August 1, 2023

Laura Greene, Co-Lead
Terrence Tao, Co-Lead
President's Council of Advisors on Science and Technology
PCAST Working Group on Generative AI

Dear Dr. Greene and Dr. Tao:

On behalf of The Leadership Conference on Civil and Human Rights (The Leadership Conference) and Anti-Defamation League (ADL), we write in response to the President's Council of Advisors on Science and Technology's request for input on generative artificial intelligence (AI).[i]

The Leadership Conference, a coalition charged by its diverse membership of more than 240 national organizations to promote and protect the rights of all persons in the United States, works to ensure that civil and human rights, equal opportunity, and democratic participation are at the center of communication and technology policy debates. ADL has been a leader in the fight against hate and antisemitism for over a century. Launched in 2017, the ADL Center for Tech and Society (CTS) provides unique expertise because of ADL's work at the intersection of civil rights, extremism, and tech. ADL is rooted in and draws upon the lived experience of a community relentlessly targeted online by extremists, bigots, and other bad actors.

As generative AI becomes more prevalent in society, we must consider both the benefits and challenges of incorporating these tools into daily life. In response to the PCAST's request, this submission addresses the following: (1) ensuring reliable access to verifiable, trustworthy information; (2) addressing the use of AI by malicious actors; (3) technologies, policies, and infrastructure for detecting and countering AI-generated disinformation; (4) ensuring public engagement with elected representatives amid AI-generated noise; and (5) developing skills to identify AI-generated misinformation, impersonation, and manipulation.

I. Ensuring Reliable Access to Verifiable, Trustworthy Information

It is not enough for verifiable, trustworthy information simply to be available to the public; rather, ensuring that this information is easily and widely accessible is essential to the preservation of democracy. As AI technologies continue to rapidly advance, verifying content is becoming increasingly difficult for end users.[ii] Therefore, it is essential for government, academia, industry, and civil society to collaborate and invest in the development and enhancement of technologies that aid in media verification and authentication. Our organizations have consistently advocated for clear disclosure mechanisms to identify

artificially generated content, especially by the programs that produce this content.[iii] Although numerous approaches for such disclosure may exist, the key is to ensure clarity for and comprehension by end users and subsequent viewers of the content. Especially in situations where artificially generated media can deceive users who rely on that information for civic participation, the harmful impacts of misleading or false AI-generated content pose too significant of a threat to democracy to ignore. We recommend that the government work to develop an effective means to authenticate information. For example, it should consider how watermarking features and advanced cryptographic technologies can support users in understanding whether they are looking at human or AI-generated content.[iv]

In addition to encouraging industry to implement more robust verification technologies, we recommend that the government support collaboration efforts by stakeholders across disciplines. Through active collaboration, society can benefit from up-to-date information, the development and implementation of best practices for information verification, and the most informed policymaking.[v]

II. Addressing the Use of AI by Malicious Actors

With the increased use of generative AI comes increased concern about the ways it can be leveraged by bad actors to engage in online hate, harassment, and abuse.[vi] Notably, some generative AI programs disallow prompts that explicitly request hateful or harassing content.[vii] Still, anti-hate policies are often either inadequate or easy for malicious actors to circumvent.[viii] Addressing the use of AI by malicious actors requires a multi-faceted approach that employs levers across policy, industry, government, academia, and civil society.

Because the widespread use of generative AI is, in many ways, a case of first impression, ADL urges legislators to consider the best ways to implement accountability measures for the malicious use of AI-generated content. Where relevant legislation already exists, statutes may need to be updated to include AI-generated content. For example, while some legal recourse at the state and federal level exists for nonconsensual distribution of intimate imagery, cyberstalking, and doxing, statutes may need to be updated to address AI-generated harm, which ultimately can have consequences similar to non-AI counterparts. Legislators should consider updating laws concerning defamation, consumer protection, election interference, and other forms of digital abuse to disincentivize bad actors from using generative AI to facilitate harm and adjust penalties appropriately when generative AI exacerbates or amplifies harms.

While lawmakers must pass laws and implement regulations that compel transparency, impose penalties, and create redress for illegal and abusive activity using AI-generated content, industry actors must do their part to mitigate harms before they arise. As industry actors are responsible for having brought generative AI technologies into the stream of commerce, the companies that develop these programs are best positioned to create and voluntarily implement safeguards that prevent online harms by bad actors from occurring in the first place. As ADL notes in its NTIA submission on AI accountability, product features designed to authenticate information or disclose its artificiality (e.g., watermarking) may mitigate some of the harms of AI-generated content after its creation and dissemination, but industry must also make proactive efforts to prevent harm from occurring through red-teaming, bias minimization, and risk assessments.[ix]

Additionally, industry may consider using AI systems to strengthen trust and safety efforts. By improving AI-powered content moderation systems that support the efforts of human moderators, social media platforms may be able to address issues of scale while effectively identifying and removing malicious AI-generated content before it causes grave harm.[x] Importantly, as tech companies employ AI tools to improve their systems, they must be transparent about these practices. For example, tech companies should publish public-facing transparency reports, provide independent researcher data access, and submit to third-party assessments and audits.[xi] They must also make sure to continuously monitor these tools for bias - current content moderation tools used by social media companies have been found to erroneously flag the content of activists and minority groups.[xii] It is critical that any system is designed in an equitable way, so that it does not discriminate against the very people it is intended to protect.

