research Projects Agency's (DARPA) AI Cyber Challenge, which will aim to identify and fix software vulnerabilities using AI.

Here's a look at our work so far.

Commitment 3: Invest in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights.

- Launched a general framework that can be applied to the safe and secure development and deployment of AI systems.
- We're developing open source tooling and infrastructure that will support the security of models, and internal standards and technical controls to guarantee the provenance, confidentiality, and integrity of models across Google.

Commitment 4: Incent third-party discovery and reporting of issues and vulnerabilities.

- Expanded the scope of our Bug Hunter Program (including Vulnerability Rewards Program) to reward and incentivize anyone to identify and report vulnerabilities in our AI systems. Last year we issued over $12 million in rewards to security researchers who tested our products for vulnerabilities.
- Announced our participation in DARPA's AI Cyber Challenge (AIxCC) to identify and fix software vulnerabilities using AI with $20 million in rewards.

Trust

Since 2018, we've been guided by our AI Principles in our bold and responsible development of AI products, and, importantly, our first Principle is "be socially beneficial." We also established a governance team to put our AI Principles into action by conducting ethical reviews of new systems, avoiding bias and incorporating privacy, security and safety. And our Responsible AI Toolkit helps developers pursue AI responsibly as well.

To continue to build trust, we're taking steps to help promote trustworthy information online, like launching a beta version of our SynthID watermarking tool and publishing reports on model or system capabilities, limitations, and domains of appropriate and inappropriate use, including discussion of societal risks. We also expanded our ads policies to require advertisers to disclose when their election ads include material that's been digitally altered or generated and depicts real or realistic-looking people or events in all countries where we have election ads verification.

None of us can get AI right on our own, which is why we're working with groups like Partnership on AI, ML Commons, and Frontier Model Forum to promote the responsible development of new Generative AI tools. We're optimistic that our continued, shared work will help solve some of our biggest challenges, because we've already seen incredible progress, like using AI to improve the accuracy and expanding the availability of breast cancer screenings, help people and cities adapt to extreme heat, and forecast floods.

Here's a look at our work so far.

Commitment 5: Develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated, including robust provenance, watermarking, or both, for AI-generated audio or visual content.

- We shared at I/O that we're investing in tools to facilitate detection of synthetic media, including watermarking and metadata, and we're building our models to include watermarking and other techniques from the start.
- Joined the Partnership on AI (PAI) Synthetic Media Framework.
- Launched a beta version of SynthID: a tool for watermarking and identifying AI-generated images.
- Expanded election ads policies to require disclosure of material that has been digitally altered or generated and depicts real or realistic-looking people or events in all countries where we have election ads verification.

Commitment 6: Publicly report model or system capabilities, limitations, and domains of appropriate and inappropriate use, including discussion of societal risks, such as effects on fairness and bias.

- Publishing regular reports on model or system capabilities, limitations, and domains of appropriate and inappropriate use. Most recently, we shared a midyear update on progress on our AI Principles in practice focused on adapting existing responsible AI practices for Generative AI.

- Sharing red team best practices so that others can learn from our experience in developing an AI Red Team. We published a report that examines our work to create an AI Red Team, which includes three areas: 1) what red teaming in the context of AI systems is and why it is important; 2) what types of attacks AI red teams simulate; and 3) lessons we have learned that we can share with others.

Commitment 7: Prioritize research on societal risks posed by AI systems, including on avoiding harmful bias and discrimination, and protecting privacy.

- Internally and externally we're committed to sharing ongoing research into AI's benefits and risks.

- We're leveraging trust and safety policy and enforcement expertise for responsible GenAI launches. This work includes setting policies and standards, conducting safety evaluations and red teaming, proactively mitigating risks related to harmful bias, discrimination, privacy, security, and more - and analyzing user feedback at scale to improve models on an ongoing basis.

