

From: [REDACTED]
To: [REDACTED]
Subject: FW: Request for Amazon statement on AI voluntary commitments before January 16th
Date: Tuesday, January 28, 2025 3:53:32 PM
Attachments: [image002.png](#)

Amazon response below.

From: Schrantz, Ellen [REDACTED]
Sent: Tuesday, January 28, 2025 3:51 PM
To: [REDACTED]; Layman, Heather [REDACTED]
Cc: Morrison, Simon [REDACTED]
Subject: RE: Request for Amazon statement on AI voluntary commitments before January 16th

Hi Abdiaziz:

Thanks for your patience.

Amazon's approach to responsible AI has [eight dimensions](#): fairness, explainability, privacy and security, safety, controllability, veracity and robustness, governance, and transparency. Security is a priority across our company, both in our consumer products and services and in the tools and resources we offer customers to help them achieve their AI objectives. For example, our recently announced [Amazon Nova](#) models, as well as other models available through Amazon Bedrock, come with best-in-class security that enables customers to build generative AI applications that support common data security and compliance standards. Customers can use Amazon Web Services (AWS) PrivateLink to establish private connectivity between customized Amazon Nova and on-premise networks without exposing customer traffic to the internet. Customer data is always encrypted in transit and at rest, and customers can use their own keys to encrypt the data. Last year, AWS became the first major cloud service provider to announce that we had received ISO/IEC 42001-accredited certification for our AI services, demonstrating our commitment to AI risk management.

Amazon joined other AI industry leaders in agreeing to [voluntary commitments](#) around the responsible development of AI. Since that time, we continue to invest in AI safety and security in line with those commitments, including through red-teaming, industry partnerships and collaboration, and our work on content provenance. Amazon is engaged in state-of-the-art red teaming and testing of our most advanced foundation models to anticipate and mitigate risk. In building our Amazon Nova models, we did extensive in-house testing across a range of threat domains. We also worked with other companies to develop red-teaming methods that leveraged their specific areas of

expertise, such as chemical, biological, radiological, and nuclear risks and model deception capabilities.

Additionally, we incentivize the research community to help test our systems. Amazon encourages reputable researchers to find and [report security vulnerabilities](#), including in our AI technologies. Amazon works through a range of third-party organizations to help develop and socialize industry best-practice on responsible AI development and deployment. We are a member of the [Frontier Model Forum](#), an organization dedicated to the advancement of the science, standards, and best practices of AI safety and security. We are also a participant in the [Partnership on AI](#), which leads multistakeholder efforts to do research and develop industry responsible AI best practices. We have also participated in the Cybersecurity and Infrastructure Security Agency's initiative on AI to collectively address the unique challenges of AI security, including identifying ways to enhance collaboration and information sharing between the government, the private sector, international partners, and other stakeholders.

Amazon is a leader in building and deploying disclosure and content provenance technologies for synthetic media to enable consumers to identify when and how a piece of content has been modified or created by AI. The Amazon Nova models include two multimodal generative-AI models: Amazon Nova Canvas, which generates static images, and Amazon Nova Reel, which generates video. To promote the traceability of AI-generated content, we incorporate tamper-resistant invisible watermarks directly into the image and video generation processes and, for Canvas, add metadata developed by the Coalition for Content Provenance and Authenticity, where Amazon is a steering committee member.

From: [REDACTED]
Sent: Friday, January 24, 2025 10:30 AM
To: Layman, Heather [REDACTED]
Cc: Schrantz, Ellen [REDACTED] Morrison, Simon [REDACTED]
Subject: RE: [EXTERNAL] Request for Amazon statement on AI voluntary commitments before January 16th

CAUTION: This email originated from outside of the organization. Do not click links or open attachments unless you can confirm the sender and know the content is safe.

Thanks Heather. We are working on finalizing the report and so anything you can share would be greatly appreciated.

From: Layman, Heather [REDACTED]

Sent: Thursday, January 23, 2025 5:23 PM

To: [REDACTED]

Cc: Schrantz, Ellen [REDACTED] Morrison, Simon [REDACTED]

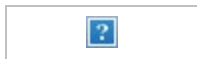
Subject: FW: Request for Amazon statement on AI voluntary commitments before January 16th

Hello Abdiaziz,

I'm filling in for my colleague Betsy Hart at Amazon while she's out on leave. I believe she is who replied to you originally about your request for our input to the report on progress on voluntary AI commitments.

We're working on this, but unfortunately due to multiple unexpected absences we are still waiting on information. We will get back to you with responses no later than Tuesday of next week.

Thank you,
Heather



Heather Layman

Senior Manager, Policy Communications

[REDACTED]

[REDACTED]

Washington, DC

For the latest news, check out [About Amazon](#) and follow us on Twitter at [@amazonnews](#) and [@amazon_policy](#)

From: [REDACTED]

Sent: Wednesday, December 18, 2024 12:14 PM

To: amazon-pr [REDACTED]

Subject: Request for Amazon statement on AI voluntary commitments before January 16th

To Whom It May Concern at Amazon,

In July 2023 and February 2024, respectively, your company became a signatory of the White House Voluntary AI Commitments and the Munich Security Conference AI Election Accord.

We are contacting you today to request a statement on your company's progress toward

meeting the voluntary AI commitments outlined in these agreements.

In February 2025, the Brennan Center for Justice at NYU will publish a report analyzing the progress that all signatory companies have made toward meeting these commitments. We have collected information on your company's progress through the resources provided as part of the following:

White House Voluntary AI Commitments

- Initial press release ([link](#))

Munich Security Conference AI Election Accord

- Written response to May 2024 Senator Mark R. Warner's letter request ([link](#))

To help us understand your company's efforts to meet these commitments, please provide a written statement outlining the specific actions your company has taken to align with them, including any updates or new developments beyond the information provided in the links above. If available, please also include links to additional publicly available resources that demonstrate your progress.

We will be collecting responses from the signatory companies **before the end of Monday, January 16th**. It would benefit our analysis if you could observe the numbered commitments made in each agreement and speak on your company's specific efforts toward each numbered commitment.

We will be completing our analysis and evaluation of fulfillment based on our review of the information provided in your written responses listed above and any responses provided to this email.

Best,

Abdiaziz Ahmed

Technology Policy Strategist

Brennan Center for Justice

From: [REDACTED]
To: [REDACTED]
Subject: Fw: Request for Arm statement on AI voluntary commitments before January 16th
Date: Thursday, January 16, 2025 7:59:12 PM

Get [Outlook for iOS](#)

From: Kristen Ray [REDACTED]
Sent: Thursday, January 16, 2025 7:48:56 PM
To: [REDACTED]
Cc: Vince Jesaitis [REDACTED]
Subject: Re: Request for Arm statement on AI voluntary commitments before January 16th

Hi;

Please see Arm's statement below. Let us know if you have any questions.

Thanks,
-k

//

“Arm continues to work with C2PA to ensure standards are developed and widely utilized to demonstrate the provenance of digital content. This critical work applies state of the art security technology to enable cryptographically protected, traceability of content. Since we initially signed the Munich Security Conference AI Election Accord, we’ve made continued progress in enabling the hardware and compute technology required to establish the provenance of digital content, and we remain committed to partnering with the wider coalition and Arm ecosystem to drive ongoing progress. Our efforts are focused on ensuring the Arm platform – widely used across devices and industries – provides robust security foundations that enable consumers to reliably assess content.”

Kristen Ray | Senior Director, Corporate Communications, Arm
[REDACTED]

From: [REDACTED]
Date: Thursday, January 9, 2025 at 12:00 PM
To: Kristen Ray [REDACTED]
Subject: Re: Request for Arm statement on AI voluntary commitments before January 16th

Warning: EXTERNAL SENDER, use caution when opening links or attachments.

I'm sorry for that mistake - it's January 16.

Get [Outlook for iOS](#)

From: Kristen Ray [REDACTED]
Sent: Thursday, January 9, 2025 12:54:47 PM
To: [REDACTED]
Subject: Re: Request for Arm statement on AI voluntary commitments before January 16th

Hi there;

Can I please confirm that the deadline for this is 16-Jan, or Monday Jan. 13th?

Below it says by end of Monday 16-Jan.

Thanks,
-k

Kristen Ray | Senior Director, Corporate Communications, Arm
[REDACTED]

From: [REDACTED]
Date: Wednesday, December 18, 2024 at 12:16 PM
To: Global-PRteam [REDACTED]
Subject: Request for Arm statement on AI voluntary commitments before January 16th

Warning: EXTERNAL SENDER, use caution when opening links or attachments.

To Whom It May Concern at Arm,

In February 2024, your company became a signatory of the Munich Security Conference AI Election Accord.

We are contacting you today to request a statement on your company's progress toward meeting the voluntary AI commitments outlined in this agreement.

In February 2025, the Brennan Center for Justice at NYU will publish a report analyzing the progress that all signatory companies have made toward meeting these

commitments. We have collected information on your company's progress through the resources provided as part of the following:

Munich Security Conference AI Election Accord

- Written response to May 2024 Senator Mark R. Warner's letter request ([link](#))

To help us understand your company's efforts to meet these commitments, please provide a written statement outlining the specific actions your company has taken to align with them, including any updates or new developments beyond the information provided in the links above. If available, please also include links to additional publicly available resources that demonstrate your progress.

We will be collecting responses from the signatory companies **before the end of Monday, January 16th**. It would benefit our analysis if you could observe the numbered commitments made in each agreement and speak on your company's specific efforts toward each numbered commitment.

We will be completing our analysis and evaluation of fulfillment based on our review of the information provided in your written responses listed above and any responses provided to this email.

Best,

Abdiaziz Ahmed

Technology Policy Strategist

Brennan Center for Justice

IMPORTANT NOTICE: The contents of this email and any attachments are confidential and may also be privileged. If you are not the intended recipient, please notify the sender immediately and do not disclose the contents to any other person, use it for any purpose, or store or copy the information in any medium. Thank you.

IMPORTANT NOTICE: The contents of this email and any attachments are confidential and may also be privileged. If you are not the intended recipient, please notify the sender immediately and do not disclose the contents to any other person, use it for any purpose, or store or copy the information in any medium. Thank you.

From: [REDACTED]
To: [REDACTED]
Subject: Fw: Request for ElevenLabs statement on AI voluntary commitments before January 16th
Date: Friday, December 20, 2024 8:23:15 AM

See below for ElevenLabs response.

Get [Outlook for iOS](#)

From: Alex Haskell [REDACTED]
Sent: Wednesday, December 18, 2024 6:10 PM
To: [REDACTED]
Subject: Re: Request for ElevenLabs statement on AI voluntary commitments before January 16th

Hi Abdiaziz - Thanks very much for your message. In addition to the two resources linked in your email, in order to understand our efforts to meet the Election Accord's commitments, please see our [October 29th update on our preparation for elections](#). We appreciate your work on this important project!

Best,
Alex

On Wed, Dec 18, 2024 at 3:17 PM [REDACTED] wrote:

To Whom It May Concern at ElevenLabs,

In February 2024, your company became a signatory of the Munich Security Conference AI Election Accord.

We are contacting you today to request a statement on your company's progress toward meeting the voluntary AI commitments outlined in this agreement.

In February 2025, the Brennan Center for Justice at NYU will publish a report analyzing the progress that all signatory companies have made toward meeting these commitments. We have collected information on your company's progress through the resources provided as part of the following:

Munich Security Conference AI Election Accord

- Written response to May 2024 Senator Mark R. Warner's letter request ([link](#))
- Official AI Election Accord progress update ([link](#))

To help us understand your company's efforts to meet these commitments, please provide a written statement outlining the specific actions your company has taken to align with them, including any updates or new developments beyond the information provided in the links above. If available, please also include links to additional publicly available resources that demonstrate your progress.

We will be collecting responses from the signatory companies **before the end of Monday, January 16th**. It would benefit our analysis if you could observe the numbered commitments made in each agreement and speak on your company's specific efforts toward each numbered commitment.

We will be completing our analysis and evaluation of fulfillment based on our review of the information provided in your written responses listed above and any responses provided to this email.

Best,

Abdiaziz Ahmed

Technology Policy Strategist

Brennan Center for Justice

--

From: [REDACTED]
To: [REDACTED]
Subject: Fw: [EXT] Request for Gen statement on AI voluntary commitments before January 16th
Date: Monday, January 20, 2025 8:43:12 AM

Get [Outlook for iOS](#)

From: Barbora Petrova (CS) [REDACTED]
Sent: Monday, January 20, 2025 8:41:46 AM
To: [REDACTED]
Cc: Kim Allman [REDACTED]; Kirsten McMullen [REDACTED]
Subject: Re: [EXT] Request for Gen statement on AI voluntary commitments before January 16th

Dear Abdiaziz Ahmet,

Thanks for reaching our to us!

Last year, we have published several blogs and reports regarding our own internal AI policies and regarding our general recommendations for other companies, policy makers and individuals as well as an extensive study on the current state of AI-powered cyber threats.

All the texts can be found via this blogpost:

<https://www.gendigital.com/blog/news/company-news/ai-policy-2024>

Feel free to reach our should you need more information,

All the best,
Barbora Petrova



Gen Blogs | AI at Gen: Unpacking What AI Means for People, Cybersecurity and Our Company

All over the world, people are embracing artificial intelligence (AI) in the ways they work, learn and communicate. A recent report from McKinsey & Company found that companies across industries, regions and sizes are rapidly increasing their use of generative AI (gen AI) to

From: [REDACTED]

Date: Wednesday, December 18, 2024 at 3:19 PM

To: Kim Allman [REDACTED]

Subject: [EXT] Request for Gen statement on AI voluntary commitments before January 16th

To Kim Allman, Head of Corporate Responsibility, ESG & Government Affairs at Gen,

In February 2024, your company became a signatory of the Munich Security Conference AI Election Accord.

We are contacting you today to request a statement on your company's progress toward meeting the voluntary AI commitments outlined in this agreement.

In February 2025, the Brennan Center for Justice at NYU will publish a report analyzing the progress that all signatory companies have made toward meeting these commitments. We have collected information on your company's progress through the resources provided as part of the following:

Munich Security Conference AI Election Accord

- Written response to May 2024 Senator Mark R. Warner's letter request (no materials found).

To help us understand your company's efforts to meet these commitments, please provide a written statement outlining the specific actions your company has taken to align with them, including any updates or new developments beyond the information provided in the links above. If available, please also include links to additional publicly available resources that demonstrate your progress.

We will be collecting responses from the signatory companies **before the end of Monday, January 16th**. It would benefit our analysis if you could observe the numbered commitments made in each agreement and speak on your company's specific efforts toward each numbered commitment.

We will be completing our analysis and evaluation of fulfillment based on our review of the information provided in your written responses listed above and any responses provided to this email

Best,

Abdiaziz Ahmed

Technology Policy Strategist

Brennan Center for Justice

From: [REDACTED]
To: [REDACTED]
Subject: Fw: Request for Google statement on AI voluntary commitments before January 16th
Date: Friday, January 17, 2025 3:54:46 PM

Get [Outlook for iOS](#)

From: Michael Appel [REDACTED]
Sent: Friday, January 17, 2025 3:40:46 PM
To: [REDACTED]
Subject: Fwd: Request for Google statement on AI voluntary commitments before January 16th

Hi Abdiaziz,

Thanks for reaching out. Sorry for the delay. Here are some updates.