III. Technologies, Policies, and Infrastructure for Detecting and Countering AI-Generated Hate and Disinformation

Experts anticipate that the rapid improvement of AI-technology will make false, hateful, and conspiratorial information more believable and tougher for the average person to dismiss as deceptive,[xiii] which means it is more important than ever for major social media platforms to implement tools to detect and policies to counter AI-generated disinformation.

Some companies, such as TikTok, have already put in place policies explicitly intended to moderate AI-generated content, while other companies, such as Meta, have not.[xiv] It is crucial for companies like Meta to develop and implement strong policies governing manipulated content. As AI-generated content becomes more popular, even platforms that have some policies in place should consider ways to strengthen their policies. For example, TikTok requires AI-generated content to be disclosed and clearly labeled using a sticker or caption.[xv] While this is a start, labeling mechanisms are currently left to the discretion of the user. Ultimately, labeling mechanisms and potentially other measures that can address the spread of manipulated content should automatically be integrated across the platform, something that has already been done across major social media platforms for other types of content.[xvi] In tandem with this, platforms should publicly verify election officials' accounts and other authoritative sources of election information, such as the National Association of Secretaries of State.[xvii] Finally, platforms can promote greater transparency into the enforcement and effectiveness of these policies by creating a database of all publicly available AI-generated content related to elections and political issues.

IV. Ensuring Public Engagement with Elected Representatives Amid AI-Generated Noise

A. *Strengthen online platforms' responsibility*

Online companies have a responsibility to reduce the impact of AI-generated noise on social media to ensure public engagement with elected representatives is authentic. First, they should tune algorithms to prioritize genuine engagement over AI-generated spam content. This may involve de-amplifying AI-generated content and prioritizing genuine human engagement. Second, online companies should strengthen their Trust and Safety and Responsible AI teams. Concerningly, these teams have actually been reduced in size or largely eliminated this past year.[xviii]

The government also has a role to play to push online companies to take responsibility when it comes to AI-generated noise. The Leadership Conference has previously urged Congress to work with federal

agencies to implement the Biden administration's Blueprint for an AI Bill of Rights. As an example, Congress could work with the Federal Elections Commission to use existing authority to regulate deliberately misleading campaign communications generated using AI - a reform contemplated by the AI Bill of Rights' emphasis on transparency.[xix]

B. *Encourage public education and awareness*

Today, policymakers across all levels of government are grappling with the ways AI affects people's daily lives. The Biden administration can and should play an important role as a convenor, source of expertise, and driver of public education.[xx] The administration should convene diverse experts to strengthen public dialogue about effective approaches to identifying and mitigating the influence of AI-generated content on public discourse. Finally, the administration must consider ways to make it easier for constituents to engage with lawmakers off social media platforms, while also building up capabilities for digital engagement beyond merely working within current systems and platforms.

Finally, the administration can promote active civic participation and critical evaluation of political information by drawing upon existing strategies used to educate the public about AI-generated hate, harassment, and dis- and misinformation. These strategies include supporting media literacy efforts in K-12 education,[xxi] and promoting high-quality journalism to counter disinformation narratives.[xxii]

V. Developing Skills to Identify AI-Generated Misinformation, Impersonation, and Manipulation

While many of the proposed courses of action in this submission concern legal and industry approaches to combating AI harms, educators also have a significant role to play in mitigating the harmful impacts of harassment and disinformation and promoting digital literacy. The administration should consider supporting trainings, providing resources, and implementing grant programs across various sectors to equip the public with educational materials on combating the impacts of AI-generated hate and disinformation. As AI becomes an aspect of everyday life, media literacy and disinformation resilience skills may become a necessary aspect of core curricula in primary, secondary, and college education: pushing for the incorporation of media literacy programs and encouraging the development of these skills early on may inoculate communities against the most pernicious impacts of mis- and disinformation, AI-generated or otherwise.[xxiii] Additionally, collaborations between educational institutions, researchers, and industry experts can lay the groundwork for initiatives that increase awareness of the risks of AI-driven hate and disinformation. By promoting critical thinking and responsible media consumption in AI use, educators can empower individuals and communities to navigate the ever-evolving digital landscape and to be more successful in efforts to discern fact from fiction.

Thank you for considering our views. Please do not hesitate to contact Dave Toomey, voting rights and technology fellow, The Leadership Conference on Civil and Human Rights at toomey@civilrights.org or Lauren Krapf, lead counsel, Center for Technology and Society, Anti-Defamation League, at lkrapf@adl.org with any questions.

