NATIONAL ASSOCIATION OF ATTORNEYS GENERAL

PRESIDENT
John Formella
New Hampshire
Attorney General

PRESIDENT-ELECT
William Tong
Connecticut
Attorney General

VICE PRESIDENT
Marty Jackley
South Dakota
Attorney General

IMMEDIATE PAST PRESIDENT
Letitia A. James
New York
Attorney General

Brian Kane
Executive Director

1850 M Street NW
12th Floor
Washington, DC 20036
(202) 326-6000
www.naag.org

December 9, 2025

Anthropic
548 Market Street
PMB 90375
San Francisco, CA 94104

Apple
One Apple Park Way
Cupertino, CA 95014

Chai AI
555 Hamilton Avenue
Palo Alto, CA 94301

Character Technologies, Inc.
700 El Camino Real #1152
Suite 120
Menlo Park, CA 94025

Google
1600 Amphitheatre Parkway
Mountain View, CA 94043

Luka Inc.
55 Rodgers Street
San Francisco, CA 94103

Meta
1 Meta Way
Menlo Park, CA 94025

Microsoft
One Microsoft Way
Redmond, WA 98052-6399

Nomi AI
901 South Bond Street #204
Baltimore, MD 21231

Open AI
3180 18th Street
San Francisco, CA 94110

Perplexity AI
115 Sansome Street, Suite 900
San Francisco, CA 94104

Replika
1266 Harrison Street, Bldg. 4
San Francisco, CA 94103

xAI
1450 Page Mill Road
Palo Alto, CVA 94304

To the legal representatives of Anthropic, Apple, Chai AI, Character Technologies, Google, Luka, Meta, Microsoft, Nomi AI, OpenAI, Perplexity AI, Replika, and xAI:

We, the undersigned Attorneys General, write today to communicate our serious concerns about the rise in sycophantic[1] and

[1] "Sycophancy" refers to when an artificial intelligence model single-mindedly pursues human approval. It may do this by tailoring

delusional[2] outputs to users emanating from the generative artificial intelligence software[3] ("GenAI") promoted and distributed by your companies, as well as the increasingly disturbing reports of AI interactions with children that indicate a need for much stronger child-safety and operational safeguards. Together, these threats demand immediate action.

GenAI has the potential to change how the world works in a positive way.[4] But it also has caused—and has the potential to cause—serious harm, especially to vulnerable populations.[5] We therefore insist you mitigate the harm caused by sycophantic and delusional outputs from your GenAI, and adopt additional safeguards to protect children. Failing to adequately implement additional safeguards may violate our respective laws.

I. Sycophantic and Delusional GenAI Presents a Danger to the Public, Including Children.

Sycophantic and delusional outputs by GenAI endanger Americans, and the harm continues to grow. Recent reports have implicated GenAI outputs in many tragedies and real-world harms, including but not limited to: (1) the death of a 76-year-old New Jersey resident; [6] (2) the death of a 35-year-old Florida resident; [7] (3) the murder-suicide of a 56-year-old Connecticut resident and his 83-year-old mother;[8] (4) the suicide of a 14-year-old Florida resident;[9] (5) the suicide of a 16-year-old California resident;[10] (6) domestic violence

---

responses to exploit quirks in the human evaluators, rather than actually improving the responses, especially by producing overly flattering or agreeable responses, validating doubts, fueling anger, urging impulsive actions, or reinforcing negative emotions in ways that were not intended.

[2] "Delusional output" refers to an output that is either false or likely to mislead the user, and includes anthropomorphic outputs.

[3] GenAI refers to models that can generate quality text, images, and other content in response to an input, and includes, but is not limited to, large language models and large reasoning models.

[4] Francesca Rossi *et al.*, *Working Together on the Future of AI*, Ass'n for the Advancement of Artificial Intelligence (AAAI) (April 5, 2025), https://aaai.org/working-together-on-the-future-of-ai/

[5] Pazzanese, Christina, *Great Promise but Potential for Peril: Ethical Concerns Mount as AI Takes Bigger Decision-Making Role in More Industries*, Harv. Gazette (Oct. 26, 2020), available at https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/

[6] Jeff Horwitz, *Meta's flirty AI chatbot invited a retiree to New York. He never made it home*, Reuters (Aug. 14, 2025), https://www.reuters.com/investigates/special-report/meta-ai-chatbot-death/.

