Snap Inc.

The Honorable Mark Warner
United States Senate
703 Hart Senate Office Building
Washington, DC 20510

May 24, 2024

Dear Senator Warner:

Since our earliest days as a company, Snap has brought an intentional and multilayered approach to protecting against threats to democratic processes across our services. Over the years, our teams have implemented a broad range of safeguards and design elements that serve to substantially mitigate risks to civic discourse and electoral processes, and these mitigations are of particular importance in 2024, as more than 50 countries are administering elections around the world.

At a high level, the core elements of our approach include:

- ***Intentional product safeguards and platform design:*** From the outset, Snapchat was designed differently from traditional social media. We've long recognized that the greatest threats from harmful digital disinformation stem from the speed and scale at which some digital platforms enable it to spread. Our platform policies and architecture limit the opportunities for unvetted or unmoderated content to achieve meaningful scale unchecked. Instead, we pre-moderate content before it can be amplified to a large audience, and broadly limit the distribution of news and political information unless it comes from trusted publishers and creators.

- ***Clear and thoughtful policies:*** We've implemented a range of policies that function to advance safety and integrity in the context of high-profile events like elections. Our Community Guidelines expressly prohibit, for example, harmful false information, hate speech, and threats or calls to violence. Our civic integrity policies expressly prohibit election-related misinformation and harmful content including procedural interference (i.e., misinformation related to actual election or civic procedures); participation interference (e.g., content that includes intimidation to personal safety or spreads rumors to deter participation in the electoral or civic process); fraudulent or unlawful participation (i.e., content that encourages people to misrepresent themselves to participate in the civic process or to illegally cast or destroy ballots); and delegitimization

of civic processes (e.g., content aiming to delegitimize democratic institutions on the basis of false or misleading claims about election results).

- ***Diligent approach to political ads:*** We've taken care to adopt rigorous practices to mitigate risks to election integrity in advertising. Most notably, every political ad on Snapchat is human-reviewed *before* it is eligible for placement on our platform, and we work with independent, nonpartisan fact-checking organizations to help assess advertisements for deceptive images or content. Each ad must clearly disclose who paid for it, and under our Political Ad Policies, we don't allow ads to be paid for by foreign governments or any individuals or entities located outside of the country (or multi-nation jurisdiction) where the election is taking place. We believe it's in the public's interest to see which political ads are approved to run and keep a Political Ads Library that includes information about targeting, costs, and other insights. To ensure compliance with all of these processes, our Commercial Content Policies disallow influencers from promoting paid political content outside of traditional ad formats.

- ***Collaborative, coordinated operations:*** Internally, we convene a cross-functional election integrity team, including misinformation, political advertising, and cybersecurity experts, to monitor all relevant developments in connection with elections throughout the world in 2024. The breadth of representation in this group reflects our whole-of-company approach we take to safeguarding platform integrity, with representatives from Trust & Safety, Content Moderation, Engineering, Product, Legal, Policy, Privacy Operations, Security, and others. We've also operationalized a crisis response protocol, to ensure operational agility in the face of high-risk global events. We routinely engage with democracy stakeholders and civil society organizations for advice, research insights, and to hear concerns or receive incident escalations. We also participate in multi-stakeholder initiatives, as we did this year, for example, working with civil society, elections authorities, and fellow industry stakeholders to help shape the Voluntary Election Integrity Guidelines for Technology Companies.

- ***Offering tools and resources to empower Snapchatters:*** As a platform that helps people express themselves and has significant reach with new and first-time voters, we make it a priority to help our community get access to accurate and trusted information about news and world events, including where and how they can vote in their local election. In support of this, we leverage our closed content platform to partner with trusted publishers and creators, who are accountable for presenting factual information to remain eligible for broadcast on Snapchat content surfaces.

Taken together, these pillars underpin our approach to mitigating and helping ensure that Snapchat is a prohibitive environment for a broad range of misinformation risks relevant to

elections (please see our Transparency Report for further information), while also ensuring Snapchatters have access to tools and information that support participation in democratic processes throughout the world.

We also recognise that advancements in generative AI technologies have elevated the importance of safeguarding information integrity in the context of democratic elections. To this end, we welcome the opportunity to address the Committee's questions on these technologies and the approach that Snap is taking to account for emergent risks across the political information landscape.