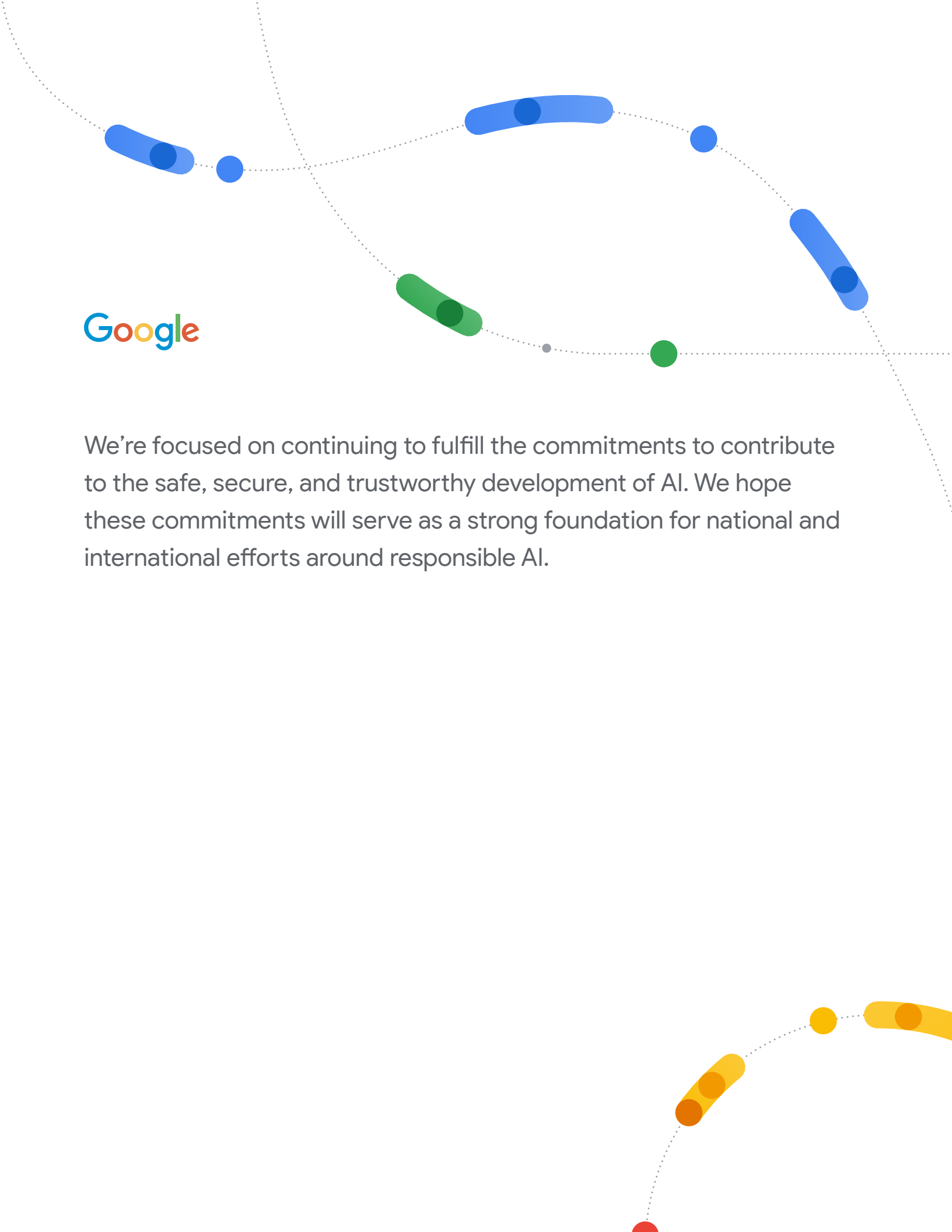
- Launched the Digital Futures Project to provide grants to leading global think tanks and academic institutions to support more research and dialogue about AI and society.

Commitment 8: Develop and deploy frontier AI systems to help address society's greatest challenges.

- We've been guided by our AI Principles since 2018, including #1 which is: "Be socially beneficial." And we've long supported research and development to use AI for good.

- For example, we have been working on AI projects helping scientists better detect breast cancer, forecast floods, limit the warming effects of contrails in air travel, accelerate nuclear fusion which could help develop clean energy, advance healthcare breakthroughs, and accurately predict 3D models of protein structures.

- Our approach to frontier models is no different–we will continue to use the most advanced AI technology we develop to push the boundaries of scientific discovery and help people use knowledge to benefit humanity.



We're focused on continuing to fulfill the commitments to contribute to the safe, secure, and trustworthy development of AI. We hope these commitments will serve as a strong foundation for national and international efforts around responsible AI.