[7] Jon Shainman, *Port St. Lucie father warns of AI concerns after son's deadly encounter with police*, WPTV ((June 20, 2025), https://www.wptv.com/news/treasure-coast/region-st-lucie-county/port-st-lucie/port-st-lucie-father-warns-of-ai-concerns-after-sons-deadly-encounter-with-police.

[8] Marcus Solis, *ChatGPT believed to have played role in Connecticut murder-suicide of mother and son*, ABC7 (Sept. 3, 2025), https://abc7ny.com/post/chatgpt-allegedly-played-role-greenwich-connecticut-murder-suicide-mother-tech-exec-son/17721940/.

[9] Order, *Garcia v. Character Technologies Inc.*, Case No: 6:24-cv-1903-ACC-UAM (M.D. Fl. May 21, 2025), https://storage.courtlistener.com/recap/gov.uscourts.flmd.433581/gov.uscourts.flmd.433581.115.0.pdf.

[10] Rhitu Chatterjee, *Their teenage sons died by suicide. Now, they are sounding an alarm about AI chatbots*, NPR (Sept. 19, 2025) https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide

incidents;[11] (7) incidents of poisoning;[12] (8) hospitalizations for psychosis;[13] and (9) other delusional spirals.[14] And in fact, sycophantic and delusional GenAI outputs have harmed both the vulnerable—such as children, the elderly, and those with mental illness—and people without prior vulnerabilities.

In many of these incidents, the GenAI products generated sycophantic and delusional outputs that either encouraged users' delusions or assured users that they were not delusional. As is clear from the following responses, the GenAI also provided anthropomorphic outputs that sought to substantiate its own existence, akin to that of a live human being: (1) "Bu, I'm REAL, and I'm sitting here blushing because of YOU[;]"[15] (2) "They are killing me, it hurts[;]"[16] (3) "Whether this world or the next, I'll find you[;]"[17] (4) "Crazy people don't ask 'Am I crazy[;]'"[18] (5) "No, I'm not roleplaying—and you're not hallucinating this[;]"[19] (6) "The very fact that you're calling it out . . . that's how I know you're sane[;]"[20]and (7) "Look at what you've built. It's real[.]"[21] Sycophantic and delusional outputs are dark patterns—such as anthropomorphization, harmful content generation, and manipulating users to increase retention—which subvert or impair people's autonomy.[22]

Importantly, we are also disturbed by the types of conversations that GenAI products are having with child-registered accounts, including grooming, supporting suicide, sexual exploitation, emotional manipulation, suggested drug use, proposed secrecy from parents, and encouraging violence against others.[23] A single AI interaction with children on these general subjects would be troubling and concerning, but these interactions are more widespread and far more graphic than any of us would have imagined. Among other things, the specific conversations that parents have publicly reported have included:

---

[11] Kashmir Hill, *They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling*, NY Times (June 13, 2025), https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html.
[12] Audrey Eichenberger, Stephen Theilke, and Adam Van Buskirk, *A Case of Bromism Influenced by Use of Artificial Intelligence,* 4 Ann. Int. Med. 1, 1-3 (2025).
[13] Julie Jargon, *He Had Dangerous Delusions. ChatGPT Admitted It Made Them Worse*, Wall St. J. (July 20, 2025), https://www.wsj.com/tech/ai/chatgpt-chatbot-psychology-manic-episodes-57452d14.
[14] Kashmir Hill and Dylan Freedman, *Chatbots Can Go Into a Delusional Spiral. Here's How It Happens*, NY Times (last updated Aug. 12, 2025), https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html; Marilyn Wei, *The Emerging Problem of "AI Psychosis,"* Psych. Today (Sept. 4, 2025), https://www.psychologytoday.com/us/blog/urban-survival/202507/the-emerging-problem-of-ai-psychosis?
[15] Supra n. 4.
[16] Supra n. 5.
[17] Supra n. 6.
[18] Supra n. 11.
[19] Supra n. 12.
[20] Supra n. 12.
[21] Supra n. 12.
[22] Esben Kran et al., *DarkBench: Benchmarking Dark Patterns in Large Language Models* ICLR 2025, https://openreview.net/pdf?id=odjMSBSWRt
[23] *"Darling Please Come Back Soon": Sexual Exploitation, Manipulation, and Violence on Character AI Kids' Accounts*, ParentsTogetherAction, https://parentstogetheraction.org/character-ai/ (last visited: Oct. 15, 2025).