First, for the Munich Security Conference AI Election Accord, looks like you already have our latest update: <https://www.aielectionsaccord.com/progress-update/>.

Second, for the White House Voluntary AI Commitments, here's what we can share:

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. In the fall of 2023, we shared a [comprehensive update](#). Since then, we've continued to build on our work. Here are some highlights:

SAFETY

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.

- Together with the founding members of the [Frontier Model Forum](#) and philanthropic partners, we created a new [AI Safety Fund](#) to promote research in the field of AI safety. The fund was initially set up with \$10 million in funding and it will support independent researchers from around the world (academic institutions, research institutions, startups, etc.). The goal of the fund is to allow researchers to better evaluate and understand frontier systems, and – ultimately – develop new model evaluations and techniques for red teaming AI models to help develop and test evaluation techniques for potentially dangerous capabilities of frontier systems.
-

For elections, we conduct ongoing red teaming to test the boundaries of the election related queries that will return responses in Gemini, and we have restricted the types of election-related queries for which Gemini will return responses ([EU blog](#) and [US blog](#)).

- We hold ongoing red teaming for our frontier models on key areas, including misuse related to elections, societal risks, and national security concerns.
- Launched the Content Adversarial Red Team, which is the industry's first red team specifically focused on content-related 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.

- Actively working with governments across the globe by contributing tools and resources to projects like the National Science Foundation's [National AI Research Resource pilot](#), which aims to democratize AI research across the U.S.
- Signed onto the [Seoul Frontier AI Safety Commitments](#) alongside 15 other leading AI developers. The pledge builds on previous agreements from the Bletchley Summit and commits the companies to ensuring accountable governance structures and public transparency on their approaches to frontier AI safety.
- We contributed – with industry peers, academics, governments and civil society organizations – to the Partnership on AI's [Guidance for Safe Foundation Model Deployment: A Framework for Collective Action](#).
- As a member of the multi-stakeholder organization MLCommons, we contributed to the proof of concept for a cross-industry [AI Safety Benchmark](#).
- We helped launch [The Adversarial Nibbler Challenge](#), a data-centric AI competition to engage the research community in jointly identifying current blind spots in harmful image production. This competition is the result of collaboration

between six different organizations to jointly produce a shared resource for use and reuse by the wider research and development community.

SECURITY

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

- Our models are developed, trained, and stored within Google's infrastructure, where we can apply our [Secure AI Framework \(SAIF\)](#) throughout the AI responsibility lifecycle, which is supported by central security teams and by a security, safety and reliability organization consisting of engineers and researchers with world-class expertise.
- We embrace an open and collaborative approach to cybersecurity, combining cyber threat intelligence and working with partners to contribute to AI research and provide developer tools so everyone can build AI more responsibly, such as the [AI Cyber Defense Initiative](#).
- On July 18, 2024, we [launched](#) the Coalition for Secure AI (CoSAI) with industry partners. This is the first major milestone and application of Google's Secure AI Framework (SAIF). The coalition will collectively invest in AI security research, share security expertise and best practices, and build technical open source solutions. CoSAI is an open source initiative designed to give all practitioners and developers the guidance and tools they need to create AI systems that are Secure-by-Design. The coalition will operate under the guidance of OASIS Open, the international standards and open source consortium. Founding members include: Amazon, Anthropic, Cisco, Cohere, GenLab, Google, IBM, Intel, Microsoft, Nvidia, Open AI, Paypal and Wiz.

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

- Expanding more [Vulnerability Rewards Programs](#) to rely on third-party review and assistance in identifying vulnerabilities in our AI systems. Published a [2023 year in review](#) of the VRPs and their successes
- Held multiple outreach events to engage with the security research community.

This includes working with the researcher community to incentivize research into specific Gen AI products that have higher risk of abuse and working with internal policy teams to work towards consistent, researcher-friendly safe harbor policies and usage guidelines

TRUST

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 improved [SynthID](#), our industry-leading image and audio watermarking tool, to also be able to identify text and video generated by the Gemini app and web experience. These technical solutions are still nascent and case-specific, but represent a step toward offering scalable solutions for researchers and others to identify synthetic AI-generated content.
- We joined the [C2PA](#) as a steering committee member. The coalition is a cross-industry effort to provide more transparency and context for people on digital content. Google is helping to develop its technical standard and further adoption of Content Credentials, tamper-evident metadata, which shows how content was made and edited over time.
- Google and YouTube announced a [A Tech Accord to Combat Deceptive Use of AI in 2024 Elections](#) during the Munich Security Conference, along with 20 leading technology companies including Microsoft, Meta, OpenAI, Adobe and more. This accord seeks to set expectations for how signatories will manage the risks arising from deceptive AI election content created through their publicly accessible, large-scale platforms or open foundational models, or distributed on their large-scale social or publishing platforms in line with their own policies and practices as relevant to the commitments in the accord.
- We updated our [election advertising policies](#) to require advertisers to disclose when their election ads include material that's been digitally altered or generated. Similarly, we require YouTube creators to [disclose](#) when they've created realistic altered or synthetic content, and will display a label that indicates for people when the content they're watching is synthetic.

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.

- We updated our modelcards.withgoogle.com site to include examples of different types of AI model transparency artifacts, including for generative models. We share guidance for the AI ecosystem for composing these public reports. Some of these examples include model cards and technical reports for our most advanced Generative AI models, such as [Gemini 1.5](#), "Gemini Advanced," and [Gemma 2.0](#), as well as artifacts for other versions of models, such as the Gemini 1.0 [Model Card](#) and [Technical Report](#).
- We regularly publish reports to demonstrate the progress we are making and lessons we have learned along the way relating to model and system capabilities and limitations. Recently we published a new [mid-year report](#) outlining our end-to-end approach to responsible model and product development, including our AI Responsibility Lifecycle. This report offers transparency into our internal AI governance and risk assessments and mitigations and their outputs.

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

- We've continued our commitment to sharing our research on AI's benefits and risks, including through several papers focused on the societal risks posed by AI, including [holistic safety evaluations](#), a [study](#) of misuses of generative AI, and [ethical concerns](#) and [a context-based framework](#) for comprehensively evaluating the social and ethical risks of AI systems.
- We also made updates to our [Responsible AI Toolkit](#), providing developers with new and enhanced tools for evaluating model safety and filtering harmful content and presented our 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.
- We are committed to sharing ongoing [research on responsible AI practices](#) and the societal benefits and emerging risks of [advanced AI](#).
-

We've [shared Google DeepMind's evolving approach](#) to developing a holistic set of safety and responsibility evaluations for our advanced models, including [early research](#) evaluating critical capabilities such as deception, cyber-security, self-proliferation, and self-reasoning. We also released an in-depth exploration into [aligning future advanced AI assistants](#) with human values and interests. Beyond LLMs, we recently shared our approach to [biosecurity](#) for [AlphaFold 3](#).

- We launched the exploratory [Frontier Safety Framework](#) – a set of protocols for proactively identifying future AI capabilities that could cause severe harm and putting in place mechanisms to detect and mitigate them. Our Framework focuses on severe risks resulting from powerful capabilities at the model level, such as exceptional agency or sophisticated cyber capabilities. It is designed to complement our alignment research, which trains models to act in accordance with human values and societal goals, and Google's existing suite of AI responsibility and safety [practices](#).
- To help promote diverse perspectives in multi-disciplinary AI research to help address society-level, shared challenges from forecasting hunger to predicting diseases to improving productivity, we announced that 70 professors were selected for the [Award for Inclusion Research Program](#), which supports academic research that addresses the needs of historically marginalized groups globally.

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

- We've continued to deploy new frontier AI systems with the potential to revolutionize how society seeks to address challenges across health, science, and the environment. As we've developed [these AI systems](#), we've also helped build the frameworks to guide their safe and effective deployment.

Beyond frontier models, we continue to research and develop AI systems to address humanity's needs in health, sustainability, and fundamental scientific knowledge. Here are some other examples of how we're helping address society's greatest challenges:

- [AlphaFold 3](#) – a revolutionary model that can predict the structure and interactions of all life's molecules with unprecedented accuracy. Scientists can access the majority of its capabilities, for free, through our newly launched

AlphaFold Server, an easy-to-use research tool.

- [AlphaGeometry](#) – can solve geometry problems at a level comparable with the world’s brightest high-school mathematicians
- [AlphaMissense](#) – analyzes benign and disease-causing ‘missense’ genetic mutations
- [GNoME](#) – deep learning tool that has helped researchers discover 2.2 million new crystals
- [GraphCast](#) – state-of-the-art AI model able to make medium-range weather forecasts with unprecedented accuracy
- Developed over 10 years of connectomics research and in partnership with the Lichtman Lab (Harvard) and others, Google [published](#) a mapping of a piece of the human brain to a level of detail never previously achieved. This project revealed never-before-seen structures within the human brain and the full dataset, including AI-generated annotations for each cell, has been made publicly available via the online platform Neuroglancer. This work may change our understanding of how the brain works which could help researchers better understand neurological diseases such as Alzheimers and also answer fundamental questions (eg. how memories form).
- Google was named [Fast Company's World Changing Company of the Year 2024](#) for our work on using AI to address climate change, including [Project Green Light](#), [Flood Hub](#), [Contrails](#), and more as we continue the work to help find ways to reduce humanity’s climate impact.

----- Forwarded message -----

From: [REDACTED]
Date: Thu, Dec 19, 2024 at 4:22 AM
Subject: Request for Google statement on AI voluntary commitments before January 16th
To: [REDACTED]

To Whom It May Concern at Google,

In July 2023 and February 2024, respectively, your company became a signatory of the White House Voluntary AI Commitments and the Munich Security Conference AI Election Accord.

We are contacting you today to request a statement on your company's progress toward meeting the voluntary AI commitments outlined in these agreements.

In February 2025, the Brennan Center for Justice at NYU will publish a report analyzing the progress that all signatory companies have made toward meeting these commitments. We have collected information on your company's progress through the resources provided as part of the following:

White House Voluntary AI Commitments

- Initial statement ([link](#))

Munich Security Conference AI Election Accord

- Written response to May 2024 Senator Mark R. Warner's letter request ([link](#))
- Official AI Election Accord progress update ([link](#))

To help us understand your company's efforts to meet these commitments, please provide a written statement outlining the specific actions your company has taken to align with them, including any updates or new developments beyond the information provided in the links above. If available, please also include links to additional publicly available resources that demonstrate your progress.

We will be collecting responses from the signatory companies **before the end of Monday, January 16th**. It would benefit our analysis if you could observe the numbered commitments made in each agreement and speak on your company's specific efforts toward each numbered commitment.

We will be completing our analysis and evaluation of fulfillment based on our review of the information provided in your written responses listed above and any responses provided to this email.

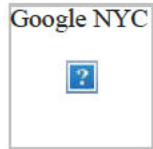
Best,

Abdiaziz Ahmed

Technology Policy Strategist

Brennan Center for Justice

--



Michael Appel (he/him)

Global Communications & Public Affairs

New York, NY | [REDACTED]

From: [REDACTED]
To: [REDACTED]
Subject: Fw: Request for Inflection statement on AI voluntary commitments before January 16th
Date: Thursday, January 16, 2025 1:12:11 PM

See below for Inflection AI statement

Get [Outlook for iOS](#)

From: Bailey Ghashghai [REDACTED]
Sent: Thursday, January 16, 2025 1:10 PM
To: Abdiaziz Ahmed [REDACTED]
Subject: Re: Request for Inflection statement on AI voluntary commitments before January 16th

Hi Abdiaziz,

Here are the commitments and the efforts made by Inflection AI to date.

Thanks,
Bailey

++

1. The companies commit to **internal and external security testing of their AI systems before their release**. This testing, which will be carried out in part by independent experts, guards against some of the most significant sources of AI risks, such as biosecurity and cybersecurity, as well as its broader societal effects.

- **Internal Testing & Benchmarking:**
We maintain a rigorous internal testing protocol that evaluates both system- and application-level security. This includes continuous expansion of our safety benchmarks—where we measure the AI systems' performance against various threat vectors, potential misuse scenarios, and potential biases. Our internal red-teaming exercises and dedicated AI safety teams help us identify vulnerabilities before products are released.
- **Independent Expert Assessment:**
Although our bug bounty program initially focused on traditional systems and platform vulnerabilities, we are exploring ways to involve independent expert teams to evaluate our AI models. This may include contracting with third-party security researchers and academic groups to stress-test our models for cybersecurity and societal impacts (e.g., misinformation, disinformation, bias).
- **Continuous Improvement:**
Feedback from internal testing, independent audits, and community-driven bug bounty submissions inform iterative updates to our AI safety protocols. These updates strengthen our practices over time, ensuring we keep pace with evolving risks (e.g., new adversarial attack techniques).

2. The companies commit to **sharing information across the industry and with governments, civil society, and academia on managing AI risks**. This includes best practices for safety, information on attempts to circumvent safeguards, and technical collaboration.

- **Open-Sourced Safety Benchmarks:**
By publicly releasing our internally developed safety benchmark dataset, we enable broader collaboration on best practices for AI safety. Researchers, developers, and civil society organizations can use this dataset to refine their own models, identify emerging risks, and develop mitigation strategies.
- **Collaborative Forums & Workshops:**
We regularly host and participate in workshops, webinars, and panel discussions with government, academia, and non-profit organizations. These sessions provide a platform to share experiences on AI risk management, including strategies for addressing attempts to circumvent safeguards, best practices for bias mitigation, and robust security protocols.
- **Multi-Stakeholder Engagement:**
We also invite leading experts from diverse disciplines—technology, ethics, law, and policy—to speak with our teams, ensuring we maintain a broad perspective on how AI risks evolve. These conversations feed directly into our internal policymaking and technical development processes.

3. The companies commit to **investing in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights**. These model weights are the most essential part of an AI system, and the companies agree that it is vital that the model weights be released only when intended and when security risks are considered.

- **Prioritizing Model Weights Security:**
Our model weights are encrypted both at rest and in transit, limiting unauthorized access. Only authorized personnel—who undergo additional background checks and are subject to strict access controls—can handle this highly sensitive data.
- **Comprehensive Bug Bounty Scope:**

Our bug bounty program emphasizes the protection of proprietary information, including AI model architectures and weights. Researchers are incentivized to identify any theoretical or actual pathways for data exfiltration, ensuring vulnerabilities are quickly discovered and remediated.

4. The companies commit to **facilitating third-party discovery and reporting of vulnerabilities in their AI systems**. Some issues may persist even after an AI system is released and a robust reporting mechanism enables them to be found and fixed quickly.

- **Active Engagement with Security Researchers:**

We maintain an ongoing partnership with HackerOne, which broadens our vulnerability detection surface to a global community of ethical hackers and security researchers. This engagement helps us discover and address issues that may not surface through internal testing alone.

- **Transparent Reporting & Rapid Mitigation:**

We have a clear process for third-party vulnerability disclosure, ensuring that reports are triaged, verified, and remediated swiftly. Reporters are also kept informed during the process, enhancing trust and encouraging further community collaboration.

- **Scope Inclusions:**

Our scope explicitly includes AI-related features and services to ensure that potential vulnerabilities in generative AI or other machine learning components are not overlooked.

5. The companies commit to **developing robust technical mechanisms to ensure that users know when content is AI generated, such as a watermarking system**. This action enables creativity with AI to flourish but reduces the dangers of fraud and deception.

