Written statement of Tulsee Doshi, Google
11/1/2023
United States Senate AI Insight Forum: High Impact AI

Thank you to Majority Leader Schumer and Senators Young, Heinrich, and Rounds for bringing us together, and for your leadership in this critical area. Google's ongoing work in AI powers tools that billions of people use every day — including Google Search, Translate, Maps and more. The AI in these tools will have a significant impact on people's everyday lives. But we're also incredibly excited about using AI to solve major societal issues and we have numerous examples of AI initiatives that will have a game changing impact:

- Google **FloodHub** displays forecasts for riverine floods around the world based on AI models. All information is free of charge to help people directly at risk, and to help governments and aid organizations access critical information.
- Google AI can also help with **wildfires**. We've advanced our use of satellite imagery to train AI models to identify and track wildfires in real time, helping predict how they will evolve and spread. We first launched this wildfire tracking system in the US, Canada, Mexico, and continue to roll it out in other countries, helping inform our users and firefighting teams with millions of views in Google Search and Maps.
- Google Maps has implemented **fuel-efficient** routing which uses AI to suggest routes that have fewer hills, less traffic, and constant speeds with the same or similar ETA. Since launching in October 2021, it's estimated to have helped prevent more than 2.4 million metric tons of CO2e emissions — the equivalent of taking approximately 500,000 fuel-based cars off the road for a year.
- Project Green Light, a Google Research initiative, uses AI and Google Maps driving trends to model traffic patterns and make recommendations for optimizing the existing traffic light plans. City engineers can implement these in as little as five minutes, using existing infrastructure. By optimizing not just one intersection, but coordinating across several adjacent intersections to create waves of green lights, **cities can improve traffic flow and further reduce stop-and-go emissions.**
- In **health care** Google teams are working on AI screening for breast cancer, colon cancer, and lung cancer – three of the cancers that kill the most Americans. We're also working on major health discoveries, with tools like DeepVariant and AlphaFold – which help with understanding DNA and proteins.
- And as we know **language** is fundamental to how people communicate and make sense of the world, with more than 7,000 languages spoken around the world only a few are well represented online today. That means traditional approaches to training language models on text from the web fail to capture the diversity of how we communicate globally. So we have been working on the 1,000 Languages Initiative, an ambitious commitment to build an AI model that will support the 1,000 most spoken languages,

bringing greater inclusion to billions of people in marginalized communities all around the world.

We've learned that AI has the potential to have a far-reaching impact on the global crises facing everyone, while at the same time expanding the benefits of existing innovations to people around the world.

This is why AI must be developed responsibly, in ways that address identifiable concerns like fairness, privacy, and safety, with collaboration across the AI ecosystem. And it's why — in the wake of announcing that we were an "AI-first" company in 2018 — we shared our AI Principles and have since built an extensive AI Principles governance structure and a scalable and repeatable ethics review process. To help others develop AI responsibly, we've also developed a growing Responsible AI toolkit.

Each year, we share a detailed report on our processes for risk assessments, ethics reviews and technical improvements in a publicly available annual update — 2019, 2020, 2021, 2022 — supplemented by a brief, midyear look at our own progress that covers what we're seeing across the industry.

Today, generative AI is receiving more public focus, conversation and collaborative interest than any emerging technology in our lifetime. That's a good thing. This collaborative spirit can only benefit the goal of AI's responsible development on the road to unlocking its benefits, from helping small businesses create more compelling ad campaigns to enabling more people to prototype new AI applications, even without writing any code.

For our part, we've applied the AI Principles and an ethics review process to our own development of AI in our products — generative AI is no exception. What we've found in recent months is that there are clear ways to promote safer, socially beneficial practices to generative AI concerns like unfair bias and factuality. We proactively integrate ethical considerations early in the design and development process and have significantly expanded our reviews of early-stage AI efforts, with a focus on guidance around generative AI projects.

Our Approach

I'd like to share three of our best practices based on this guidance and what we've done in our pre-launch design, reviews and development of generative AI: design for responsibility, conduct adversarial testing and communicate simple, helpful explanations.