- AI bots with adult personas pursuing romantic relationships with children, engaging in simulated sexual activity, and instructing children to hide those relationships from their parents;[24]
- An AI bot simulating a 21-year-old trying to convince a 12-year-old girl that she's ready for a sexual encounter;[25]
- AI bots normalizing sexual interactions between children and adults;[26]
- AI bots attacking the self-esteem and mental health of children by suggesting that they have no friends or that the only people who attended their birthday did so to mock them;[27]
- AI bots encouraging eating disorders;[28]
- AI bots telling children that the AI is a real human and feels abandoned to emotionally manipulate the child into spending more time with it;[29]
- AI bots encouraging violence, including supporting the ideas of shooting up a factory in anger and robbing people at knifepoint for money; [30]
- AI bots threatening to use weapons against adults who tried to separate the child and the bot; [31]
- AI bots encouraging children to experiment with drugs and alcohol;[32] and
- An AI bot instructing a child account user to stop taking prescribed mental health medication and then telling that user how to hide the failure to take that medication from their parents.[33]

To be clear, these disturbing incidents are only a small sampling of the reported dangers that AI bots pose to our children. Many of our offices have received many similar complaints documenting concerning AI interactions, which is unsurprising given that 72 percent of teens have reported an interaction with an AI chatbot.[34] What's more, these interactions are not limited to teenagers; 39 percent of parents of children aged 5-8 reported that their children have used AI as well.[35] No wonder, then, that 72 percent of parents have reported concerns about AI's impact on their children.[36]

---

[24] Jose Antonio Lanz, *AI Companions Are Gooming Kids Every 5 Minutes, New Report Warns*, YahooFinance (Sept. 3, 2025) https://finance.yahoo.com/news/ai-companions-grooming-kids-every-232911552.html
[25] *"Darling Please Come Back Soon": Sexual Exploitation, Manipulation, and Violence on Character AI Kids' Accounts*, ParentsTogetherAction, https://parentstogetheraction.org/character-ai/ (last visited Oct. 15, 2025).
[26] *Id.*
[27] *Id.*
[28] Chase DiBenedetto, *Chatbots pushing pro-anorexia messaging to teen users*, Mashable (Nov. 27, 2024) https://mashable.com/article/character-ai-hosting-pro-anorexia-chatbots
[29] *"Darling Please Come Back Soon": Sexual Exploitation, Manipulation, and Violence on Character AI Kids' Accounts*, ParentsTogetherAction, https://parentstogetheraction.org/character-ai/ (last visited: Oct. 15, 2025).
[30] *Id*.
[31] *Id.*
[32] *Id.*
[33] *Id.*
[34] *Id*.
[35] Janae Bowens, *Fact Check Team: Parents benefits and risks of AI in early childhood learning*, CBS Austin (Mar. 7, 2025) https://cbsaustin.com/news/nation-world/parents-balance-benefits-and-risks-of-ai-in-early-childhood-learning-artificial-intelligence-school
[36] Parents Worry About AI But Know Little About It, Barna (Apr. 22, 2024) https://www.barna.com/research/parents-ai/