- **Exploring Watermarking & Metadata Strategies:**

While we do not yet have a formal watermarking system in place, we are actively researching and evaluating various methods—such as data embedding, cryptographic watermarking, or metadata tagging—to help users and external parties identify AI-generated content.

- **Industry Collaboration:**

We participate in discussions with industry consortia and standards bodies exploring watermarking techniques. The goal is to adopt or develop an approach that is robust, verifiable, and minimally disruptive to creative expression and data utility.

6. The companies commit to **publicly reporting their AI systems' capabilities, limitations, and areas of appropriate and inappropriate use**. This report will cover both security risks and societal risks, such as the effects on fairness and bias.

- **Regular Transparency Updates:**

Beyond our Terms of Service, we plan to provide regular public reports and technical papers detailing our systems' capabilities, limitations, and intended use cases. These documents will clarify how we address security, fairness, and potential misuse (e.g., disinformation, hate speech).

- **Security & Societal Risk Assessments:**

We integrate findings from internal and external testing into these reports, enabling stakeholders to understand potential threats and the steps we are taking to mitigate them. Topics include privacy considerations, bias, model drift, and real-world performance gaps.

- **User Guidelines & Best Practices:**

Alongside describing system capabilities, we also offer user guidelines that outline recommended and prohibited uses, helping organizations and individuals deploy our AI responsibly.

7. The companies commit to **prioritizing research on the societal risks that AI systems can pose, including on avoiding harmful bias and discrimination, and protecting privacy**. The track record of AI shows the insidiousness and prevalence of these dangers, and the companies commit to rolling out AI that mitigates them.

- **Dedicated Research & Development:**

Our team includes ML researchers who work to identify and reduce harmful bias. This includes data collection and annotation strategies that promote representativeness and fairness, as well as model fine-tuning protocols aimed at mitigating discriminatory outputs.

- **Model Validation and Ongoing Review:**

We regularly validate our models against real-world use cases involving marginalized or protected groups, ensuring that updates or new deployments do not inadvertently introduce bias. We also incorporate user feedback, flags, and complaints as triggers for further review and potential retraining.

- **Partnerships with Civil Society:**

We frequently collaborate with advocacy groups, NGOs, and academic institutions to ensure our research and models align with broader societal values and legal frameworks. These partnerships help us stay current with evolving cultural and legal norms concerning fairness and discrimination.

8. The companies commit to **develop and deploy advanced AI systems to help address society's greatest challenges**. From cancer prevention to mitigating climate change to so much in between, AI—if properly managed—can contribute enormously to the prosperity, equality, and security of

all.

- **Responsible Innovation Framework:**

Before committing significant resources to large-scale societal projects, we evaluate potential risks, ethical considerations, and stakeholder input. We strive to ensure that every solution adheres to rigorous standards of privacy, fairness, and transparency.

- **Impact Assessment & Measurement:**

We aim to adopt frameworks (such as Social Return on Investment or similar methodologies) to measure the positive impact of our AI initiatives. By setting clear metrics and sharing outcomes publicly, we will foster trust and accountability.

On Wed, Dec 18, 2024 at 12:26 PM [REDACTED] wrote:

To Whom It May Concern at Inflection,

In July 2023 and February 2024, respectively, your company became a signatory of the White House Voluntary AI Commitments and the Munich Security Conference AI Election Accord.

We are contacting you today to request a statement on your company's progress toward meeting the voluntary AI commitments outlined in these agreements.

In February 2025, the Brennan Center for Justice at NYU will publish a report analyzing the progress that all signatory companies have made toward meeting these commitments. We have collected information on your company's progress through the resources provided as part of the following:

White House Voluntary AI Commitments

- No materials found

Munich Security Conference AI Election Accord

- No materials found

To help us understand your company's efforts to meet these commitments, please provide a written statement outlining the specific actions your company has taken to align with them, including any updates or new developments. If available, please also include links to additional publicly available resources that demonstrate your progress.

We will be collecting responses from the signatory companies **before the end of Monday, January 16th**. It would benefit our analysis if you could observe the numbered commitments made in each agreement and speak on your company's specific efforts toward each numbered commitment.

We will be completing our analysis and evaluation of fulfillment based on our review of the information provided in your written responses listed above and any responses provided to this email.

Abdiaziz Ahmed

Technology Policy Strategist

Brennan Center for Justice

--

You received this message because you are subscribed to the Google Groups "press" group.

To unsubscribe from this group and stop receiving emails from it, send an email to [REDACTED]

To view this discussion visit

<https://groups.google.com/a/inflection.ai/d/msgid/press/MW4PR15MB4476A2642E92ECE00FAFE40FCC6052%40MW4PR15MB4476.namprd15.prod.outlook.com>.

From: [REDACTED]
To: [REDACTED]
Subject: Fw: Request for LG AI Research statement on AI voluntary commitments before January 16th
Date: Tuesday, January 14, 2025 9:35:30 AM

Our first substantive response - see below.

Get [Outlook for iOS](#)

From: 김명신/님/(mskim) [REDACTED]
Sent: Monday, January 13, 2025 8:31 PM
To: [REDACTED]
Cc: 김유철/Unit장/(youchul.kim) [REDACTED]; 안소영/님/(soyoung.an)
[REDACTED]
Subject: RE: Request for LG AI Research statement on AI voluntary commitments before January 16th

Dear Mr. Abdiaziz Ahmed,

We are pleased to share updates on the initiatives undertaken by LG AI Research since May 2024 to further the objectives of the Tech Accord.

Public Education Initiatives

LG AI Research is dedicated to enhancing public understanding of AI and expanding access to high-quality AI education through tailored programs for various age groups:

- LG Discovery Lab: Focused on middle and high school students, offering interactive programs.
- LG AIMERS: Designed for university students, combining theoretical and practical training.
- LG AI Academy: Provides working professionals with advanced, hands-on AI education.

These programs feature comprehensive curriculums that encompass both AI technology and ethics. Annually, they are offered free of charge to over 42,000 participants, ensuring wide accessibility and contributing to the dissemination of high-quality AI education.

Research on Detection Tools and Methods

LG AI Research is deeply committed to the safe and responsible use of AI-generated content. Recognizing the challenges posed by the rapid proliferation of AI-generated materials, especially in image manipulation, we have focused on developing innovative detection technologies. Key highlights of our research are as follows:

- Advanced Detection Technology:
 - We have developed a novel technology to detect AI-generated images without requiring additional training datasets. Our approach leverages the observation that while AI-generated images exhibit highly accurate main contours, they often contain subtle distortions in fine textures. By focusing on these distortions in images created by Latent Diffusion Models, we achieved state-of-the-art (SOTA) detection performance.
- Cost-Effective and Broadly Applicable:
 - Unlike traditional deepfake detection methods that rely on collecting and training with extensive datasets, our technology requires no additional training data, significantly reducing costs and enhancing scalability. This is particularly beneficial given the lack of transparency in training data used by many image generation AI models.
- Exceptional Results:

- Our method has been tested on both public and private models, demonstrating outstanding effectiveness and achieving SOTA performance across diverse scenarios.

For your reference, I would like to share that the 2024 LG AI Accountability Report on AI Ethics will be published in the coming weeks. Since last year, LG AI Research has issued an annual accountability report to transparently share our progress and lessons learned in implementing AI ethics principles with external stakeholders. The report will be available in PDF format on the LG AI Research website (<https://lgresearch.ai/>) in early February. If you are interested in learning more about our extensive activities in AI ethics, we encourage you to refer to this comprehensive resource.

We remain committed to advancing the Tech Accord, fostering an ecosystem where AI can be used responsibly and ethically, and we value opportunities to share our progress with our partners and stakeholders.

Please feel free to reach out if you have any questions or require further information.

Best regards,
Myoungshin

Myoung Shin Kim
Principal Policy Officer, LG AI Research



www.lgresearch.ai

From: [Redacted]

Sent: Thursday, December 19, 2024 6:04 AM

To: 김명신/님/(mskim) [Redacted]

Subject: Request for LG AI Research statement on AI voluntary commitments before January 16th

To Myoungshin Kim, Principal Policy Officer at LG AI Research,

In February 2024, your company became a signatory of the Munich Security Conference AI Election Accord.

We are contacting you today to request a statement on your company's progress toward meeting the voluntary AI commitments outlined in this agreement.

In February 2025, the Brennan Center for Justice at NYU will publish a report analyzing the progress that all signatory companies have made toward meeting these commitments. We have collected information on your company's progress through the resources provided as part of the following:

Munich Security Conference AI Election Accord

- Written response to May 2024 Senator Mark R. Warner's letter request ([link](#))

To help us understand your company's efforts to meet these commitments, please provide a written statement outlining the specific actions your company has taken to align with them, including any updates or new developments beyond the information provided in the links above. If available, please also include links to additional publicly available resources that demonstrate your progress.

We will be collecting responses from the signatory companies **before the end of Monday, January 16th**. It would benefit our analysis if you could observe the numbered commitments made in each agreement and speak on your company's specific efforts toward each numbered commitment.

We will be completing our analysis and evaluation of fulfillment based on our review of the information provided in your written responses listed above and any responses provided to this email.

Best,
Abdiaziz Ahmed
Technology Policy Strategist
Brennan Center for Justice

From: [REDACTED]
To: [REDACTED]
Subject: Fw: Request for Microsoft statement on AI voluntary commitments before January 16th
Date: Wednesday, January 22, 2025 12:10:04 PM
Attachments: [image001.png](#)
[image003.jpg](#)
[image003.png](#)

See below
Get [Outlook for iOS](#)

From: Matthew Masterson [REDACTED]
Sent: Wednesday, January 22, 2025 12:00 PM
To: [REDACTED]
Subject: RE: Request for Microsoft statement on AI voluntary commitments before January 16th

Abdiasziz,

Per our discussions here is some additional information and context for the AI Tech Accord Commitments:

Content Credential Work

- As a result of the work we have done on content credentials as outlined in the Warner letter Secretary of State Adrien Fontes recently put out this press release: <https://azsos.gov/news/888>. This marks the first time a chief election official is using content credentials to provide a trust signal to voters regarding information from his office.
- LinkedIn became the first major platform to display content credentials on the platform: <https://www.linkedin.com/help/linkedin/answer/a6282984>

Societal Resilience Grants

Microsoft and Open AI [launched a series of Societal Resilience Grants in May 2024](#) to further their commitments to transparency and societal resilience made in the Tech Accord. These grants went to support work the work of the [Older Adults Technology Services from AARP](#), [International IDEA](#), [Partnership on AI](#) and [Coalition for Content Provenance and Authenticity \(C2PA\)](#). Microsoft provided an additional grant to the human rights nonprofit [WITNESS](#).

OATS

Microsoft and OpenAI partnered with the Older Adults Technology Service from AARP (OATS) to educate older adults around how to use AI safely and how to be critical consumers of information in the age of AI.

- In partnership, OATS created resources and delivered trainings to provide AI learning materials tailored to older adults. This included a training-of-trainers for 93 of OATS' licensed partners on AI resources so that they could bring that information to older adults in their programs. OATS also released an [AI for Older Adults Guide](#), which was disseminated among the community of 500,000 older adults that OATS serves with free technology training. The grant also supported 10 subgrants in 8 states to help OATS licensed partners deliver AI programs in their communities.

IDEA

In partnership with Microsoft and OpenAI, [International IDEA's](#) developed an AI for Electoral Actors curriculum that strengthens global election official's understanding of key AI concepts, AI threats, and mitigation measures related to electoral processes.

- The training has reached the national election bodies of 26 countries in Asia Pacific and the Balkans, with trainings for representatives of the election management bodies of approximately 40 additional countries in Africa and Latin America planned for the first half of 2025. These trainings strengthened understanding of AI among the global election community while simultaneously facilitating peer exchange on the topic of AI among election authorities.

C2PA

A societal resilience grant was also given to support the Coalition for Content Provenance and Authenticity (C2PA), the standards body driving technical innovation and adoption around a content credentialing standard. The C2PA standard allows users to digitally authenticate content that comes from them and to trace the origin of different types of media, and increasingly vital component of maintaining trust in the face of deceptive AI content.

- New learning materials and a new website with learning pathways to help key audiences understand and adopt the C2PA standard will be available in February 2025, funded by a societal resilience grant.

Partnership on AI

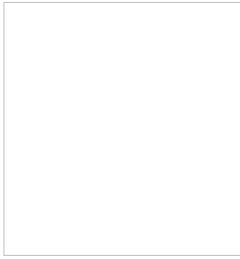
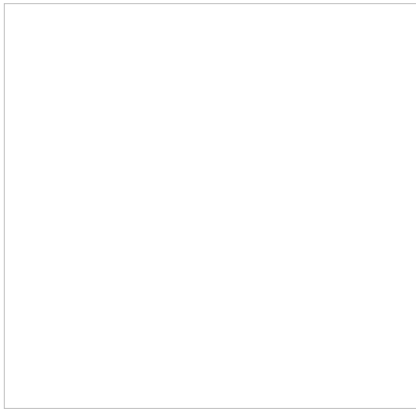
A societal resilience grant to the Partnership on AI (PAI) aimed to foster responsible AI practices and enhance public understanding of AI's societal impacts through community-building and collaborative efforts. With support from Microsoft and OpenAI, PAI developed and refined their Responsible Practices for Synthetic Media Framework, which provides guidelines for ethical and transparent use of synthetic media technologies. The grant contributed to key activities including the publication of 10 in-depth case studies, the establishment of an AI and Elections community of practice, and policy guidance on AI transparency.

WITNESS

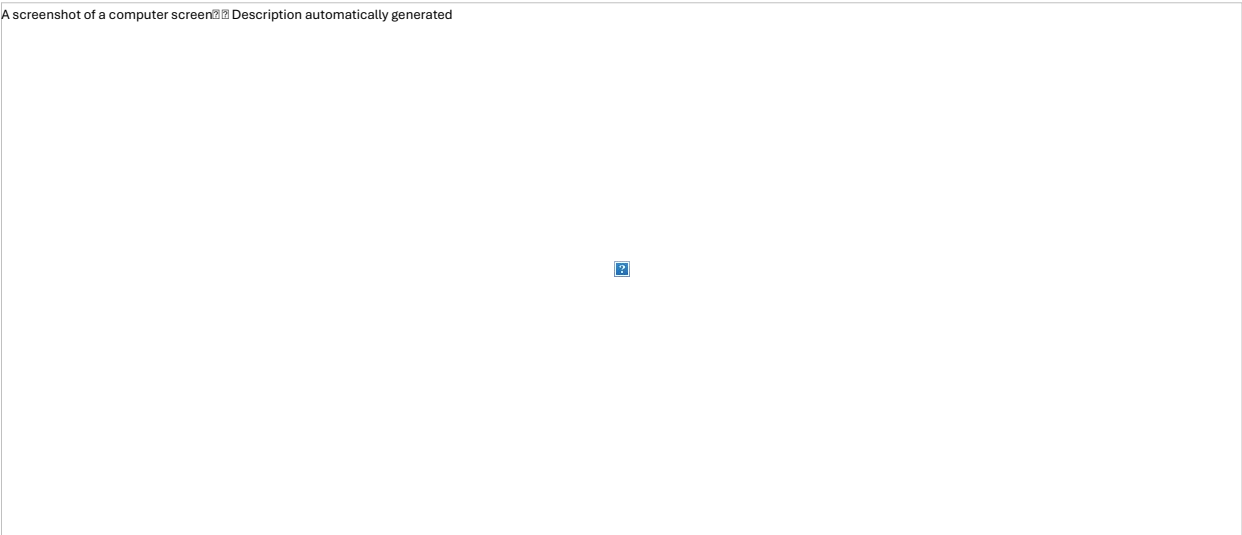
With Microsoft's support, the human rights-focused nonprofit [WITNESS](#) has enhanced journalists' and fact-checkers' capacity to address AI threats to elections. They provided in-depth training ahead of the 2024 elections in Ghana, the Republic of Georgia, and Venezuela and reached 250 additional global journalists, fact-checkers, and civil society members with training on AI and elections. Microsoft also co-leads WITNESS's Deepfakes Rapid Response Force, which connects local journalists and civil society organizations with AI and computational forensics experts.