1. *What steps is your company taking to attach content credentials, and other relevant provenance signals, to any media created using your products? To the extent that your product is incorporated in a downstream product offered by a third-party, do license terms or other terms of use stipulate the adoption of such measures? To the extent you distribute content generated by others, does your company attach labels when you assess – based on either internal classifiers or credible third-party reports – to be machine-generated or machine-manipulated?*

At Snap, we are constantly working on new ways to enhance Snapchatters' experience with generative AI. Many new features on our platform are powered by generative AI, like AI Lenses, My AI, Dreams, and AI Snaps. Across these features, we have taken steps to indicate that a feature in Snapchat is powered by generative AI in a number of ways, including using the sparkle icon ✨, specific disclaimers, Context Cards, or tool tips.

Some generative AI-powered features, like Dreams and AI Snaps, allow you to create or edit images. When you export or save a generated image to Camera Roll, a watermark of a Snap Ghost with sparkles may be added to those images. The purpose of these watermarks is to provide transparency that the image was created with generative AI and is not real or based on real events, even if it is a realistic style.

We provide guidance to Snapchatters that when they see these contextual symbols or other indicators in Snapchat, they should know they are interacting with AI and not a human, or viewing content that has been produced using Snap's AI tools. We also engage in a number of multistakeholder and industry collaborations to remain abreast of the latest capabilities in content provenance and authenticity, including the AI Futures Initiative (convened by the Information Technology Industry Council) and the Tech Coalition committee on GenAI Content.

Many of Snap's AI features are integrated with upstream models developed by third parties; these features leverage model-level safeguards developed by the third-party service, as well as additional measures that Snapchat layers on top of those models to ensure the experience is

appropriate for our community. These additional safeguards include AI-specific content policies, as well as adversarial testing measures to ensure these features are meeting our safety standards.

Snap has long had in place robust measures to identify and remove harmful content, whether it is AI-generated, deceptively manipulated, or wholly user-generated. Rather than attach labels to harmful content generated by others, we remove it entirely and take appropriate, risk-based enforcement action against accounts that share such content in violation of our policies.

2. *What specific public engagement and education initiatives have you initiated in countries holding elections this year? What has the engagement rate been thus far and what proactive steps are you undertaking to raise user awareness on the availability of new tools hosted by your platform?*

Earlier this year, Snap began publishing a series of public-facing resources related to our approach to election integrity in 2024. This includes multiple blog posts (including this one and this one); a publicly-released letter to global civil society organizations regarding Snap's commitment and approach to election integrity; and public statements in support of specific cross-industry measures to mitigate election-related risks, including the Munich Accord and the Voluntary Election Integrity Guidelines for Tech Companies (released by the International Foundation for Electoral Systems). In addition, Snap has published a support site article and blog post to educate users about our generative AI products and safeguards.

Snap often partners with election authorities in countries and jurisdictions around the world to promote credible election-related resources and information. For example, in the UK, we work closely with the Electoral Commission on communicating voter information to Snapchatters; we are also partnering with the European Parliament in support of AR tools and other resources to raise awareness of this year's elections in the European Union.

In preparation for the United States elections, Snap has partnered with Vote.org to provide resources for voter registration. Snap has also pledged financial and advertising support for a Public Service Announcement to be produced by a cohort of nonpartisan, non-profit organizations to raise awareness of the risks of deceptive, political deepfakes; the PSA is expected to launch late this summer.

In advance of elections in Europe, India, and the United Kingdom, Snap has provided numerous briefings and evidentiary submissions to ensure stakeholders are aware of the measures our company has in place to ensure the integrity of our services in the context of sensitive events like elections.

3. *What specific resources has your company provided for independent media and civil society organizations to assist in their efforts to verify media, generate authenticated media, and educate the public?*

Snap supports efforts by independent media and civil society organizations to ensure the validity and authenticity of media consumed by the public. Snap is pleased to contribute funding, as well as in-kind ad credits, to a cohort of non-partisan, non-profit organizations who are developing a public awareness campaign to educate Americans regarding the risks of deceptive synthetic media in advance of the 2024 elections. Snap has similarly offered ad credits to the European Commission to promote their election integrity efforts ahead of the European Parliamentary elections June 6-9.