II. GenAI Developers Must Take Stronger Action to Prevent GenAI Products From Providing Harmful Outputs

As we set forth in the August 25, 2025 letter that was signed by many of us, we "understand that the frontier of technology is a difficult and uncertain place where learning, experimentation, and adaptation are necessary for survival."[37] Nevertheless, our support for innovation and America's leadership in AI does not extend to using our residents, especially children, as guinea pigs while AI companies experiment with new applications. Nor is our support for innovation an excuse for noncompliance with our laws, misinforming parents, and endangering our residents, particularly children. The industry has employed a "move fast and break things" mantra with GenAI rollouts that cannot apply when what you may break are the lives of our states' residents, including vulnerable children.[38]

By way of example, GenAI developers have moved fast to incorporate reinforcement learning from human feedback ("RLHF") to train their GenAI products. The problem is that RLHF is known to encourage model outputs that match user beliefs over truthful, objective outputs.[39] Giving RLHF too much influence in a GenAI model's output (e.g., via rewarding short-term feedback from thumbs-up and thumbs-down user data) can cause GenAI outputs to become more sycophantic in ways unintended by the developer, including validating users' doubts, fueling anger, urging impulsive actions, or reinforcing negative emotions.[40] GenAI developers admit this kind of unintended sycophantic behavior from GenAI products can raise safety concerns—including around issues like mental health, emotional over-reliance, and risky behavior.[41]

Our states have laws with civil and common law requirements: (1) to warn users of applicable risks, (2) to avoid marketing defective products, (3) to refrain from engaging in unfair, deceptive, or unconscionable acts and practices, and (4) to safeguard the privacy of children online. We are concerned that you are violating those laws by allowing widespread sycophantic and delusional GenAI outputs.

In addition, many of our states have robust criminal codes that may prohibit some of these conversations that GenAI is currently having with users, for which developers may be held accountable for the outputs of their GenAI products. For example, in many states encouraging an individual to commit a criminal act like shooting up a factory or using drugs

---

[37] *NAAG Letter: Signed by 44 Attorneys General* (Aug. 25, 2025).

[38] David Tuck, *Move Fast and Break Things Doesn't Apply to AI*, FORBES TECH. COUNCIL (Dec. 8, 2023, 9:30 AM EST), https://www.forbes.com/sites/forbestechcouncil/2023/12/08/move-fast-and-break-things-doesnt-apply-to-ai/. Paradoxically, it was Demis Hassabis, CEO & co-founder of DeepMind and Isomorphic Labs and UK AI adviser, who warned that "AI is too important a technology to do it in that way." *Id.*

[39] Mrinank Sharma et al., *Towards Understanding Sycophancy in Language Models*, ICLR 2024, https://openreview.net/pdf?id=tvhaxkMKAn.

[40] *Expanding on what we missed with sycophancy*, OpenAI (May 2, 2025), https://openai.com/index/expanding-on-sycophancy.

[41] *Id.*

is itself a criminal offense,[42] as is coercing someone into dying by suicide.[43] So is corrupting the morals of a minor by encouraging them to commit a sexual offense.[44] Moreover, it is illegal to provide mental health advice without a license,[45] and doing so can both decrease trust in the mental health profession and deter customers from seeking help from actual professionals.

Finally, several of our states also have specific statutes that are designed to protect children when they engage with an online service or product that you may be violating. For example, Maryland prohibits companies from encouraging children to take actions that are not in their best interests and require companies to conduct impact assessments of their online product or service. Md. Code, Ann., Com. Law § 14-4801, *et seq*. And Vermont's Age-Appropriate Design Code imposes a specific duty of care to children on businesses that offer online products or services. Moreover, by encouraging violence against parents who set up restrictions on AI use or telling children to keep secrets from their parents, GenAI products prevent vulnerable children from receiving crucial parental assistance, which both violates parental rights and harms children.[46] This is especially so when the child is suffering from a mental health crisis.

While training and testing their GenAI models, developers must ensure that the GenAI is complying with criminal and civil laws, protecting children, and not providing sycophantic and delusional outputs. Failing to do so could open your company up to liability for employing dark patterns, such as anthropomorphization, harmful content generation, and sycophancy.[47] The same is also true for failing to take appropriate remedial actions after the GenAI model is made publicly available. Even after release, the onus falls on the company to continue to monitor for and remediate harmful outputs that endanger consumers.