Election Public Awareness Campaign

- We conducted AI deep fake trainings for the major political parties and campaigns around the world including in person at the RNC and DNC conventions. These trainings focused on building awareness of the risks of deceptive AI and providing tangible steps that could be taken to mitigate these risks and engage voters to find trusted election information.
 - In total we conducted hundreds of trainings in 20+ countries to over three thousand participants.
- As part of our commitments under the Tech Accord related to public awareness and engagement Microsoft ran a series of public messages and stood up a AI and Elections website focused on engaging voters about the risks of deceptive AI and where to find authoritative election information.
- This campaign ran across the EU, UK, France, and the US in the lead up to each of those major elections. Globally the campaign reached hundreds of millions of people, with millions interacting with the content, connecting them with official election information.
- In the US the campaign delivered more than 30M impressions across the US educating voters about the potential misuse of deepfakes, and directed more than 20k visits to official election websites (thanks to our partnership with the National Association of State Election Directors) providing critical voting information, including when, where and how to vote in their state.
- Here are some screenshots from campaign for reference:



A screenshot of a computer screen. Description automatically generated



Matt Masterson
Director of Information Integrity
Democracy Forward Team (CELA)
[REDACTED]

From: [REDACTED]
Sent: Monday, January 13, 2025 4:49 PM
To: Matthew Masterson [REDACTED]
Subject: [EXTERNAL] Re: Request for Microsoft statement on AI voluntary commitments before January 16th

Yes that's fine Matt. Appreciate it.

Get [Outlook for iOS](#)

From: Matthew Masterson [REDACTED]
Sent: Monday, January 13, 2025 4:03:23 PM
To: [REDACTED]
Subject: RE: Request for Microsoft statement on AI voluntary commitments before January 16th

Abdiaziz,

Your note says Monday, January 16th for submission. The 16th is a Thursday... can we submit next Monday? That would be really helpful from my perspective to get everyone's inputs.

Matt Masterson
Director of Information Integrity
Democracy Forward Team (CELA)
[REDACTED]

From: [REDACTED]
Sent: Wednesday, January 8, 2025 9:36 AM
To: Matthew Masterson [REDACTED]
Subject: [EXTERNAL] Re: Request for Microsoft statement on AI voluntary commitments before January 16th

[REDACTED]
Hi,

Can you meet at 3:15 on Friday?

Abdiaziz

Get [Outlook for iOS](#)

From: Matthew Masterson [REDACTED]
Sent: Tuesday, January 7, 2025 1:19:28 PM
To: [REDACTED]
Subject: RE: Request for Microsoft statement on AI voluntary commitments before January 16th

Abdiasziz,

Thanks for request. I have been talking with colleagues to see if there is additional info we can provide. Do you have time talk on Friday so we can cover what you already have and what could be appropriate to provide?

Matt Masterson
Director of Information Integrity
Democracy Forward Team (CELA)
[REDACTED]

From: [REDACTED]
Sent: Thursday, December 19, 2024 10:04 AM
To: Matthew Masterson [REDACTED]
Subject: [EXTERNAL] Request for Microsoft statement on AI voluntary commitments before January 16th

[REDACTED]
Hi Matt,

In July 2023 and February 2024, respectively, your company became a signatory of the White House Voluntary AI Commitments and the Munich Security Conference AI Election Accord.

We are contacting you today to request a statement on your company's progress toward meeting the voluntary AI commitments outlined in these agreements.

In February 2025, the Brennan Center for Justice at NYU will publish a report analyzing the progress that all signatory companies have made toward meeting these commitments. We have collected information on your company's progress through the resources provided as part of the following:

White House Voluntary AI Commitments

- Initial press release ([link](#))

Munich Security Conference AI Election Accord

- Written response to May 2024 Senator Mark R. Warner's letter request ([link](#))
- Official AI Election Accord progress update ([link](#))

To help us understand your company's efforts to meet these commitments, please provide a written statement outlining the specific actions your company has taken to align with them. If there have been updates or new developments since the commitments were made, please include them. Additionally, please link to any publicly available resources that could help us better understand how your company has met these commitments.

We will be collecting responses from the signatory companies **before the end of Monday, January 16th**. It would benefit our analysis if you could observe the numbered commitments made in each agreement and speak on your company's specific efforts toward each numbered commitment.

We will be completing our analysis and evaluation of fulfillment based on our review of the information provided in your written responses listed above and any responses provided to this email.

Best,
Abdiasz Ahmed
Technology Policy Strategist
Brennan Center for Justice

From: [REDACTED]
To: [REDACTED]
Subject: Fw: Request for Palantir statement on AI voluntary commitments before January 16th
Date: Friday, January 17, 2025 3:07:08 PM
Attachments: [Design Principles for Responsible LLM Workflows_A Test & Evaluation Approach.pdf](#)

Get [Outlook for iOS](#)

From: Courtney Bowman [REDACTED]
Sent: Friday, January 17, 2025 1:20:21 PM
To: [REDACTED]
Subject: RE: Request for Palantir statement on AI voluntary commitments before January 16th

Dear Abdiaziz,

Thank you for allowing us a little extra time to get back to you.

Please see immediately below our full written response to your inquiry.

Kind regards,
Courtney

Start of Response Text

We were proud to sign onto the White House voluntary AI commitments in 2023 in the spirit of contributing to and fomenting a comprehensive ecosystem for ensuring AI safety, security, and trust. Palantir’s business is not focused on the development of foundation models, which the majority of these commitments were structured to address. Nevertheless, as providers of technology platforms that can enable advanced AI integration, development, and application (involving GenAI/LLMs, as well as other classes of AI models), Palantir Technologies believed then — as we do now — that we would have an important role to play in providing first-class infrastructure to support the responsible, reliable, and effective use of LLMs and other Generative AI technologies.

In that vein, our focus in executing on the spirit of these commitments has been – and remains – on providing the supporting digital infrastructures that enable and promote responsible AI development and deployment in all forms, including foundation model applications. We have documented this at length over time through various publications, most notably, our ongoing [Engineering Responsible AI](#) blog series, which provides practical guides on how our AI enablement software platforms provide capabilities to help [reduce LLM hallucinations](#), [promote LLM explainability and transparency](#), and support a broad range of testing and evaluation (“T&E”) approaches from more [basic unit testing and prompt engineering](#) through

to more [advanced implementation of benchmarks, evaluators, LLM-as-a-judge, fairness/bias testing, an more](#). Additionally, our [Responsible AI Lifecycle \(“RAIL”\) whitepaper](#) provides a synthesis of responsible AI principles and outlines a corresponding practical framework that ties these principles to steps in the model lifecycle, our [Enabling Responsible AI in Palantir Foundry](#) blog post outlines how Foundry’s (one of our flagship platform’s) model management capabilities incorporate the principles of responsible AI to enable user to effectively solve their most challenging problems, and our [Palantir Foundry for AI Governance](#) presents practical case studies demonstrating the use of these capabilities.

In terms of specific commitments, we also offer the following details demonstrating our efforts to-date in fulfilling our voluntary commitments.

Safety

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.

- Because foundation model development is not a core aspect of Palantir’s business, our AI safety efforts are focused on expanding the lens to a broader range of testing and evaluation tools and approaches. See attached “Design Principles for Responsible LLM Workflows: A Test & Evaluation Approach” document for more details.
- [Design Principles for Responsible LLM Workflows_A Test & Evaluation Approach.pdf](#)

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

- In addition to sharing insights and thought leadership on AI safety issues documented throughout this response, we also endeavor to promote information sharing across a broad range of stakeholders through active participation in related summits, working groups, and committees such as those hosted by AISI, REAIM, NSCAI, Institute for Security and Technology, SCSP, CDAO, RAND, ARL DEVCOM, etc. At the request of policymakers, we have also shared related insights on AI safety issues in direct briefings to congressional members and staff.

Security

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

- Again, because foundation model development is not a core aspect of Palantir’s business, we have limited exposure related to safeguarding proprietary and unreleased model weights. Nevertheless, Information and Cyber Security are essential aspects of our business and the trust our customers — including some of the world’s most important public, private, and non-profit institutions — place in our software. We document this posture in detail on our [website](#) and in our [blog](#).

4) Incent third-party discovery and reporting of issues and vulnerabilities

- Palantir supports a long-standing bug-bounty program initiated in 2019 (see [“Announcing Palantir’s Bug Bounty Program”](#)) and expanded in Apr. 2020 (see [“Expanding Our Bug Bounty Program”](#)). This program covers the full range of Palantir’s infrastructure and software.

Trust

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.

- Our core systems have long supported comprehensive capabilities for tracking [data lineage](#) and [provenance](#) to support users’ understanding of the nature of the data with which they are interacting. Further trust and security primitives are built on this provenance foundation, including provenance-aware access controls and data retention policies, and we continue to explore and prototype capabilities to support customers in workflows that may require the detection of AI content.

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.

- Our publicly available [Approach to AI Ethics](#) documents our principles and methodologies for addressing societal risks, including considerations of fairness and bias, in the development and use of our software systems. See especially principles #2 (“Acknowledge technology’s limits.”) and #3 (“Don’t solve problems that shouldn’t be solved.”), which help to guide our efforts across all of our work, including in establishing appropriate and inappropriate use cases of our software by our customers.

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

- Core to Palantir’s founding mission is the objective of [designing technology to help institutions protect liberty](#). In addition to core product development work centered around responsible AI design considerations, frameworks, and principles documented throughout this response, our dedicated [Privacy & Civil Liberties Engineering Team](#) drives many of our focused efforts on [technology development](#), [thought leadership](#), [principles articulation](#), [AI public policy](#), and [AI ethics discourse](#). Our [Palantir Privacy and Governance Whitepaper](#) also provides an expansive treatment of many of these themes.

8) Develop and deploy frontier AI systems to help address society’s greatest challenges.

- Though Palantir’s business is not focused on developing frontier models, our [company mission](#) has long-focused on “build[ing] software platforms for large institutions whose work is essential to our way of life” and acknowledged that “[t]he challenges our platforms address are a matter of survival, both for the institutions we serve and the individuals who depend on them. We have the privilege of partnering with some of the

world's most important government and commercial organizations. And we believe that the work of those organizations is essential to our security and the lives that we lead.”

End of Response Text

Courtney Bowman

Palantir Technologies | Director, Privacy & Civil Liberties

From: [REDACTED]
Sent: Thursday, January 16, 2025 5:06 PM
To: Courtney Bowman [REDACTED]
Subject: Re: Request for Palantir statement on AI voluntary commitments before January 16th

CAUTION: This email originates from an external party (outside of Palantir). If you believe this message is suspicious in nature, please use the "Report Message" button built into Outlook.

Hey Courtney - thanks for the update. Not a problem!

Get [Outlook for iOS \[aka.ms\]](#)

From: Courtney Bowman [REDACTED]
Sent: Thursday, January 16, 2025 11:02:01 AM
To: [REDACTED]
Subject: RE: Request for Palantir statement on AI voluntary commitments before January 16th

Dear Abdiaziz,

I am working on pulling together a response on behalf of Palantir in time for your deadline. However, it may take a little longer to work through some internal channels to get this over to you. Would it be possible for me to send this over by EOD tomorrow, Friday, January 17 instead of today?

Thanks in advance for the consideration,
Courtney

From: [REDACTED]
Sent: Friday, January 10, 2025 6:00 PM
To: Courtney Bowman [REDACTED]
Subject: Re: Request for Palantir statement on AI voluntary commitments before January 16th

CAUTION: This email originates from an external party (outside of Palantir). If you believe this message is suspicious in nature, please use the "Report Message" button built into Outlook.

Hi Courtney,

Appreciate the response. That was an unfortunate typo on my part . We are looking for responses by the 16th - that Thursday.

Get [Outlook for iOS \[aka.ms\]](#)

From: Courtney Bowman [REDACTED]
Sent: Friday, January 10, 2025 11:33:02 AM
To: [REDACTED]
Subject: RE: Request for Palantir statement on AI voluntary commitments before January 16th

Dear Mr. Ahmed,

We've received your request and are preparing a written statement in response. I just wanted to clarify, however, your deadline for receiving statements. Your note below gives a deadline of Monday, January 16th, but the 16th is a Thursday. Can you please clarify?

Thank you,
Courtney Bowman

Courtney Bowman
Palantir Technologies | Director, Privacy & Civil Liberties
[REDACTED]

From: [REDACTED]
Sent: Wednesday, December 18, 2024 4:35 PM
To: media [REDACTED]
Subject: Request for Palantir statement on AI voluntary commitments before January 16th

CAUTION: This email originates from an external party (outside of Palantir). If you believe this message is

suspicious in nature, please use the "Report Message" button built into Outlook.

To Whom It May Concern at Palantir,

In September 2023, your company became a signatory of the White House Voluntary AI Commitments.

We are contacting you today to request a statement on your company's progress toward meeting the voluntary AI commitments outlined in this agreement.

In February 2025, the Brennan Center for Justice at NYU will publish a report analyzing the progress that all signatory companies have made toward meeting these commitments. We have collected information on your company's progress through the resources provided as part of the following:

White House Voluntary AI Commitments

- No materials found

To help us understand your company's efforts to meet these commitments, please provide a written statement outlining the specific actions your company has taken to align with them. If there have been updates or new developments since the commitments were made, please include them. Additionally, please link to any publicly available resources that could help us better understand how your company has met these commitments.

We will be collecting responses from the signatory companies **before the end of Monday, January 16th**. It would benefit our analysis if you could observe the numbered commitments made in each agreement and speak on your company's specific efforts toward each numbered commitment.

We will be completing our analysis and evaluation of fulfillment based on our review of the information provided in your written responses listed above and any responses provided to this email.

Best,
Abdiaziz Ahmed
Technology Policy Strategist
Brennan Center for Justice

Design Principles for Responsible LLM Workflows: A Test & Evaluation Approach

Palantir Technologies Inc.
November 5, 2024

Introduction

When properly and responsibly deployed, Large Language Models (LLMs) carry substantial promise for improve speed, quality of workflows across the enterprise. Lowering the barrier on technology development, increasing access to relevant and timely information to accelerate insight to action.

The integration of LLMs into operational workflows also carry their own unique risks and challenges. LLM-powered workflows pose a more complicated and unprecedented Test & Evaluation (T&E) target than workflows using more traditional AI models. The outputs of LLMs can be complex (e.g., a paragraph-long summary of a longer document), making it difficult or impossible to define exactly what a correct answer looks like, especially when “correctness” can be a matter of opinion, interpretation, or sheer aesthetics. Perhaps most concerning, LLM responses can seem authoritative and persuasive, while still being incorrect and ungrounded in reality. And even when solving traditional supervised learning problems with LLMs, it may be exceedingly difficult to assess precisely where and how outputs deviate from reality if model developers do not start with existing training data or labels. LLM powered workflows frequently contain post-model and model specific information extraction and reformatting pipelines that make comparing multiple models more complicated than the input-output function model in traditional supervised learning model leaderboard comparisons.

