MICHAEL F. BENNET
COLORADO

COMMITTEES:
AGRICULTURE, NUTRITION, AND FORESTRY
FINANCE
INTELLIGENCE

United States Senate
WASHINGTON, DC 20510–0609

WASHINGTON, DC:
261 RUSSELL SENATE OFFICE BUILDING
WASHINGTON, DC 20510
(202) 224-5852

COLORADO:
CESAR E. CHAVEZ BUILDING
1244 SPEER BOULEVARD
DENVER, CO 80204
(303) 455-7600

http://www.bennet.senate.gov

June 29, 2023

Mr. Mark Zuckerberg
Chairman and Chief Executive Officer
Meta
1 Hacker Way
Menlo Park, CA 94025

Mr. Elon Musk
Owner
Twitter
1355 Market Street, Suite 900
San Francisco, CA 94103

Mr. Sam Altman
Chief Executive Officer
OpenAI
3180 18th Street, Suite 100
San Francisco, CA 94110

Mr. Shou Zi Chew
Chief Executive Office
TikTok
5800 Bristol Parkway, Suite 100
Culver City, CA 90230

Mr. Sundar Pichai
Chief Executive Officer
Alphabet Inc. and Google LLC
1600 Amphitheatre Parkway
Mountain View, CA 94043

Mr. Satya Nadella
Chief Executive Officer
Microsoft
One Microsoft Way
Redmond, WA 98052

Mr. Emad Mostaque
Chief Executive Officer
Stability AI
88 Notting Hill Gate
London, U.K. W11 3HP

Mr. David Holz
Chief Executive Officer
Midjourney
611 Gateway Boulevard, Suite 120
San Francisco, CA 94080

Dr. Dario Amodei
Chief Executive Officer
Anthropic
548 Market Street
San Francisco, CA 94104

Dear Mr. Zuckerberg, Mr. Musk, Mr. Altman, Mr. Chew, Mr. Pichai, Mr. Nadella, Mr. Mostaque, Mr. Holz, and Dr. Amodei:

I write with concerns about your current identification and disclosure policies for content generated by artificial intelligence (AI). Americans should know when images or videos are the

product of generative AI models, and platforms and developers have a responsibility to label such content properly. This is especially true for political communication. Fabricated images can derail stock markets, suppress voter turnout, and shake Americans' confidence in the authenticity of campaign material. Continuing to produce and disseminate AI-generated content without clear, easily comprehensible identifiers poses an unacceptable risk to public discourse and electoral integrity.

Online misinformation and disinformation are not new. But the sophistication and scale of these tools has rapidly evolved and outpaced our existing safeguards. In the past, creating plausible deepfakes required significant technical skill; today, generative AI systems have democratized the ability–opening the floodgates to anyone who wants to use or abuse the technology.

We have already seen evidence of generative AI being used to create and share false images.[1] In some instances, these have been relatively benign–such as Pope Francis depicted wearing a large white down jacket.[2] Others are more disturbing. In May, an AI-generated image of a purported explosion at the Pentagon went viral, causing a dip in major stock indices.[3] Fake news accounts recirculated these images alongside real outlets, including *RT*, a Russian state-backed media organization.

The proliferation of AI-generated content poses a particular problem for political communication. In his recent testimony before the Senate Judiciary Committee, OpenAI CEO Sam Altman identified the ability of AI models to provide "one-on-one interactive disinformation" as an area of greatest concern.[4] We have entered the beginning of this era.

In June, the official rapid response Twitter account of Florida Governor Ron DeSantis, a candidate for the 2024 Republican nomination for president, shared images that experts say appear to be AI-generated.[5] Both official and unaffiliated accounts supporting former President Trump have posted AI-generated content targeting his political rivals.[6]

---

[1] Tiffany Hsu and Steven Lee Myers, "Can We No Longer Believe Anything We See?" *New York Times*, April 8, 2023; Clare Duffy, "Puffer coat Pope. Musk on a date with GM CEO. Fake AI 'news' images are fooling social media users," CNN, April 2, 2023.

[2] Daysia Tolentino, "AI-generated images of Pope Francis in puffer jacket fool the internet," NBC News, March 27, 2023.

[3] Shannon Bond, "Fake viral images of an explosion at the Pentagon were probably created by AI," NPR, May 22, 2023.

[4] Cat Zakrzewski, Cristiano Lima and Will Oremus, "CEO behind ChatGPT warns Congress AI could cause 'harm to the world,'" *Washington Post*, May 16, 2023.

[5] Benjy Sarlin and Shelby Talcott, "DeSantis campaign shares fake Trump/Fauci images, prompting new AI fears," *Semafor*, June 8, 2023; Bill McCarthy, "Ron DeSantis ad uses AI-generated photos of Trump, Fauci," *Agence France-Presse*, June 7, 2023.

[6] Aditi Bharade, "Someone made a hyper-realistic deepfake of Ron DeSantis as Michael Scott from 'The Office' wearing women's clothes. It's the latest instance of AI being weaponized to take DeSantis down," *Business Insider*, May 29, 2023; Matthew Loh, "The Trump-DeSantis showdown is now official, and artificial intelligence is right in the middle of it," *Business Insider*, May 25, 2023.