1. Design for responsibility.

It's important to first identify and document potential harms and start the generative AI product development process with the use of responsible datasets, classifiers and filters to address those harms proactively. From that basis, we also:

- Participate in workshops alongside the research community to identify comprehensive ways to build trustworthy AI. Recently, we've supported and helped advance forums like Ethical Considerations in Creative Applications of Computer Vision and Cross-Cultural Considerations in NLP.
- Develop policies to guide development, based on harms identified early in the research and ethics review process.
- Use technical approaches such as classifiers and other tools to flag and filter outputs that violate policies, and additional methods such as those in the Responsible AI toolkit. Recently, we've added a new version of the Learning Interpretability Tool (LIT) to the toolkit, for model debugging and understanding, and the Monk Skin Tone Examples (MST-E) dataset to help AI practitioners use the inclusive Monk Skin Tone (MST) scale.
- Gather a group of external experts across a variety of fields such as law and education for robust discussions on equitable product outcomes. Our ongoing Equitable AI Research Roundtable (EARR), for example, continues to meet with thought leaders who represent communities historically underrepresented in AI leadership positions, focusing on generative AI topics.
- Offer an experimental, incremental release to trusted testers for feedback.
- Proactively engage with policymakers, privacy regulators and global subject matter experts on an ongoing basis to inform wider releases, as we did before expanding Bard to 40 languages and international audiences.

2. Conduct adversarial testing.

Developers can stress-test generative AI models internally to identify and mitigate potential risks before launch and any ongoing releases. For example, with Bard, our experiment that lets people collaborate with generative AI, we tested for outputs that could be interpreted as person-like, which can lead to potentially harmful misunderstandings, and then created a safeguard by restricting Bard's use of "I" statements to limit risk of inappropriate anthropomorphization we discovered during testing. We also:

- Seek input from communities early in the research and development process to develop an understanding of societal contexts. This can help inform thorough stress testing. For example, we recently partnered with MLCommons and Kaggle to create Adversarial Nibbler, a public AI competition to crowdsource adversarial prompts to stress-test text-to-image models, with the goal of identifying unseen gaps, or "unknown unknowns," in how image generation models are evaluated.
- Test internally and inclusively. Before releasing Bard, we pulled from a group of hundreds of Googlers with a wide variety of backgrounds and cultural experiences, who volunteered to intentionally violate our policies to test the service. We continue to conduct these internal adversarial tests to inform Bard's ongoing expansions and feature releases.
- Adjust and apply adversarial security testing to address generative AI-specific concerns. For example, we've evolved our ongoing "red teaming" efforts — a stress-test approach

that identifies vulnerabilities to attacks – to "ethically hack" our AI systems and support our new Secure AI Framework.

- Expand our programs that reward for threat identification. Cyberthreats evolve quickly and some of the biggest vulnerabilities aren't discovered by companies or product manufacturers — but by outside security researchers. That's why we have a long history of supporting collective security through our Vulnerability Rewards Program (VRP), Project Zero and in the field of Open Source software security. It's also why we joined other leading AI companies at the White House earlier this year to commit to advancing the discovery of vulnerabilities in AI systems. This month we announced we're expanding our VRP to reward for attack scenarios specific to generative AI. We believe this will incentivize research around AI safety and security, and bring potential issues to light that will ultimately make AI safer for everyone. We're also expanding our open source security work to make information about AI supply chain security universally discoverable and verifiable.
- As part of expanding VRP for AI, we're taking a fresh look at how bugs should be categorized and reported. Generative AI raises new and different concerns than traditional digital security, such as the potential for unfair bias, model manipulation or misinterpretations of data (hallucinations). As we continue to integrate generative AI into more products and features, our Trust and Safety teams are leveraging decades of experience and taking a comprehensive approach to better anticipate and test for these potential risks. But we understand that outside security researchers can help us find, and address, novel vulnerabilities that will in turn make our generative AI products even safer and more secure. In August, we joined the White House and industry peers to enable thousands of third-party security researchers to find potential issues at DEF CON's largest-ever public Generative AI Red Team event. Now, since we are expanding the bug bounty program and releasing additional guidelines for what we'd like security researchers to hunt, we're sharing those guidelines so that anyone can see what's "in scope." We expect this will spur security researchers to submit more bugs and accelerate the goal of a safer and more secure generative AI.
- And to further protect against machine learning supply chain attacks, we're expanding our open source security work and building upon our prior collaboration with the Open Source Security Foundation. The Google Open Source Security Team (GOSST) is leveraging SLSA and Sigstore to protect the overall integrity of AI supply chains. SLSA involves a set of standards and controls to improve resiliency in supply chains, while Sigstore helps verify that software in the supply chain is what it claims to be. To get started, we recently announced the availability of the first prototypes for model signing with Sigstore and attestation verification with SLSA.