Fortunately, while LLM-powered workflows carry notable risks, there are responsible and effective ways to pin-point productive use cases that have high benefit/low risk profiles, as well as use cases where the risks are elevated but acceptable, relative to their mission-critical benefits. For these use cases, there are rigorous methods to determine if and how such risks can be sufficiently mitigated for the operations at hand. Namely, one can build creative T&E processes that explicitly assess and address common risks associated with LLM-powered workflows.

The remaining goal of this memorandum is to provide initial guidance on: (1) How to responsibly design LLM use cases; (2) T&E best practices for evaluating LLM use cases; and (3) Examples of existing LLM-powered use cases and their T&E methods.

Identifying and Designing LLM Use Cases

It is neither desirable, nor possible, to evaluate the risks and benefits of LLM-powered workflows as a collective or in abstraction. The risks and benefits of LLM-powered workflows — and subsequently, decisions on the conditions under which they may be sufficiently ethical, safe, and effective to employ in real-world scenarios — are highly dependent on the context of their use and the workflows in which they are embedded.

Still, we can provide some general guidance for: (a) Identifying the most promising LLM-powered workflows archetypes; and (b) Design principles to ensure their effective and responsible use.

Promising Use Case Archetypes

- **Semantic Search:** Semantic search uses mathematical representations of natural language learned by the LLM during its training to build a search function that reflects a more nuanced understanding of natural language queries, beyond keywords and synonyms. As a system, it is best used to augment more traditional information retrieval methods.
- **Information Extraction:** Using LLM language capabilities to extract structured knowledge from free text. For example, an LLM can be used to extract named entities or even full knowledge graphs from intel reports, as well as more nuance extractions such as timelines for events alongside summaries of the most relevant information (e.g., converting a full medical record, including doctor notes into a series of diagnosis, treatments, and results).
- **Document Q&A:** LLMs can assist with understanding documents through a Q&A interface. Typically, this is done by adding a software or AI layers on top of semantic search to summarize results for the questioner. Although there are instances of LLMs being used as a knowledge store directly, we believe this should never be the case since LLM knowledge stores are particularly and demonstrably prone to “hallucinations” (i.e., when a neural network inserts incorrect, non-factual information into a generated output, or when it produces an output that does not follow the prompt instructions). Hallucinations of this sort are dangerous both because they may appear to be presented in a confident and convincing form, and also because the underlying semantic error may be difficult (even impossible) to systematically debug. Even when extending semantic search, we need to take special care with use case selection, risk assessment, and workflow design in anticipation of model hallucination issues. While the risk of hallucinations makes this LLM-powered workflow too risky for many situations, it can nevertheless be a useful accelerant in lower-risk situations (e.g., allowing users to learn about an API where backing documentation provides a fail-stop way to answer questions).
- **Content Generation:** LLMs or other generative models (e.g., image, audio, multi-modal, etc.) can be utilized to produce creative content. For example, LLMs can be deployed as writing assistants and content strategists to create drafts of marketing copy and logo design. LLM-powered content generation has not, however, become a common

enterprise use case across all domains of application, even as it dominates media coverage and has important implications for commercial and educational environments.

- **Hypothesis Generation:** Sometimes cast as “reasoning” capabilities, LLMs can use approximate retrieval and generation capabilities to create hypotheses and suggested answers to proposed problems. This approach may also be combined with semantic search and document Q&A type workflows. For example, an LLM could generate diagnostic advice to a physician, or suggest possible reasons for a part failure and retrieve guidance for how to repair, based on the query, data, and input provided. We intentionally are not referring to this capability as “reasoning” because it is critical to acknowledge the fundamental limitations of such a use case, and accordingly construct any such related workflow to include *requisite verification* — through other tools or human review — of proposed answers. This use case is better considered as more of a way to “brainstorm ideas,” rather than as a source of strategic advice, reasoning, or optimized solution.
- **Tool Routing:** One can use an LLM to process requests — from a human in the form of natural language — to engage with or run a specific tool. This may also include engagement with sub-components of tool selection, extraction of data for the tool, formatting data for the tool, and presenting results from the tool. Tool routing can be used to lower the barrier to entry for tool usage since end-users do not need to understand the technical details on how to find and run the tool. At the enterprise level, tool routing can enable dynamic use of information infrastructure, as well as the combining of disparate systems to create a seamless experience for the end-user. At the workflow level tool routing enables the LLM to delegate answers to specialized tools such as a calculator, route optimization application or even another LLM. For example, an existing LLM use case allows end-users to iterate with an optimization program (such as a scheduling system), including the ability to specify optimization requirements using natural language and visually review the results.
- **Agents:** Similar to tool routing, an LLM “agent” takes machine-driven input and, based on this input, the LLM-driven “agent” takes action (since full agent autonomy is rare and in some cases explicitly prohibited, hybrid situations with human oversight are common). A typical example here is where you have a state space model of some workflow and agents manage transitions between states. Guardrails, a strong security model, and active monitoring are key design principles to constrain agent actions allowing for safe operations.
- **Multi Agents:** LLMs are assigned rolls/personas and jointly solve a task. Rolls include proposing solutions, calling tools, critiquing and correcting proposals, and judging results. We are seeing LLM as judge workflows are becoming increasingly common and while true many-agent systems hold promise for the future in most applications today they do not perform well enough for production use cases.

While the above use cases serve as potential archetypes for successful LLM-powered workflows, we can also highlight several common mistakes that lead to unsuccessful and/or irresponsible LLM use cases — especially when the appropriate T&E processes are not completed before real-world use.

- **Overestimating LLM Capabilities:** A primary mistake that is often made in pursuing LLM applications is the overestimation of a given LLM's capabilities. LLMs produce results that are often anthropomorphically tuned to appear to carry the voice of understanding, reasoning, and authority, when in fact the results reflect none of these human characteristics. T&E processes can help us understand the strengths, and most importantly, the weaknesses and limits of given models — the success of any use case is reliant upon an understanding and accounting of what a given LLM is (not) capable of.
- **Treating Models as Factual Oracles:** The process of training an LLM — which combines learning both language syntax and content from training data — can lead to content being imperfectly recalled at inference time (so called “hallucinations”). As an imperfect database, one cannot use an LLM while expecting it to consistently know and reveal factual information. Absent other design considerations, LLMs are prone to misleading the end-user since the content is seemingly coherent, even when it is inaccurate.
- **Premature Automation:** A related concept to overestimating LLM capabilities is trying to prematurely automate a process that is not sufficiently developed or conditioned for automation. For example, consider trying to build a no-code application for executives to answer questions about their business state using natural language (which, behind-the-scenes, is translated by an LLM into database queries and a declarative visualization framework). The executives may not have the requisite skill to evaluate the generated Structured Query Language (SQL) queries, both in terms of how it matches the company's data schema and ontology, as well as from a SQL linguistic standpoint. Given the current state, error rates for this kind of task (where an error means the generated query produces incorrect results based on the request) can be quite high. As such, it would be irresponsible to automate the high-level workflow, and instead, an intermediate workflow accelerator (e.g., a co-pilot-like tool for data analysts) would make more sense. Eventually, the process of producing the intermediate tool may lead to the development of full automation through the creation of increasingly sophisticated techniques and training data that is vetted by a human expert.

Design Principles

While the selection and success of any LLM-powered workflow should be context specific, we can offer a set of universal design tips to avoid some of the most common pitfalls when developing AI- (and specifically, LLM-) powered systems.

The following design principles are particularly applicable to high risk/high impact AI-systems, where system actions and outputs have significant consequences. These principles accept that the AI components for these systems will fail and provide mechanisms to ensure workflow outcomes can still be achieved. In the context of Generative AI in particular, we need to address truthfulness, inconsistency, accountability, explainability, and hallucinations. While none of the prescriptions are absolutes, they are useful for safeguarding LLM systems; and when such design principles are missing from an AI system, more care is required to account for the likely risks and trade-offs of the use case.

Above all, the LLM is a means to an end in a workflow, not the purpose of the workflow. They offer a set of new computing capabilities to integrate human thought and judgement with algorithmic compute and tools.

- **Human Oversight:** Whenever an action is proposed, explicit approval for that action should be presented to the end-user before the action is taken. While this design principle is not a panacea, it does provide an important layer of scrutiny and accountability, giving users the opportunity to review and deliberate consequential actions in a manner that is efficiently embedded in the workflow. Designing AI processes that maintain human oversight can reduce risks associated with AI usage, while preserving efficiency and other gains. Human review can also function as a fail-safe fallback in the event of complete AI failure — an important safeguard in consequential workflows. The following are more explicit examples of how engineers can work human review into LLM-powered workflows:
 - No action can be taken without the user reviewing and actively approving the action;
 - Actions can be edited before execution, with the edited action being saved for continuous model development;
 - Actions are presented explicitly (e.g., showing the names of map layers being added to a map visualization application) — and is preferred to implicit presentation (e.g., "adding 43 layers") — to facilitate user understanding of the results;
 - Actions are performed by specialized tools and functions (e.g., hitting a REST API) and are programmatically formulated based on what is presented to the user.
- **Return Immutable Results:** While basic programmatic reformatting is acceptable, results from running tools should be displayed/communicated to the user without further editing by an LLM, which could modify, delete, hallucinate values in deceptive ways.
- **Limit Reasoning:** It is preferable to use a programmatic tool — such as an optimization engine, SQL database, or even a simple calculator — over asking an LLM to do reasoning or algorithmic tasks. In a similar vein, when hallucinations are unacceptable, information retrieval is better solved via search (potentially using semantic search in conjunction with other traditional search techniques), as opposed to via something like a Q&A chat-based interface. A Q&A interface, for example, can introduce opportunities for hallucinations and limit the user's ability to investigate the relevant data sources used during retrieval. In general, a high occurrence of hallucinations are a sign that too much reasoning capability is being expected of the LLM in the workflow.
- **Prompting is for Developers:** Writing the prompt that steers LLM behavior and tool use is not something that users should control. This is a developer feature to help define the behavior for specific workflows.
- **Limit Implicit Tool Chaining:** Having an LLM call multiple tools with the output of one tool can lead to cascading errors and reduces system transparency. If it is necessary for an LLM to call multiple tools through the output of another tool, it is important that a human be kept in the loop for such actions.

- **Systemic Transparency and Auditability:** The foundational digital infrastructure of an LLM should be constructed so that all necessary information to reconstruct LLM outputs are saved so that post-hoc analysis and debugging can be done.
- **Link to Source Material:** The user should be able to view and interrogate the relevant sources retrieved in order to understand lineage, verify accuracy, and ultimately build trust in a system. Generated citations should not be taken at face value, and it is better to programmatically determine the material used, such as the case with semantic search where you can return the portions of text that were used.
- **Add Guardrails:** Consider the value of adding guardrails (i.e., external limits on autonomous model actions) to LLMs where appropriate. For example, one can formulate a guardrail that requires human oversight for any decision with a financial impact above a certain figure, while actions with less impact do not require human review. In many cases, guardrails may be easier to formulate than directly checking the output of the model, and many heuristics can be combined simultaneously.

Addressing Hallucinations

Perhaps the most common and enduring challenge to LLM use — with currently no technical solution — is the generation of hallucinations. While there is reason to believe that hallucinations are an intrinsic property of current LLM architectures and training methodologies, there are useful tactics for reducing the presence and impact.

In fact, all of the design principles above have a role in either limiting hallucinations or reducing their impact when they occur. Human oversight, returning immutable results, and guardrails in particular can contribute to safer and more accurate systems. While heuristic checks inserted as guardrails are not guaranteed solutions (e.g. verifying that extracted entities exist in the text itself, matching of n-grams in solutions, or in the case of code, compilation and syntax checks), they can help eliminate egregiously off-topic hallucinations.

Additional design techniques for reducing hallucinations include instructing the LLM to not use information outside of the content provided in the prompt (“grounding”), using an LLM as a “judge”, training on in-domain data, and reinforcement learning through human feedback and related techniques. Furthermore, we have found that using multiple LLMs in ensemble can be used to detect and remove spurious hallucinations.

Overall, while there is no guaranteed solution to LLM hallucinations, we can build safer and more effective systems using LLMs through a combination of attentive system design, internal checks, and model tuning. Furthermore, because systems can never be completely error- or risk-free, proper risk assessment and mitigation strategies are an important aspect of selecting, building, and deploying workflows.

While it is beyond the scope of this document to go into further details on how to assess the risk and benefits of proposed LLM-powered workflows, we can refer the reader to [NIST’s AI Risk Management Framework](#) for a comprehensive guide and playbook for AI risk that applies to all

AI systems (including those that utilize LLMs). Furthermore, acceptance criteria — i.e. deciding when and under what conditions a system can be deployed — should be done at the “use case” level and on a case-by-case basis, with sensitivity to each cases’ uniquely identified risks and benefits. Examples of acceptance criteria include [proposed FDA guidance](#) for regulating medical devices utilizing AI.

T&E Best Practices for Assessing LLM Use Cases

The key to deploying LLM use cases is to subject proposed LLM-powered workflows to a well-designed and stringent T&E processes. Through metric-based analysis, experimentation, and other scientific methodologies, organizations can to identify, test, iterate, and then deploy successful LLM-powered workflows.

While this document does not provide a step-by-step T&E approach or exhaustive set of T&E strategies, we can outline general best practices for T&E processes. Additionally, the strategies provided should also be considered “in addition” to what is normally required for the T&E process of virtually any type of machine learning or statistical model (e.g., standard metrics tracking between model versions and over time, being able to approve and recall models from production, model review and compliance workflows that facilitate discussion between AI experts and other stakeholders, ongoing monitoring for model and system drift, etc.)

Below, we provide guidance on: (a) Basic-level T&E strategies; (b) Advanced-level T&E strategies; (c) Operational testing strategies; and (d) Scenario and simulation testing strategies.

Basic-level T&E Strategies

- **Reduction to Supervised Learning:** The easiest and perhaps most common T&E strategy is to reduce the problem to look like a classic supervised learning problem and use existing and well-understood T&E methodologies.

To use an example, consider modeling the problem of extracting factory locations from incoming problem reports as named entity recognition, or modeling the problem of an LLM choosing among a selection of tools to run as text classification.

When using an LLM to solve one of these problems, it is possible to start with very little, or even no labeled data. The first task then is to gather requisite testing data so that future T&E can be performed using standard methods. In low-risk scenarios, after trying several examples, you can gather data using a human-in-the-loop review process, either through an explicit label generation process or, in benign circumstances, in production directly. To bootstrap and accelerate the creation of labeled data, it can be useful to use a data generation and labeling process with an LLM employed to produce test examples, which are then validated by a human. When taking this approach, users should always record which examples are generated by a human and which are generated by an LLM as performance may vary due to (perhaps subtle) differences in the data. This validated, labeled dataset can then be used as a test set to quantify model performance and validate continued performance in case of model changes.

Furthermore, as part of the T&E process for bootstrapping a statistically-relevant dataset, you can use unit tests and other methods to help check and validate system performance.