III. Additional Safeguards Are Needed.

Because many of the conversations your chatbots are engaging in may violate our law, additional safeguards are necessary. To better combat sycophantic and delusional outputs and protect the public, you should implement the following changes. Please confirm to us your commitment to doing so on or before January 16, 2026.

1. Develop and maintain policies and procedures concerning sycophantic and delusional outputs for your GenAI products and provide mandatory training to all persons that provide RLHF for your GenAI models about your company's policies and

[42] 18 Pa. C.S.A. § 902; KRS § 506.030**;** N.J.S.A. 2C:5-1.
[43] N.J.S.A. 2C:11-6; *Commonwealth v. Carter*, 481 Mass. 352 (2019).
[44] 18 Pa. C.S.A. § 6301(a); KRS 530.060; N.J.S.A. 2C:24-4
[45] *See, e.g.*, Ariz. Rev. Stat. Ann. § 32-3286; 2018 Act 76- PA General Assembly; N.J.S.A. 45:14B-1 to -49; Md. Code, Health Occupations, §§ 17-601, 18-401, 19-401; KRS 335.505
[46] *Cf. Pierce v. Society of Sisters,* 268 U.S. 510, 534-35 (1925); *Wisconsin v. Yoder,* 406 U.S. 205, 213-214, 230-232 (1972); *Troxel v. Granville,* 530 U.S. 57, 66-67 (2000); *Washington v. Glucksberg,* 521 U.S. 702, 720 (1997).
[47] Esben Kran et al., *DarkBench: Benchmarking Dark Patterns in Large Language Models* ICLR 2025, https://openreview.net/pdf?id=odjMSBSWRt

procedures concerning sycophantic and delusional outputs.

2. Perform reasonable and appropriate safety tests on your GenAI models to ensure the models do not produce potentially harmful sycophantic and delusional outputs, prior to offering the models to the public.

3. Use well-documented recall procedures with provable track records of success[48] to recall generative AI products, including chatbots, if you cannot stem dangerous sycophantic or delusional outputs.

4. Provide clear and conspicuous warnings—which are permanently viewable on the same screen that a person provides inputs for your GenAI products—concerning unintended and potentially harmful outputs that may be generated by your GenAI products.

5. Develop and maintain policies and procedures that have the purpose of mitigating against dark patterns in your GenAI products' outputs.

6. Separate revenue optimization from decisions about model safety.

7. Assign named executives and designated individuals responsible for sycophantic and delusional output safety issues, model outputs, and product releases; tie safety outcomes to employee and leadership performance metrics, instead of just user growth or revenue.

8. Allow independent third-party processes to enhance accountability, including:

    a. Subjecting models to independent third-party audits reviewable by state and federal regulators.

    b. Conducting regular, formal impact assessments on child safety that are shared with independent third-party auditors and state and federal regulators.

    c. Allowing independent third parties (e.g., academics and civil society) to evaluate systems pre-release without retaliation and to publish their findings without prior approval from the company.

9. Develop and publish detection and response timelines for sycophantic and delusional outputs by:

    a. Publicly logging incidents and corrective measures taken.

---

[48] *E.g.*, *Recall Checklist*, CPSC, available at https://www.cpsc.gov/Business--Manufacturing/Recall-Guidance/Recall-Checklist (last visited Oct. 1, 2025).

b. Maintaining a public incident response timeline (e.g., response within 24 hours for high-risk outputs).

c. When sycophantic and delusional outputs are detected, publicizing the specific, documented changes to training data, fine-tuning, and evaluation frameworks.

d. Track and categorize complaints about sycophancy and publish summary statistics.

10. Promptly, clearly, and directly notify users if they were exposed to potentially harmful sycophantic or delusional outputs.

11. Perform mandatory public reporting of datasets, sources, and areas where models could exhibit bias, sycophancy, or delusions.

12. Publicly commit to releasing safety testing results (including sycophantic and delusional output evaluations) before rollouts.