4. *What has been your company's engagement with candidates and election officials with respect to anticipating misuse of your products, as well as the effective utilization of content credentialing or other media authentication tools for their public communications?*

Upon request, Snap has provided formal briefings for election authorities in the United States and in Europe. In addition to these briefings, we have attended meetings and provided written submissions on these topics to multiple authorities in the United States, the United Kingdom, and the European Union.

In March, Snap joined a small number of platforms to participate in a multi-day convening with election authorities from around the world and global civil society stakeholders working to advance election integrity and security. An outcome of this convening is the Voluntary Election Integrity Guidelines for Tech Companies, which Snap is pleased to have helped inform.

5. *Has your company worked to develop widely-available detection tools and methods to identify, catalogue, and/or continuously track the distribution of machine-generated or machine-manipulated content?*

In developing our approach to platform integrity, Snap's priority has been to identify and remove harmful content, irrespective of whether it is machine-generated or machine-manipulated. Consistent with our commitments under the Tech Accord—and with our long-standing values and policies—this priority extends to identifying and removing any content that undermines the integrity of civic processes and participation in democratic elections. (Note: we describe our approach to mitigating and enforcing against such harmful content in greater detail in our response to Question 6 below.)

Snap uses internal signals to understand whether an image was created with Snap AI-powered image generation tools. These tools enable our teams to understand if the image was created with Snap Gen AI so that we can append a visual watermark when generated images are shared off-platform or saved to Camera Roll.

Within the Snapchat app, Snapchatters may see various visual indicators when content was created with Snap AI-powered image generation tools. For example, if a Snapchatter posts a

selfie to their story using one of our AI-powered Lenses, other Snapchatters will see an entry point into that same Lens.

We are exploring the possibility of ingesting signals to identify content created off-platform with other generative AI tools. The Coalition for Content Provenance and Authenticity recently proposed a metadata standard for this purpose; we are currently evaluating this against other potential approaches.

> 6. *(To the extent your company offers social media or other content distribution platforms) What kinds of internal classifiers and detection measures are you developing to identify machine-generated or machine-manipulated content? To what extent to these measures depend on collaboration or contributions from generative AI vendors?*

In developing Snap's platform integrity mitigations, we do not index for identifying "machine-generated" or "machine-manipulated content" specifically; rather, our priority is identifying and removing harmful content, irrespective of whether it is AI-manipulated. Snap maintains robust policies—applicable to both the dissemination and the creation of generative AI content—that function to mitigate risk and advance safety.

Creation

In the context of on-platform features for creating generative AI content, Snap has developed several internal policies relating to generative AI. In particular,

(1) Content and Product policies: We have developed a suite of policies that disallow the generation of harmful content (including deceptive political content). Our policy and moderation teams work in partnership with engineering and data science colleagues to ensure that our AI products are responsibly trained on these policy parameters.
(2) Acceptable Use: We have similarly developed Acceptable Use Policies that prohibit the use of our AI tools to attempt to generate violative content at the prompt-level.

Dissemination

In the context of dissemination of content on Snapchat, we have processes in place to mitigate the risks of digital disinformation or deceptive content relating to political matters or processes.

Our Community Guidelines and Terms of Service set out the rules on what content is allowed on Snapchat. They are focused on preventing harm to Snapchatters and the broader community from content and behavior, whether or not caused by generative AI or any other form of IT tools (such as Photoshop). These rules apply to all content formats across our platform, including content that is AI-generated. While the rules are agnostic to content format or creative tools, the Community Guidelines specifically note: "We implement safeguards designed to help keep generative AI content in line with our Community Guidelines, and we expect Snapchatters to use

AI responsibly. We reserve the right to take appropriate enforcement action against accounts that use AI to violate our Community Guidelines, up to and including the possible termination of an account."

Our rules and internal enforcement guidance include clear provisions related to content risks for civic discourse and electoral processes. In particular, our Community Guidelines prohibit spreading false information that causes harm or is malicious, such as denying the existence of tragic events, unsubstantiated medical claims, undermining the integrity of civic processes, or manipulating content for false or misleading purposes (whether through generative AI or through deceptive editing).

Our Community Guidelines rules on false information refer to a more detailed Explainer that prohibits content that undermines the integrity of civic processes, or deep fake content or other media that is manipulated for false or misleading purposes. The Community Guidelines further explain that these prohibitions extend to the following types of harmful content:

- *Procedural interference:* misinformation related to actual election or civic procedures, such as misrepresenting important dates and times or eligibility requirements for participation.
- *Participation interference:* content that includes intimidation to personal safety or spreads rumors to deter participation in the electoral or civic process.
- *Fraudulent or unlawful participation:* content that encourages people to misrepresent themselves to participate in the civic process or to illegally cast or destroy ballots.
- *Delegitimization of civic processes:* content aiming to delegitimize democratic institutions on the basis of false or misleading claims about election results, for example.