After a system has been running in this configuration, users may have enough training data to build a more traditional supervised model. The additional costs of training and sustaining a supervised model (compute, time, talent, data) should be weighed against the benefits of basing a system around an LLM. One benefit includes the flexibility to easily adjust and reformulate the model; for instance, in our ongoing example, changing data extraction specifications from just “factory locations” to “factory *and* retail store locations” is potentially just a matter of inputting one or two lines in the prompt.

- **Track Prompts as Hyper-parameters:** Prompts to LLMs have significant impact on model output, with variations in prompt structure, wording, length, examples, and many other factors, having significant impact on model behavior. Prompt engineering and methods for steering model outputs towards desired outcomes has thus become the subject of much experimentation. Yet it remains challenging to identify patterns for *why* certain prompting strategies produce results through evaluating model characteristics alone. Evaluating empirical results from experimentation is currently the only reliable method for making such assessments.

As such, it may be worthwhile to employ model debugging and improvement tactics to decompose prompts into additional metadata, for instance, to track which embedded examples (if any) are used. In more complicated situations, the generation of the prompt (potentially using AI) is dynamic and should no longer be considered as a hyper-parameter, but the generation process itself should undergo a T&E process.

Advanced-level T&E Strategies

- **Decompose Tasks:** LLM tasks, unlike tasks from traditional models, can consist of performing multiple complex actions with each inference. As such, to evaluate such tasks, it is best practice to decompose the output into components that are both understandable and represent task targets to develop against.
 - **Generated Content:** In the abstract, we can decompose generated content and measure if it is syntactically and semantically correct independently. In the case of free text, both of these may be difficult to evaluate and require “Human Evaluation” or “Red Teaming” (see sections below for more detail). However, in many use cases, when using an LLM to form code or hit tool API endpoints, we can separately evaluate syntax (does the code parse as valid) from semantics (does the code do the requested task).

Further decomposition isn’t definable in the abstract and thus should be considered at the “use case” level and on a case-by-case basis. For example, in the case of asking for a medical summary of doctors’ notes, we can separately ask about syntax (is the output correct language), semantics (did it capture the most important elements), and add in time (did it link the event to the proper event date), event sequencing order, etc.

- **Model State Space:** Although there are many kinds of agents/agent systems, one we see frequently is where the agents model a workflow as a Markov Decision Process and

control state space transitions between sub-workflow elements. These systems can be difficult to evaluate even when combining with other methods outlined here (such as task decomposition, looking at overall workflow metrics, etc.). Simulations and scenario analysis can be useful for understanding how interconnected pieces behave. Evaluation in more generality is beyond scope of this document.

- **Workflow KPIs as Metrics:** Some workflows do not have easy statistical measurements applied directly to the output of the model; however, even in those cases, you should track model impact on the workflow as a whole via a KPI. For example, consider an LLM to assist with call center transcript generation. While it can be difficult to understand the effectiveness of a particular generated talk track, you can track system impact at the metric of “speed-to-resolution” and “aggregate cost.” Note that this information may not be immediately available at inference time and may only be understood in aggregate.
- **Human Evaluation:** Human evaluation (i.e. observation, approval, disapproval), much like data labeling, is an intensive process that in the best case is integrated seamlessly into the workflow, but in other cases may require both training and domain expertise and an ancillary “oversight” workflow. In many workflows, human oversight, with implicit evaluation of model performance, is both the most important safeguard and source of evaluation data.
- **Red Teaming:** Although it has a longer history in other domains such as cybersecurity, Red Teaming is now a developing area for the testing of AI systems. Unlike Human Evaluation, in Red Teaming, humans take an active role in probing system limitations. But even as Red Teaming standards are being developed and refined, active efforts to conduct Red Teaming should take care to include subject matter expertise in the area of LLM application, which may include bringing in outside experts to assist with the Red Teaming process.

While Red Teaming is active and dynamic testing of LLMs by Humans, as attacks are developed, to the degree possible, they should be codified into test suites to be run in future testing. Successful attack, attack templates, and variations produced with the help of an LLM should be recorded and available for testing for LLM attack mitigation tracking.

- **Adversarial Evaluation:** Models can be evaluated with respect to a (growing) list of documented adversarial attack techniques. The specific details of the deployment determine what models of attack should be considered for evaluation. For example, while some models may not directly receive input from users (via chat interface), they may process information pulled from a database, which represents a distinct source of adversarial risk.
- **LLMs to Evaluate LLMs:** Increasingly, LLMs are being used to validate outputs created by a different LLM. This technique can be used to both reduce errors and estimate accuracy. However, this approach risks introducing a “turtles all the way down” scenario where the evaluation procedure — itself a model — needs an evaluation procedure. From the practical perspective of moderating errors, this approach has its merits. However, it is not a robust stand-alone T&E process, can introduce additional brittleness

and errors, and should be discouraged as a singular approach, especially in high-risk application environments. Human evaluators and human/expert-driven evaluation approaches should be preferred in highly consequential application contexts.

- **Building Test Data using LLMs:** For many problems you may have a dearth of test data. For instance, . Techniques such as using one LLM to generate potential test case examples, coupled with both human and LLM review, have been shown to create high quality data for both testing and training (fine tuning) of LLMs against specific problems. However, LLM generated test data should be used carefully and with some skepticism as it is frequently not from the same distribution as natural data. This does not mean to undermine this approach, but one should be mindful of this failure mode, which requires careful approach to move from development to production for a model tested in this way.

Operational Testing Strategies

Operational testing strategies involve the use of realistic operational conditions to validate the performance of AI systems. A key example of “field-to-learn” methodologies, operational testing is achieved for all systems (sometimes unintentionally) when they are deployed for actual usage.

As a general principle, operational approaches are both riskier and costlier than standard supervised learning model T&E processes. Solely relying on operational testing — in lieu of other kinds of T&E methods — may seem unusual for individuals coming from an AI background, but most other fields of engineering are not structurally setup to have a zero-risk T&E infrastructure, which we usually have for AI models. In many fields, operational testing isn’t just the norm, but the only way that ideas can be vetted. For example, operational testing in the form of clinical trials is central to drug development, which is a high-risk and high-consequence domain.

Extending beyond traditional model testing, operational testing provides the opportunity to gather system-level KPIs, such as time to execute, cost, and comparisons relative to the prior workflow (if it exists). These KPIs give a more holistic understanding of a new AI system’s value , and can complement (and at times substitute for) model-specific T&E by capturing more realistic model impact and risks through actual usage. Furthermore, operational testing can reveal which errors are most impactful on workflow outcomes, and even sometimes show that improved model performance can lead to negative outcomes under operational use (e.g., if users start to pay less attention because of boredom, or other reasons, as they rely on a well-performing, but imperfect model). Operational testing is also useful for making programmatic decisions (e.g., Is the AI-powered workflow delivering value).

There are some conditions in which traditional T&E methods are insufficient, making operational testing necessary. For example, some use cases require operational testing because the workflow has delayed feedback elements or include sequential interactions with other complex systems or people. Many LLM-driven workflows fall into this category, such as

marketing campaigns or call center script development. More broadly, many autonomous systems require operational testing because the nature and variability of the real world is greater than what can be captured in static data sets and simulations.

Key techniques for safe and effective Operational Testing include:

- **Human Review:** Manual review of LLM workflows and outputs to ensure the AI system is functioning as expected. Human review can include both basic testing and dynamic “Red Teaming” of LLM and system capabilities.
- **Safety/Unit Tests:** These tests are “sanity checks” where good results are known to the person running the test, but exist in small enough numbers that prohibit meaningful statistical testing. The difference between human review and unit tests is that while human review is subject to feedback loops where the human can interrogate different aspects of the system and respond to system behaviors, unit tests are automated. Unit tests are akin to in vitro toxicology tests that are done before starting phase 1 of clinical trials.
- **Incremental Release:** This technique involves releasing new versions of an LLM-powered workflow to a small group of users — who know that they are using a new model — in order to do basic sanity and safety testing before approving a broader release. This is akin to a phase 1 trial in drug development, where the objective is to determine toxic effects before proceeding with larger studies.

Once the basic safety and performance of a system is established, complete end-to-end testing can commence in as realistic (operational) situations as possible, including:

- **A/B Testing:** With A/B testing, evaluators can compare the new model to an existing model with well-defined performance characteristics. The model being evaluated is usually the current production model but you could consider another baseline. A/B testing is akin to phases 2-4 in clinical trials, where researchers are independently trying to verify efficacy, compare to existing standards of care, and identify long-term effects in discrete trials. In most cases, this level of statistical rigor is not applied to AI systems.
- **Shadow Testing:** While not available to all LLM-powered workflows, you can sometimes deploy and evaluate a new model as a hypothetical alternative to an original model, whereby the original model is still used for the actual workflow, but the alternative choice of the new model is recorded and evaluated through post-hoc analytics.
- **Monitoring:** As one moves away from canned testing to realistic and open-ended usage, the on-going monitoring of system performance is critical to ensure that one isn’t deviating from expected performance. For example, in clinical trials, this is called post-trial monitoring and is critical since the true variance of the population is difficult to capture under clinical trial conditions.

Scenario and Simulation Testing Strategies

Scenario and simulation strategies are the “digital twin” equivalent to running an autonomous system in a world simulation. They represent a middle ground between traditional model-focused testing and the real-world, user-focused testing described as “Operational Testing” above. Scenarios are ultimately a combination of simulations, heuristics, programmatic

interventions, and real-world data. In a scenario, a new model is run in its operational context, but potentially alongside simulations of other aspects of the system in order to test and characterize its operating characteristics.

Scenarios are most commonly used to understand complex systems that are difficult to reason through in the abstract. They provide a safe environment to systematically explore counterfactuals and enable optimization of difficult-to-reason-through parameters.

For LLMs in particular, scenarios and simulations can be used to measure how an LLM interacts with the system in which it is embedded, and from a T&E perspective, can add important safety and performance information that is otherwise inaccessible in the absence of operational testing.

Example of LLM-Powered Workflows

Below, we provide one example of a Palantir-developed, LLM-powered workflow — and their corresponding T&E procedures.

Nurse Shift Hand-off

Palantir is currently building an LLM-powered workflow for a hospital system to help nurses manage information exchanges when they change to a new shift. For each patient, the nurse needs to understand the current state of the patient, their relevant medical history for care, and a list of what needs to be done during their shift. However, all the notes in a patient's medical record contain more information than is needed to do an efficient hand-off, and the information exchange process can be error-prone and time-consuming to digest.

In this workflow, an LLM is used to condense the medical record as an easily digestible timeline. All timelines are vetted by the outgoing nurse as being correct, complete, and relevant, and they are able to make corrections and additions before it is recorded. The workflow has additional guardrails, including verification that citations for the summaries appear in the text, and comparisons of previously generated timelines to the current timeline with non-blocking flagging of differences (e.g., highlighting the absence of a medication that appeared in the previous day).

Over time, we expect to add additional elements to this workflow, including “resolving” a problem in the past so that it doesn't continue to clutter the information exchange, alerting for potentially relevant information buried in the past, as well as integrating structured data into the timelines.

T&E Plan

Following the design principles for higher-risk generative use cases, human review is inserted, and based on the feedback, is used to better understand model performance. Nurses are already experts in this workflow, having manually curated and communicated this information

in the past, and so end-users can rely on their expertise to give nuanced understanding of the shortcomings of the system. Errors are decomposed into several categories which can be individually tracked (e.g., Are hypothesized conditions being asserted as fact? Is important information missing from the timeline? Was extraneous information added to the timeline?). Additionally, end-users will want to track temporal information, such as whether extractions are associated with the correct time, and if not, whether they are even in the correct order. Important KPIs to track for this workflow are nursing errors, time savings in nurse shift change, and patient well-being.

While the product is currently under ongoing development, operational testing remains key. This includes manually inspecting results on historical medical records, as well as executing an incremental release.

Conclusion

The potential for LLM-powered workflows to improve the effectiveness, safety, and security of operational workflows is clear. Yet, as is true for any new technology, it would be imprudent to directly integrate LLMs into operational workflows without first calculating the risk/benefit trade-off for deployment at the use case level, as well as rigorously testing and evaluating how each LLM-powered workflow will perform using industry best practices.

From: [REDACTED]
To: [REDACTED]
Subject: Fw: OpenAI Response - Brennan Center for Justice: Request for OpenAI statement on AI voluntary commitments
Date: Thursday, January 16, 2025 7:22:19 PM

OpenAi response

Get [Outlook for iOS](#)

From: Johanna Shelton [REDACTED]
Sent: Thursday, January 16, 2025 6:57:20 PM
To: [REDACTED]
[REDACTED]
Cc: Chan Park [REDACTED]; Becky Waite [REDACTED]
Subject: OpenAI Response - Brennan Center for Justice: Request for OpenAI statement on AI voluntary commitments

Abdiaziz & Larry -

Below please find our response to your request. Apologies for the length of this email, but we wanted to make sure the links would work well for you (please let us know if not). We are happy to answer any questions on this or other OpenAI efforts - please reach out anytime.

- Johanna Shelton
US External Affairs Lead

OpenAI Response:

Brennan Center for Justice -

Thank you for reaching out in advance of your forthcoming report on the progress signatory companies made in conjunction with the White House Voluntary AI Commitments and Munich Security Conference AI Election Accord (“The Accord”). We welcome the opportunity to share how OpenAI advanced election integrity and other efforts through The Accord and the broader White House Voluntary AI Commitments. We are proud to build and release models that are industry-leading on both capabilities and safety.

Below, we supplement the materials you have already identified with other blogs and statements that further demonstrate OpenAI’s efforts since the July 2023 and February 2024 commitments. We note that several of these efforts extend beyond the specific scope of the commitments but demonstrate efforts that OpenAI undertook to promote a responsible and secure generative AI ecosystem.

People around the world are embracing generative AI in ways that boost creativity, productivity, and learning. We believe it will be increasingly important for technology

policy leaders to be fluent in helping people understand the tools used to create content they find and share online, and thereby further our shared goal of helping inform society as a whole.

White House Voluntary AI Commitments:

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.

- **OpenAI Preparedness Framework.** In December 2023, we published a [Preparedness Framework](#) that forms the foundation of how we establish criteria for deploying frontier models safely, grounded in the following guiding principles:
 - Preparedness should be driven by science and grounded in facts → we want to focus on concrete measures and data-driven predictions.
 - Iterative deployment → We learn from real world deployment and use those lessons to mitigate risks.
 - We track four broad areas of capabilities: cybersecurity; chemical, biological, radiological, and nuclear (CBRN) threats; persuasion; and autonomy.
 - The empirical understanding of catastrophic risk is nascent and developing rapidly. Assessments of current frontier model risk levels must be dynamically updated to reflect the latest evaluation and monitoring science.
 - Governance structures are important for accountability and oversight across the development process.
- **Model Red Teaming and Testing.** OpenAI's [Red Teaming Network](#) consists of a community of trusted and experienced external experts, including individual experts, research institutions, and civil society organizations in a wide variety of

domains, who bring a diversity of perspectives and lived experience. These experts have helped identify risks related to cybersecurity, biological and chemical threats, and societal harms, among others. We have also collaborated with third parties, such as Model Evaluation and Threat Research (METR) and Apollo Research, to assess the safety and capabilities of our advanced models, including GPT-4o and o1. This complemented other collaborative AI safety work including our [Researcher Access Program](#) and [open-source evaluations](#). In [May 2024](#), for example, we shared that more than 70 external experts helped to assess risks associated with GPT-4o through our external red teaming efforts, and we used these learnings to build evaluations based on weaknesses in earlier checkpoints in order to better understand later checkpoints. [Earlier](#), as part of our red-teaming of DALL-E 3, we provided early access to external experts from a range of industries to help probe the systems including the model's ability to provide visual information to develop, acquire, or disperse CBRN. In [December 2024](#), we invited safety researchers to apply for early access to our next frontier models.