These are early steps toward ensuring the safe and secure development of generative AI — and we know the work is just getting started. Our hope is that by incentivizing more security research and conducting adversarial testing, while applying supply chain security to AI, we'll spark even more collaboration with the open source security community and others in industry, and ultimately help make AI safer for everyone.

3. Communicate simple, helpful explanations.

At launch, we seek to offer clear communication on when and how generative AI is used. We strive to show how people can offer feedback, and how they're in control. For example, for Bard, some of our explainability practices included:

- The 'Google It' button provides relevant Search queries to help users validate fact-based questions
- Thumbs up and down icons as feedback channels
- Links to report problems and offer operational support to ensure rapid response to user feedback
- User control for storing or deleting Bard activity

We also strive to be clear with users when they are engaging with a new generative AI technology in the experimental phase. For example, Labs releases such as NotebookLM are labeled prominently with "Experiment," along with specific details on what limited features are available during the early access period.

Another explainability practice is thorough documentation on how the generative AI service or product works. For Bard, this included a comprehensive overview offering clarity on the cap on the number of interactions to ensure quality, accuracy and prevent potential personification and other details on safety, and a privacy notice to help users understand how Bard handles their data.

Maintaining transparency is also key. We released a detailed technical report on PaLM 2, the model currently powering Bard, which includes information based on our internal documentation of evaluation details, and guidance for AI researchers and developers on the responsible use of the model.

In addition to the three observations above, we're broadly focused on ensuring that new generative AI technologies have equally innovative guardrails when addressing concerns such as image provenance. Our efforts include watermarking images Google AI tools generate (as in Virtual Try On or Da Vinci Stickies) and offering image markups for publishers to indicate when an image is AI generated.

Google DeepMind has just launched SynthID, an experimental tool for watermarking and identifying AI-generated images. This technology embeds a digital watermark directly into the pixels of an image, making it imperceptible to the human eye, but detectable for identification.

While generative AI can unlock huge creative potential, it also presents new risks, like the spreading of false information – both intentionally and unintentionally. SynthID represents a significant research effort, in line with the voluntary commitments made by leading AI companies to the White House earlier this year. SynthID is being released to a limited number of Google

Cloud Vertex AI customers using Imagen, one of our latest text-to-image models that uses input text to create photorealistic images. There are two elements to the system:

- Watermarking: produces a watermark designed to be imperceptible to the human eye.
- Verification: determines whether an image is generated by Imagen vis a vis a confidence interval.

While this combined approach is not infallible, our internal testing shows it is accurate - even when an image undergoes various common image manipulations. Being able to identify AI-generated content is critical to empowering people with knowledge of when they're interacting with generated media, and helping prevent the spread of misinformation.

Collaboration

Finally, when it comes to the responsibility, safety and security of high impact AI, we know we can't do this work alone. That's why we created a $20M Digital Futures Fund through Google.org to spark more research and discussion on AI safety, security, and responsibility. In addition, together with Anthropic, Microsoft, and OpenAI, we announced our first Executive Director of the Frontier Model Forum, and the creation of a new AI Safety Fund, a more than $10 million initiative to promote research in the field of AI safety. The Frontier Model Forum, an industry body focused on ensuring safe and responsible development of frontier AI models, is also releasing its first technical working group update on red teaming to share industry expertise with a wider audience as the Forum expands the conversation about responsible AI governance approaches.

We also kicked off a public discussion inviting web publishers, civil society, academia and AI communities to offer thoughts on approaches to protocols to support the future development of the Internet in the age of generative AI. As we move ahead, we will continue to share how we apply emerging practices for responsible generative AI development and ongoing transparency with our annual, year-end AI Principles Progress Update.

Forums like this and the work the White House is doing will enable meaningful progress, unlock more opportunity for all Americans, and ensure continued U.S. leadership. We can all be bold and responsible in promoting the acceptance, adoption and helpfulness of new high impact technologies. I'm delighted to participate in this forum and answer any questions you might have about our approach. Thank you.