In May, I joined colleagues to introduce the REAL Political Ads Act, which would require a disclaimer on political ads for federal campaigns that use content generated by AI.[7] However, as political media increasingly shifts from regulated television, print, and radio advertising to the free-for-all of social media, broader disclosure requirements must follow.

AI system developers and platforms will have to collaborate to combat the spread of unlabeled AI content. Developers should work to watermark video and images at the time of creation, and platforms should commit to attaching labels and disclosures at the time of distribution. A combined approach is required to deal with this singular threat.

Companies have started taking steps to better identify AI-generated content for users. For example, non-profit organizations, like the Partnership on AI, have released suggested guidelines.[8] Microsoft has committed to watermark AI-generated content, and Google will begin attaching a written disclosure on Google Images.[9] OpenAI's DALL-E 2 adds a watermark to images it generates, and Stable Diffusion embeds watermarks into its content by default.[10] Midjourney, Shutterstock, and Google have committed to embedding metadata indicators in AI-generated content.[11]

However, these policies remain easily bypassed or alarmingly reliant on voluntary compliance. Google's process for labeling AI-generated images from third-party systems depends on self-disclosure.[12] Stable Diffusion's open source structure allows users to circumvent the watermarking code.[13] DALL-E 2's watermarks are inconspicuous and easily removed.[14] And while some platforms –including Meta,[15] Twitter,[16] and TikTok[17]– have existing policies for AI-generated images and video, such content continues to appear on users' feeds.

---

[7] Office of Senator Michael Bennet, "Bennet, Klobuchar, Booker Push to Regulate AI-Generated Content in Political Ads," May 15, 2023.
[8] Partnership on AI, *PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action*, February 27, 2023.
[9] Kyle Wiggers, "Microsoft pledges to watermark AI-generated images and videos," *TechCrunch*, May 23, 2023; Cory Dunton, "Get helpful context with About this image," Google, May 10, 2023.
[10] Benj Edwards, "AI image generation tech can now create life-wrecking deepfakes with ease," *Ars Technica*, December 9, 2022.
[11] IPTC, "Midjourney and Shutterstock AI sign up to use of IPTC Digital Source Type to signal generated AI content," May 11, 2023.
[12] Dunton, "Get helpful context with About this image."
[13] Hany Farid, "ChatGPT and Dall-E Should Watermark Their Results," *Gizmodo*, April 2, 2023.
[14] OpenAI, "How should I credit DALL·E in my work?" accessed June 28, 2023.
[15] Meta, "Manipulated Media," accessed June 28, 2023.
[16] Twitter, "Synthetic and manipulated media policy," April 2023.
[17] TikTok, "Integrity and Authenticity," last updated March 2023.

Platforms must update their policies for a world where everyone has access to generative AI tools. They should require clear, conspicuous labels for AI-generated video and images, and where users fail to comply, should label AI-generated content themselves. Platforms should consider particular rules for official political accounts, and should release regular reports detailing their efforts to identify, label, or remove AI-generated content.

Similarly, generative AI system developers must scrutinize whether their models can be used to manipulate and misinform, and should conduct public risk assessments and create action plans to identify and mitigate these vulnerabilities. We cannot expect users to dive into the metadata of every image in their feeds, nor should platforms force them to guess the authenticity of content shared by political candidates, parties, and their supporters.

Continued inaction endangers our democracy. Generative AI can support new creative endeavors and produce astonishing content, but these benefits cannot come at the cost of corrupting our shared reality.

To that end, I request answers to the following questions by July 31, 2023:

For generative AI developers:

- What technical standards, features, or requirements do you currently employ to watermark or otherwise identify content created using your systems?
    - When were these standards, features, or requirements developed?
    - When were these standards, features, or requirements last updated?
    - What auditing processes, if any, does your organization have in place to evaluate the effectiveness of these standards, features, or requirements?

- What policies do you currently have in place for users that repeatedly violate a watermarking or identifying requirement, either by removing the identifier or avoiding it in some other way?
    - How many accounts, if any, have you suspended or removed for violating a watermarking or identifying requirement?

- What tracking system do you currently have in place, if any, to monitor the distribution of content created using your systems?

- Before deploying a model, what tests or evaluations do you use to estimate potential capabilities relating to misinformation, disinformation, persuasion, and manipulation?

- What processes do you use to estimate risks associated with misinformation, disinformation, persuasion, and manipulation? Under what circumstances would you delay or restrict access to a generative AI system due to concerns about these risks?

- What interoperable standards currently offer the highest degree of provenance assurance?

For social media platforms and search engines:

- Will you commit to removing AI-generated content designed to mislead users?

- What technical processes are currently in place to identify AI-generated content?

- How many pieces of AI-generated content did you identify in 2022 and the first quarter of 2023?
    - Of those identified, how many were removed for violating a policy?
    - If removed, what policy did they violate?
    - If not removed, was a label or other clear identifier affixed?
    - If not labeled, please provide a rationale for declining to do so.

- Do you have specific policies in place for AI-generated content posted by an official political campaign account?
    - If so, what are they?
    - If not, describe why not.

- Do you have specific policies in place for AI-generated content related to campaigns and elections?
    - If so, what are they?
    - If not, describe why not.

I appreciate your attention to this important matter and look forward to your response.

Sincerely,

U.S. Senator Michael F. Bennet