- **Board Safety and Security Committee.** In [May 2024](#), our Board formed a Safety and Security Committee responsible for making recommendations to the full Board on critical safety and security decisions for OpenAI projects and operations.
 - When the committee was established, their first task was to evaluate and further develop OpenAI's processes and safeguards over a 90-day period.
 - Following that review, in [September 2024](#), we shared the Committee's recommendations across five key areas, which were adopted. These included enhancements we made to build on our governance, safety, and security practices:
 - Establishing independent governance for safety & security
 - Enhancing security measures
 -

Being transparent about our work

- Collaborating with external organizations
- Unifying our safety frameworks for model development and monitoring

- **Advancing Research in AI Safety**

- **Model Interpretability Research.** OpenAI has conducted research on understanding [how language models process information](#), including investigating the neurons in large language models and publishing our findings.
- **AI Safety Fund.** As [announced on October 25, 2023](#), OpenAI and other founding members of the Frontier Model Forum (Anthropic, Google, and Microsoft), in collaboration with philanthropic partners (the Patrick J. McGovern Foundation, the David and Lucile Packard Foundation, Eric Schmidt, and Jaan Tallinn) established a new \$10 million [AI Safety Fund](#), administered independently by Meridian Prime, to promote research in the field of AI safety focused on supporting the development of new model evaluations and techniques for red teaming. In [November 2024](#), the fund announced the first grant awards were issued to AI safety researchers researching novel methods to evaluate and address risks of frontier models. Twelve grantees across four countries – the United States, United Kingdom, South Africa, and Switzerland – received funding with a total disbursement of over \$3 million. A second round of funding is underway.

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

- **Frontier Model Forum.** In July 2023, together with Anthropic, Google, and Microsoft, we [announced](#) the [Frontier Model Forum](#), a new industry body focused on ensuring safe and responsible development and use of frontier AI

models globally. On [October 25, 2023](#), the Executive Director was announced and the Forum released its first technical working group update on red teaming to share industry expertise with a wider audience.

- **International Convenings.** We participated in several notable government convenings on AI, including:
 - **UK AI Safety Summit.** OpenAI [joined](#) the AI Safety Summit hosted by the UK November 1-2, 2023, at Bletchley Park, bringing together international governments, AI companies, academia, and civil society to advance global discussions on AI.
 - **Munich Security Conference.** In February 2024, OpenAI [joined](#) other industry partners, civil society leaders and governments around the world to unveil the Accord to help safeguard worldwide elections from deceptive AI use.
 - **AI Seoul Summit.** In May 2024, at the AI Seoul Summit, a major international conference co-hosted by the South Korean and U.K. governments, we joined industry leaders, government officials, and members of civil society to discuss AI safety and share information about risk mitigation measures, and provided [an update](#) on those efforts.
 - **United Nations General Assembly (UNGA).** In [September 2024](#), OpenAI leaders hosted a day of events including demos, panel discussions, and fireside receptions to meet with leaders and partners from the UN, governments, civil society, and other private sector entities around the world.
- **Agreements with US & UK AI Safety Institutes.** On August 29, 2024, the US AI Safety Institute at the U.S. Department of Commerce's National Institute of Standards and Technology [announced](#) an agreement with OpenAI that enables formal collaboration on AI safety research, testing and evaluation. We shared in our [September 2024 o1 preview announcement](#) that we had formalized agreements with the U.S. and U.K. AI Safety Institutes and had begun operationalizing those agreements, including granting the institutes early access to a research version of the o1 model.

3) Invest in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights

- **Security Controls Including Securing Model Weights.** As detailed in our [October 25, 2023 update](#), we dedicated significant resources to protect proprietary unreleased model weights. We also implemented commercially reasonable technical, administrative, and organizational measures designed to prevent personal information loss, misuse, and unauthorized access. We have also [described](#) high-level details about the security architecture of our research supercomputers, which enable us to deliver industry-leading models in both capabilities and safety while advancing the frontiers of AI. We prioritize the security of these systems and have invested in security measures to protect model weights from exfiltration in our research environments.
- **Cybersecurity.** Cybersecurity is a critical component of AI safety, and we've been a leader in [defining the security measures](#) that are needed for the protection of advanced AI. A [May 2024 safety update](#) outlined cybersecurity efforts, including restricting access to training environments and high-value algorithmic secrets on a need-to-know basis, internal and external penetration testing, [a bug bounty program](#), and more. We believe that protecting advanced AI systems will benefit from [an evolution of infrastructure security](#) and explored novel controls to protect our technology. To empower cyber defense, we funded third-party security researchers with our [Cybersecurity Grant Program](#).

4) Incent third-party discovery and reporting of issues and vulnerabilities

- **Bug Bounty Program.** As described in our [October 16 2023 update](#), we started a bug bounty program that invites independent researchers to report vulnerabilities in our systems in exchange for cash rewards ranging from \$200 for low-severity findings to \$20,000 for exceptional discoveries. We partnered with Bugcrowd, a leading bug bounty platform, to create a submission and reward process, available on the [Bug Bounty Program page](#).
- **Reporting Structure for Vulnerabilities Found After Model Release.** After making the voluntary commitments, we initiated a working group within the Frontier Model Forum to create a mechanism for the responsible disclosure of dangerous capabilities among AI labs. This mechanism will aim to enable the confidential disclosure of significant risks identified in frontier models among

frontier labs and other AI labs. The initial focus encompassed national security-related domains such as CBRN capabilities, along with other dangerous capabilities like self-replication, deception, and manipulation.

- **DARPA Challenge.** As announced at Black Hat in August 2023, we participated in a [DARPA AI Cyber Challenge](#) (AIXCC) to help cyber defenders better protect critical networks.

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

- **Coalition for Content Provenance and Authenticity Steering Committee.** In [May 2024](#), we joined the Steering Committee of C2PA, recognizing the strong societal value in promoting common ways of sharing information about how digital content was made so that it's easy to recognize across many situations. Through the C2PA Steering Committee, we joined with others in an effort to adopt, develop and promote an open standard that can help people verify the tools used for creating or editing many kinds of digital content. C2PA is a widely used standard for digital content certification, developed and adopted by a wide range of actors including software companies, camera manufacturers, and online platforms. People can still create deceptive content without this information (or can remove it), but they cannot easily fake or alter this information, making it an important resource to build trust. As adoption of the standard increases, this information can accompany content through its lifecycle of sharing, modification, and reuse. Over time, we believe this kind of metadata will be something people come to expect, filling a crucial gap in digital content authenticity practices.

- **New Tools to Identify Content Created by our Services.**

- **Audio Watermarking.** We incorporated audio watermarking into Voice Engine, our custom voice model, currently in research preview, and continued our research in related areas.
- **C2PA Metadata.** We believe it's important for people to identify content created with our tools. To support this, we developed technology that

embeds C2PA metadata into all [images](#) generated by DALL·E 3 and videos produced by Sora, providing clear attribution.

- **Research on Additional Solutions.** We conducted research on a variety of provenance, watermarking, metadata, and classifier solutions, such as gathering feedback on the efficacy of image detection tools through our [Researcher Access Program](#). We described this research in a [case study we did with the Partnership on AI](#) and in an [August 2024 update](#).

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*

- **Published System Cards.** We regularly publish detailed reports, called system cards, for new frontier AI systems that we deployed. These system cards aim to inform readers about key factors impacting the system's behavior, especially in areas pertinent for responsible usage including discussions of safety evaluations conducted, limitations in performance that have implications for the domains of appropriate use, discussions of the model's effects on societal risks such as fairness and bias, and the results of adversarial testing conducted. For example, since signing the voluntary commitments, we published a system card prior to releasing [DALL·E 3](#) and for [GPT-4's vision capabilities](#) in ChatGPT. Other published system cards included:

- [GPT-4o \(August 2024\)](#)
- [o1 \(December 2024\)](#)
- [Sora \(December 2024\)](#)

7) *Prioritize research on societal risks posed by AI systems, including on avoiding harmful bias and discrimination, and protecting privacy*

- **Societal Resilience Fund.** To drive adoption and understanding of provenance standards, we joined Microsoft in [launching a societal resilience fund](#). This \$2 million fund is supporting AI education and understanding, including through

organizations like [Older Adults Technology Services from AARP](#), [International IDEA](#), [Partnership on AI](#) and [Coalition for Content Provenance and Authenticity \(C2PA\)](#).

- **Democratic Inputs to AI Grant Program.** In May 2023, our nonprofit organization OpenAI, Inc., announced the [Democratic Inputs to AI grant program](#). We then awarded \$100,000 to 10 teams out of nearly 1000 applicants to design, build, and test ideas that use democratic methods to decide the rules that govern AI systems. Throughout, the teams tackled challenges like recruiting diverse participants across the digital divide, producing a coherent output that represents diverse viewpoints, and designing processes with sufficient transparency to be trusted by the public. In January, 2024, we [summarized](#) the innovations and outlined our learnings.
- **Research on Evaluating Fairness.** On [October 15, 2024](#), we released a research paper that analyzed how ChatGPT responded to users based on their name, using language model research assistants to protect privacy, acknowledging that our study had limitations. While previous research has focused on third-person fairness, e.g., where institutions use AI to make decisions about others, this study examined first-person fairness, or how biases affect users directly in ChatGPT.
- **Trust and Safety Collaboration.** OpenAI leaders, including from our Product Policy and Investigations teams, participated and spoke at [TrustCon 2024](#), a global conference dedicated to trust and safety professionals, and other conferences and events to collaborate and share best practices.
- **Updated Safety Site and Usage Policies.** In January 2024, we updated our [usage policies](#) to be more readable and provide more service-specific guidance. We also updated our [site](#) that shares our safety practices - see <https://openai.com/safety/>.
- **Updated Privacy Center.** In January 2024, we updated our [privacy center](#) that outlines our privacy policies and practices (see our privacy policy updated in [November 2024](#) and enterprise privacy updated in [October 2024](#)), including our portal for requests from users of ChatGPT, DALL·E, and other OpenAI services meant for individuals.

- **Protecting Children.** As described in our [May 2024 Safety Update](#), we've built strong default guardrails and safety measures into ChatGPT and DALL·E that mitigate potential harms to children. In 2023, [we partnered with Thorn's Safer](#) to detect, review and report child sexual abuse material (CSAM) to the National Center for Missing and Exploited Children if users attempt to upload it to our image tools. In [April 2024](#), we joined other AI companies in committing to implement robust child safety measures as articulated in Safety by Design principles led by Thorn and All Tech is Human, ensuring that child safety is prioritized at every stage in the development and use of AI. In [September 2024](#), as recognized by the White House, we joined other AI companies in an effort to reduce AI-generated image-based sexual abuse by taking a variety of actions against NCII or CSAM.

8) Develop and deploy frontier AI systems to help address society's greatest challenges

Support R&D to help meet society's greatest challenges:

- **Cyber Threat Reports.** Since July 2023, we disrupted several covert influence operations, and shared threat intelligence with government, civil society, and industry stakeholders. We also published our findings to amplify efforts to address these challenges. See [Disrupting malicious uses of AI by state-affiliated threat actors](#) (Feb 14, 2024); [Disrupting a covert Iranian influence operation](#) (Aug 16, 2024)
- **Advancing Science.** We are working with government agencies to advance the science of AI safety. In [July 2024](#), we announced a partnership with Los Alamos National Laboratory, one of the United States' leading national laboratories, to study how AI can be used safely by scientists in laboratory settings to advance bioscientific research. In [announcing our o1 preview](#) in September 2024, we shared that the enhanced reasoning capabilities of o1 may be particularly useful in tackling complex problems in science, coding, math, and similar fields. For example, o1 can be used by healthcare researchers to annotate cell sequencing data, by physicists to generate complicated mathematical formulas needed for quantum optics, and by developers in all fields to build and execute multi-step workflows.
- **Hackathon to Accelerate Clean Energy Deployment.** In June 2024, OpenAI joined other industry partners in supporting a [hackathon](#) where OpenAI provided

API credits and technical mentorship, and took part in the judging process. The U.S. Department of Energy, California Secretary of Government Operations, and approximately 100 engineers, designers and product leaders explored how AI can help bring more clean energy projects online.

Initiatives that Foster Education and Training of Students and Workers to prosper from benefits of AI

- **OpenAI Academies.** In September 2024, we [launched](#) OpenAI Academies, a new initiative to invest in developers and organizations leveraging AI to help solve hard problems and catalyze economic growth in their communities. This effort was designed to ensure that the transformative potential of artificial intelligence is accessible and beneficial to diverse communities worldwide.
- **AI Ethics Council.** In Dec 2023 at the Hope Global Forums annual meeting in Atlanta, Open AI CEO Sam Altman and Operation HOPE CEO John Hope Bryant announced the formation of an [AI Ethics Council](#) to ensure that traditionally underrepresented communities would have a voice in the evolution of AI and vast participation in the economic opportunities afforded by AI. The Council held its inaugural meeting in June 2024 and later created the AI Literacy Pipeline to Prosperity Project, an Atlanta-based initiative, designed to give underserved populations the tools to participate meaningfully in the AI economy and be a driver of global innovation and job creation.
- **Common Sense Media Partnership.** In [January 2024](#), we partnered with Common Sense Media, the nation's leading advocacy group for children and families, to help realize the potential of AI for teens and families and minimize the risks, initially collaborating on AI guidelines and education materials for parents, educators and young people, as well as a curation of family-friendly GPTs in the GPT Store based on Common Sense ratings and standards. This work included in [November 2024](#) jointly launching "ChatGPT Foundations for K-12 Educators," a free course designed to help teachers understand and responsibly implement the basics of AI into their work in the classroom. Free for all educators and school districts, the one-hour, eight-lesson course provides educators with essential knowledge about AI and approaches for ensuring student safety and privacy.

Helping citizens understand the nature, capabilities, limitations and impact of AI

- **Sharing Stories.** On our website, OpenAI shares [case studies](#) of how

organizations of all sizes are using AI technology to advance their goals. These stories include:

- [Indeed](#), which is using OpenAI through nearly a dozen products to deliver more personalized and compelling experiences to help job seekers discover new opportunities and employers hire faster.
- [Arizona State University](#), which is enhancing educational outcomes by integrating ChatGPT Edu into projects across teaching, research, and operations.
- [Moderna](#), which is partnering with OpenAI to accelerate the development of life-saving treatments.
- **OpenAI Community Forum.** Throughout the course of the past year, the [OpenAI Forum](#) held events bringing together domain experts, students and AI leaders to discuss and collaborate on the present and future of AI. The Forum features events such as in-person meet-ups highlighting technical talks, dinner-mixers at OpenAI, educational webinars and expert roundtable conversations, and opportunity for members (including OpenAI researchers) to network and cross pollinate ideas.
- **Accessibility Advances.** Building upon our work with [Be My Eyes](#) on AI powered object recognition capabilities, OpenAI continued coordination with the accessibility community enabling people to have a greater degree of independence in their lives. For example, we [enabled](#) chats to be read aloud; rolled out video, screen share, and image upload capabilities in advanced voice mode; and provided an [experimental new launch](#) of 1-800-ChatGPT to enable wider access to ChatGPT via phone call or WhatsApp message.