Sincerely,

Anti-Defamation League

[i] https://www.whitehouse.gov/pcast/briefing-room/2023/05/13/pcast-working-group-on-generative-ai-invites-public-input/

[ii] Clare Duffy, "With the Rise of AI, Social Media Platforms Could Face Perfect Storm of Misinformation in 2024," CNN (July 17, 2023), https://www.cnn.com/2023/07/17/tech/ai-generated-election-misinformation-social-media/index.html.

[iii] Anti-Defamation League, "Submission for NTIA's AI Accountability Policy Request for

Comment, Docket No. NTIA-2023-0005" (June 12, 2023), https://www.adl.org/sites/default/files/pdfs/2023-06/CTS-Comment-to-NTIA.pdf.

[iv] *Id.* ADL specifically cites Microsoft's pledge to cryptographically watermark AI-generated outputs as an example of a company's efforts to prevent its outputs from being used to deceive or mislead others.

[v] While some of these interactions may eventually be impeded by Judge Doughty's *Missouri v. Biden* injunction, prohibitions on the Biden administration's interactions with social media platforms and civil society organizations are not currently in effect because of the temporary hold in place. Moreover, as the injunction may have carved out an exception for interactions related to limiting voter suppression, its reinstatement should not preclude the communications that are necessary to ensure that artificial intelligence is not leveraged in ways that could discourage democratic engagement.

[vi] Anti-Defamation League, "Six Pressing Questions We Must Ask About Generative AI" (May 14, 2023)

[vii] *Id.*

[viii] *Id.*

[ix] Anti-Defamation League, "Submission for NTIA's AI Accountability Policy Request for

Comment, Docket No. NTIA-2023-0005" (June 12, 2023), https://www.adl.org/sites/default/files/pdfs/2023-06/CTS-Comment-to-NTIA.pdf; The Leadership Conference on Civil and Human Rights, "Comments to OSTP on National AI Strategy" (July 7, 2023), https://civilrights.org/resource/leadership-conference-comments-to-ostp-on-national-ai-strategy/.

[x] *Id.*

[xi] *Id.* The ADL Center for Tech & Society has long advocated for increased platform transparency and accountability; most of our publications invoke the need for one or the other as a safeguard against bad actors, deliberate indifference, and other harms.

[xii] Merlyna Lim and Ghadah Alrasheed, "Beyond a Technical Bug: Biased Algorithms and Moderation Are Censoring Activists on Social Media," Carleton University Newsroom (May 16, 2021), https://newsroom.carleton.ca/story/biased-algorithms-moderation-censoring-activists/

[xiii] Clare Duffy, "With the Rise of AI, Social Media Platforms Could Face Perfect Storm of Misinformation in 2024, CNN (July 17, 2023), https://www.cnn.com/2023/07/17/tech/ai-generated-election-misinformation-social-media/index.html.

[xiv] Irene Benedicto, "AI-Generated Election Content Is Here, and the Social Networks Are Not Prepared," Forbes (July 6, 2023), https://www.forbes.com/sites/irenebenedicto/2023/07/05/ai-generated-2024-election-content-social-media/?sh=757fed6137c1.

[xv] James Vincent, "TikTok Bans Deepfakes of Nonpublic Figures and Fake Endorsements in Rule Refresh," The Verge (Mar. 21, 2023), https://www.theverge.com/2023/3/21/23648099/tiktok-content-moderation-rules-deepfakes-ai.

[xvi] Ivan Mehta, "Twitter Will Now Show Labels on Tweets With Reduced Visibility," TechCrunch (Apr. 25, 2023), https://techcrunch.com/2023/04/25/twitter-will-now-show-labels-on-tweets-with-reduced-visiblity/.

[xvii] Mekela Panditharatne and Noah Giansiracusa, "How AI Puts Elections at Risk - And the Needed Safeguards," Brennan Center for Justice (June 13, 2023), https://www.brennancenter.org/our-work/analysis-opinion/how-ai-puts-elections-risk-and-needed-safeguards.

[xviii] Clare Duffy, "'It's an Especially Bad Time:' Tech Layoffs Are Hitting Ethics and Safety Teams," CNN (Apr. 6, 2023), https://www.cnn.com/2023/04/06/tech/tech-layoffs-platform-safety/index.html.

[xix] Press Release, Public Citizen, Public Citizen Submits Petition to Federal Election Commission Calling for Rulemaking on AI (July 13, 2023), https://www.citizen.org/news/public-citizen-submits-petition-to-federal-election-commission-calling-for-rulemaking-on-ai/.

[xx] https://civilrights.org/resource/leadership-conference-comments-to-ostp-on-national-ai-strategy/

[xxi] Tiffany Hsu, "When Teens Find Misinformation, These Teachers Are Ready," New York Times (Sept. 8, 2022), https://www.nytimes.com/2022/09/08/technology/misinformation-students-media-literacy.html.

[xxii] Darrell M. West, "How to Combat Fake News and Disinformation," Brookings (Dec. 18, 2017), https://www.brookings.edu/articles/how-to-combat-fake-news-and-disinformation/.

[xxiii] PEN America, "Building Resilience: Identifying Community Solutions to Targeted Disinformation." https://pen.org/report/building-resilience/