Sharing such content will violate Snap's Community Guidelines irrespective of whether it is AI-generated or user-generated, or whether it is generated on Snapchat or on another platform.

Snap has a suite of internal policies and guidelines to help our content review and trust and safety teams apply the Community Guidelines to user generated content disseminated via our online platforms (such as Spotlight and Discover/For You). They provide more granular information for our content review teams. For example, we explain that obvious jokes, memes, satire and non-libelous comments about prominent social figures are not violative; whereas false political narratives meant to undermine elections are NOT OK.

In addition to our Community Guidelines, Snap applies our stricter Content Guidelines for Recommendation Eligibility to content before it may be algorithmically recommended to a broad audience, e.g., on Discover or Spotlight. Our platform does not widely distribute an unvetted

feed of algorithmically curated political information; we disallow all political content[1] from Spotlight (our broadcast platform for User Generated Content) unless it's from trusted news partners and creators, and pre-moderate that surface to ensure that other such political content is not broadly distributed. As an additional safeguard, we monitor any content that is achieving large-scale reach—and ensure a human reviews it—as a "virality circuit breaker" and a means of checking that our pre-moderation systems are working effectively. These safeguards ensure that Snap is not algorithmically promoting political statements from unvetted sources, and generally reflects Spotlight's function as an entertainment platform. (Consistent with our commitments to fundamental rights of expression and access to information, Snapchat provides other, non-algorithmically amplified spaces for users to express their views and political observations, such as Chat and My Story; users can also seek access to political information from known publishers and creators whom Snap has on-boarded for distribution on the Stories tab).

7. *(To the extent your company offers social media or other content distribution platforms) What mechanisms has your platform implemented to enable victims of impersonation campaigns to report content that may violate your Terms of Service? Do you maintain separate reporting tools for public figures?*

Snapchat has reporting mechanisms in place across our services to enable Snapchatters to easily report deceptive behavior and other violations of our Community Guidelines. Reports can also be submitted on our support site at: https://help.snapchat.com/hc/en-us/requests/new.

Additional information about Snap's reporting mechanisms is available here. We also maintain a Trusted Flagger program, through which trusted organizations and government authorities can escalate high-priority concerns.

8. *(To the extent your company offers generative AI products) What mechanisms has your platform implemented to enable victims of impersonation campaigns that may have relied on your models to report activity that may violate your Terms of Service?*

We offer various mechanisms to allow reporting of inappropriate content. Every generative AI product that generates an image has an option for users to provide feedback in-app. Reports can also be submitted via this support site: https://help.snapchat.com/hc/en-us/requests/new.

9. *(To the extent your company offers social media or other content distribution platforms) What is the current status of information sharing between platforms on detecting machine-generated or machine-manipulated content that may be used for*

[1] For these purposes, "political content" means content related to political campaigns and elections, government activities, and/or viewpoints on issues of ongoing debate or controversy. This includes content about candidates or parties for public office, ballot measures or referendums, and political action committees, as well as personal perspectives on candidate positions, government agencies/departments or the government as a whole.

malicious ends (such as election disinformation, non-consensual intimate imagery, online harassment, etc.)? Will your company commit to participation in a common database of violative content?

Snap has information sharing agreements in place with other companies to allow for privacy-respecting signal-sharing and information-sharing related to a variety of online harms, including election-related threats. In addition to industry-peer agreements, we participate alongside other companies in multi-partner information-sharing programs facilitated by NGOs, including StopNCII and Take It Down, which are dedicated specifically to identifying non-consensual intimate imagery. Snap is also a member of the Tech Coalition, which facilitates ThreatExchange (a signal-sharing platform) and Lantern (a cross-platform sharing program related to Child Sexual Exploitation and Abusive Imagery).

Snap is very open to opportunities for responsible information-sharing, and will commit to reviewing any such opportunities or common databases that may become available.

Please let us know if you have any further questions.

Sincerely,

Gina Woodworth

Gina Woodworth
Director, Americas Public Policy