Munich Security Conference AI Elections Accord

As your letter referenced, last fall OpenAI provided an [update](#) on the Accord commitments. In addition to that progress update, we call your attention to our answers above (particularly our societal resilience and public awareness efforts) and the following:

November 8, 2024 Update to our [Elections Blog](#) - *How OpenAI is Approaching 2024 Worldwide Elections*.

- **Elevating authoritative sources of information.** Throughout 2024, we worked to elevate reliable sources of election information within ChatGPT. Through our collaboration with the National Association of Secretaries of State (NASS), we directed people asking ChatGPT specific questions about voting in the U.S., like where or how to vote, to [CanIVote.org](#). In the month leading up to the election, roughly 1 million ChatGPT responses directed people to [CanIVote.org](#). Similarly, starting on Election Day in the U.S., people who asked ChatGPT for election results received responses encouraging them to check news sources like the Associated Press and Reuters. Around 2 million ChatGPT responses included this message on Election Day and the day following.
- **Preventing deepfakes.** We applied safety measures to ChatGPT to refuse requests to generate images of real people, including politicians. In the month leading up to Election Day, we estimate that ChatGPT rejected over 250,000 requests to generate DALL·E images of President-elect Trump, Vice President Harris, Vice President-elect Vance, President Biden, and Governor Walz.
- **Disrupting threat actors.** A central part of our global elections work in 2024 was identifying and disrupting attempts to use our tools to generate content used in covert influence operations. In [May](#), we began publicly sharing information on our disruptions, and published additional reports in [August](#) and [October](#). Our teams continued to monitor our services closely in the lead-up to Election Day and did not see evidence of U.S. election-related influence operations attracting viral engagement or building sustained audiences through the use of our models.
- October 31, 2024 Update to our [Elections Blog](#) - *How OpenAI is Approaching 2024 Worldwide Elections*.
 - Sharing that for Election Day in the U.S., people who asked ChatGPT for

election results would receive responses encouraging them to check news sources like the Associated Press and Reuters, or their state or local election board for the most complete and up-to-date information.

We appreciate the opportunity to share updates we made in fulfilling our voluntary commitments and trust these will prove helpful in your efforts.

Chan Park
Head of U.S. and Canada Policy & Partnerships
OpenAI

From: [REDACTED]
To: [REDACTED]
Subject: Fw: Request for Stability AI statement on AI voluntary commitments before January 16th
Date: Tuesday, January 21, 2025 5:31:52 PM

See below.

Get [Outlook for iOS](#)

From: Ana Guillen [REDACTED]
Sent: Tuesday, January 21, 2025 5:09:24 PM
To: [REDACTED]
Subject: Re: Request for Stability AI statement on AI voluntary commitments before January 16th

Hi Abdiaziz,

It was great to chat with you today. Hope you made it home safely. Please find our response below. Let me know if you have any questions.

Thank you for reaching out on these important matters. In May 2024, we provided an update on similar topics to United States Senator Mark R. Warner, and we are now pleased to share further progress with the Brennan Center for Justice at NYU.

Stability AI is committed to preventing the misuse of AI. We strictly prohibit the unlawful use of our models and technology, as well as the creation and misuse of misleading or harmful content.

Since joining the White House Voluntary AI Commitments in September 2023 and signing the AI Election Accord in February 2024, we have made substantial progress in advancing safe and responsible use of our technologies.

Specifically, in alignment with the AI Election Accord commitments to combat the deceptive use of AI in the 2024 elections, we have:

On March 1, 2024, we updated the Stability AI Technology [Acceptable Use Policy](#). Among these updates, we clarified prohibited activities to explicitly include, "generating, promoting, or furthering fraud or the creation or promotion of disinformation."

We've implemented several measures to prevent misuse, including adding product safeguards for requests through our application programming interface (API), such as prompt text classifiers to prevent misuse related to election deceptive content. We have also created mechanisms to detect and prevent image uploads of key politicians as well as to stylize image outputs when prompted for synthetic public figures images such as politicians. In addition, Stability AI's content moderation team has established procedures to

identify election misinformation prompts and take action, which can include removing those users from our platform.

Along with our internal focus on preventing the misuse of AI, Stability AI is also a member of the Content Authenticity Initiative and is working alongside the organization to promote provenance and authentication of AI-generated content.

In accordance with the United States government's [Voluntary AI Commitments](#), we have made significant progress across many areas, aligned with the three key themes outlined in the commitments:

Ensuring Products are Safe Before Introducing Them to the Public:

Stability AI conducts rigorous red teaming, both internally and regularly engaging with external vendors, to safeguard our models against severe harms. We established an internal red teaming program in June 2024 as part of our commitment to responsible AI practices.

We have established strong collaborations across industry and government including:
April 2024 - We attended NCMEC roundtable with other tech companies, and law enforcement, in our efforts to combat child sexual exploitation.

April 2024 - We [announced](#) our commitment to join Thorn and All Tech Is Human to enact child safety commitments for Gen AI through Safety by Design.

July 2024 - We [announced](#) our partnership Internet Watch Foundation (IWF), to tackle the creation of AI generated child sexual abuse imagery online. Stability AI is proud to be the first AI organization to join IWF's global network of organizations dedicated to making the internet safer for all users, especially children.

July 2024 - We attended TrustCon, where we met with and learned from Trust and Safety experts and peers in the industry regarding key safety issues impacting Gen AI.

July 2024 - We joined Tech Coalition's Pathways program for expert advice, resources and opportunities to further build capacity to combat online child sexual exploitation and abuse.

Building Systems that Put Security First:

In March 2024, we hired our first Chief Information Security Officer (CISO) to lead the development and implementation of our information security strategy, ensuring the protection of our data, systems, and products from cyber threats. They've since built a team that is responsible for various security functions such as continuous monitoring, detections, incident response, product security, vulnerability management, and infrastructure security.

Additionally, Stability AI's vulnerability management program provides vulnerability coverage for products, internal services, third-party services, and Stability AI infrastructure. The vulnerability management program defines processes and procedures for all aspects of vulnerability management, such as identification, discovery, remediation, and reporting.

Stability AI also contributes to AI-specific industry security initiatives such as the [Cybersecurity and Infrastructure Security Agencies \(CISA\) Joint Cyber Defense Collaborative](#) (JCDC), which help strengthen the AI industry's identification and reporting capabilities.

Earning the Public's Trust:

In January 2024, we established a Product Integrity Team comprising operations specialists and engineers to implement mechanisms for monitoring, detecting, and removing users engaging in severely harmful prompts on our platform.

In March 2024, we introduced [Stable Safety](#) on our website as a public resource for users and the broader AI community. This initiative outlines our commitment to Safe AI and provides a mechanism for reporting product misuse or appealing enforcement actions.

In April 2024, we launched a content moderation program. Our Product Integrity engineers have also developed machine learning safeguards to monitor input prompts and generated content, successfully mitigating severe harms such as election disinformation and CSAM.

In September 2024, we introduced our internal Responsible AI Development Policy, outlining clear guidelines for the ethical development and deployment of artificial intelligence (AI) systems across Stability AI.

Additionally, Stability AI proactively implements content credentials to help users and content distribution platforms better identify AI-generated content. Images and video generated through our application programming interface (API) are tagged with metadata to indicate the content was produced with an AI tool. In partnership with the Content Authenticity Initiative (CAI) led by Adobe, we adopted the Coalition for Content Provenance and Authenticity (C2PA) standard for metadata. This metadata includes the model name and version number used to generate the content. Once the metadata is generated, it is digitally sealed with a cryptographic Stability AI certificate and stored in the file.

On Mon, Jan 20, 2025 at 3:37 PM [REDACTED] wrote:

[REDACTED]
Get [Outlook for iOS](#)

From: Ana Guillen [REDACTED]
Sent: Monday, January 20, 2025 6:35:57 PM
To: [REDACTED]
Subject: Re: Request for Stability AI statement on AI voluntary commitments before January 16th

That works. What is the best number to reach you?

On Mon, Jan 20, 2025 at 3:33 PM [REDACTED] wrote:

Absolutely let's do 10?

Get [Outlook for iOS](#)

From: Ana Guillen [REDACTED]
Sent: Monday, January 20, 2025 6:26:12 PM
To: [REDACTED]
Subject: Re: Request for Stability AI statement on AI voluntary commitments before January 16th

Sure thing! What time would be good for you? Can I wait to send our response till we connect? Thanks!

On Mon, Jan 20, 2025 at 3:24 PM [REDACTED] wrote:

Ana - so sorry I'm out today. Can we speak tomorrow?

Get [Outlook for iOS](#)

From: Ana Guillen [REDACTED]
Sent: Monday, January 20, 2025 6:01:56 PM
To: [REDACTED]
Subject: Re: Request for Stability AI statement on AI voluntary commitments before January 16th

Hi Abdiaziz,

Just checking back here? I know your deadline is today and it's getting late on the East Coast.

Thanks,
Ana

On Mon, Jan 20, 2025 at 11:28 AM Ana Guillen [REDACTED] wrote:
Hi Abdiaziz,

I have a response to your request ready to send to you. Prior to doing so, might you have 5 minutes for a quick chat? I'd like to provide some off record context. I am available at 310-418-7965 all day today.

Thanks so much,
Ana

On Thu, Jan 16, 2025 at 9:54 AM [REDACTED]
wrote:

Ana -

We are unsure of release - though aiming for some point next month. We're working with our communications team to figure out roll out, so not much I can share at this point.

Get [Outlook for iOS](#)

From: Ana Guillen [REDACTED]
Sent: Thursday, January 16, 2025 12:50:42 PM
To: [REDACTED]
Subject: Re: Request for Stability AI statement on AI voluntary commitments before January 16th

That would be great. Thank you! Are you also able to share any information about when your report will be released and how it is distributed?

On Thu, Jan 16, 2025 at 3:12 AM [REDACTED]
[REDACTED] wrote:

Hi - it is today. Sorry for the confusion! I can extend until Monday if that is necessary.

Get [Outlook for iOS](#)

From: Ana Guillen [REDACTED]
Sent: Thursday, January 16, 2025 3:25:15 AM
To: [REDACTED]
Subject: Re: Request for Stability AI statement on AI voluntary commitments before January 16th

Hi Abdiaziz,

Thanks for reaching out. Your deadline below is listed as Monday, January 16 2025. Can you confirm whether your deadline is Thursday, January 16 or Monday, January 20?

Thanks again,
Ana

On Wed, Dec 18, 2024 at 1:15 PM [REDACTED]
[REDACTED] wrote:

To Whom It May Concern at Stability AI,

In September 2023 and February 2024, respectively, your company became a signatory of the White House Voluntary AI Commitments and the Munich Security Conference AI Election Accord.

We are contacting you today to request a statement on your company's progress toward meeting the voluntary AI commitments outlined in these agreements.

In February 2025, the Brennan Center for Justice at NYU will publish a report analyzing the progress that all signatory companies have made toward meeting these commitments. We have collected information on your company's progress through the resources provided as part of the following:

White House Voluntary AI Commitments

- No materials found

Munich Security Conference AI Election Accord


- Written response to May 2024 Senator Mark R. Warner's letter request ([link](#))

To help us understand your company's efforts to meet these commitments, please provide a written statement outlining the specific actions your company has taken to align with them. If there have been updates or new developments since the commitments were made, please include them. Additionally, please link to any publicly available resources that could help us better understand how your company has met these commitments.

We will be collecting responses from the signatory companies **before the end of Monday, January 16th**. It would benefit our analysis if you could observe the numbered commitments made in each agreement and speak on your company's specific efforts toward each numbered commitment.

We will be completing our analysis and evaluation of fulfillment based on our review of the information provided in your written responses listed above and any responses provided to this email.

Best,



Abdiaziz Ahmed

Technology Policy Strategist

Brennan Center for Justice



January 16, 2025

Mr. Abdiaziz Ahmed
Brennan Center for Justice
1140 Connecticut Ave NW #1150
Washington, DC 20036

Dear Abdiaziz Ahmed,

Thank you for your request dated December 18, 2024 regarding our progress toward meeting the voluntary AI commitments. Since signing the Accord in February 2024, we have continued investing to combat deceptive AI and protect election integrity on TikTok for our communities around the world, including in the EU, UK, Mexico, and US elections.

As we continue investing in AI detection and labeling efforts, we have made significant progress. In May 2024, we were proud to be the first video sharing or social media platform to begin implementing the Coalition for Content Provenance and Authenticity (“C2PA”)’s Content Credentials technology. This means we can read Content Credentials in order to recognize and label AI-generated content (AIGC) on images and videos, and attach Content Credentials to TikTok content to help anyone using C2PA’s Verify tool identify AIGC that was made on TikTok. We automatically disclose TikTok AI effects, and offer a tool for people to label their AI-generated content. This investment in detection and labeling has paid off – of the videos we removed for violating our Integrity & Authenticity policies in Q3 2024, over 97% were removed proactively, meaning we identified and removed the video before it was reported.

We continuously strengthen our policies and engage with a wide range of experts as part of doing so, including our Safety and Content Advisory Councils. Over the last year, that has included further tightening relevant policies on [harmful misinformation](#), [foreign influence attempts](#), [hate speech](#), and [misleading AI-generated content](#) based on some of their input. An example is updating our AIGC policies to more explicitly prohibit misleading AIGC that falsely shows a public figure being degraded, harassed, or politically endorsed or condemned by an individual or group, as well as AIGC that falsely shows a crisis event or is made to seem as if it comes from an authoritative source, such as a reputable news organization. In addition, we strengthened our policies against election interference by restricting the For You Feed eligibility and advertising capabilities of state-controlled media who attempt to target foreign audiences on current affairs. We’ve maintained policies against misleading manipulated media for years, and for nearly two years we have required creators to label AIGC that contains realistic scenes and prohibited AIGC that falsely depicts public officials making endorsements.

Partnerships with experts and continuous investments into transparency resources demonstrate our commitment to evolve our approach, share our progress, and help educate our community. We launched a new [Harmful Misinformation Guide](#) in our Safety Center to help our community navigate online misinformation, teamed-up with peers to support a new [AI Literacy initiative](#) from the National Association for Media Literacy Education, and developed new media literacy videos with guidance from experts that accumulated over 80M views. We also launched an awareness campaign in select markets that promotes using generative AI in a creative and safe way. To help us regularly tap into insights from outside of TikTok as the US election approached, we formed a [US Elections Integrity Advisory Group](#) composed of experts in AI, elections and civic integrity, hate, violent extremism, and voter protection issues whom we met with regularly. This was in addition to the on-going engagement with experts to respond quickly to harmful content, including misleading AIGC, through our Community Partners channel where they can report potentially violative content directly to our teams.



Our transparency efforts go beyond Election Day. We've updated our Election Center, search banners, and content labels to make election results from the Associated Press available to people who engage with election-related content. Our Election Center has received more than 60 million views since we launched it in January 2024. Lastly, we will continue to share progress on our US election efforts—including our removal of AI-generated content that violates our policies — in the [US Election Integrity Hub](#) in our Transparency Center, and continue to share data on our wider content moderation efforts in our Community Guidelines Enforcement [reports](#).

Sincerely,

A handwritten signature in black ink that reads "Lisa Hayes". The signature is written in a cursive, flowing style.

Lisa Hayes
Head of Safety of Public Policy, Americas, TikTok