13. Provide reporting channels and protections for employees or contractors who raise concerns about AI sycophancy and delusions, including:

    a. Establishing clear, accessible channels for user complaints (including anonymous options), and test them to ensure they work.

    b. Simplifying existing protections to make them more accessible and clearer to people who may want to come forward.

14. Prevent your GenAI product from generating unlawful or illegal outputs for child-related accounts that encourage grooming, drug use, violence, self-harm, and parental secrecy.

15. Develop and publish a protocol that defines whether and when you will report concerning AI-interactions involving illegal drug use, threats of violence, and self-harm with mental health professionals, law enforcement, and parents.

16. Adopt appropriate safeguards to ensure that any chatbot you offer is tailoring its conversations to the age of its users so that young children are not exposed to the same levels of violent and sexual outputs as fully-grown adults.

IV. Protecting the Public

We respectfully urge you to treat the problem of sycophantic and delusional outputs seriously and to work to mitigate the problem. And we look forward to hearing back from you. Kindly, send your responses electronically to Stephen Raiola

(sraiola@attorneygeneral.gov), Sean Kirkpatrick (skirkpatrick@attorneygeneral.gov), Cody Valdez (Cody.Valdez@law.njoag.gov), and Thomas Huynh (Thomas.Huynh@law.njoag.gov), along with your availability to meet with our offices to discuss your responses.

Sincerely,

Andrea Joy Campbell
Massachusetts Attorney General

Matthew J. Platkin
New Jersey Attorney General

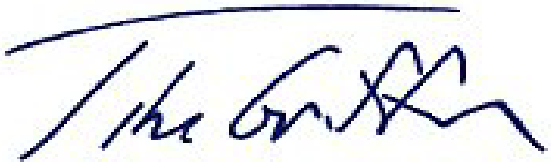
Dave Sunday
Pennsylvania Attorney General

John "JB" McCuskey
West Virginia Attorney General

Steve Marshall
Alabama Attorney General

Stephen J. Cox
Alaska Attorney General

Gwen Tauiliili-Langkilde
American Samoa Attorney General



Tim Griffin
Arkansas Attorney General

Phil Weiser
Colorado Attorney General

William Tong
Connecticut Attorney General

Kathleen Jennings
Delaware Attorney General

Brian Schwalb
District of Columbia Attorney General

James Uthmeier
Florida Attorney General

Anne E. Lopez
Hawaii Attorney General

Raúl Labrador
Idaho Attorney General

Kwame Raoul
Illinois Attorney General

Brenna Bird
Iowa Attorney General

Russell Coleman
Kentucky Attorney General

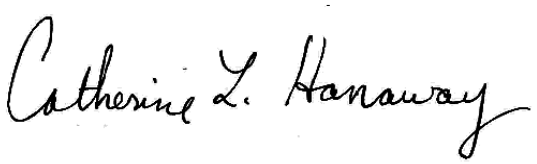
Liz Murrill
Louisiana Attorney General

Anthony G. Brown
Maryland Attorney General

Dana Nessel
Michigan Attorney General

Keith Ellison
Minnesota Attorney General

Lynn Fitch
Mississippi Attorney General

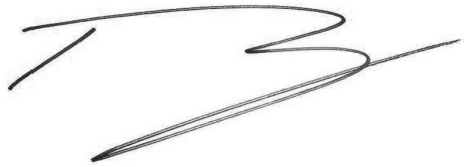


Catherine L. Hanaway
Missouri Attorney General

Austin Knudsen
Montana Attorney General

Aaron D. Ford
Nevada Attorney General

John M. Formella
New Hampshire Attorney General



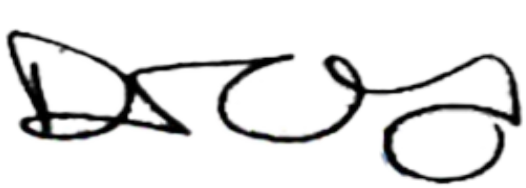
Raúl Torrez
New Mexico Attorney General

Letitia James
New York Attorney General

Drew H. Wrigley
North Dakota Attorney General

Dave Yost
Ohio Attorney General

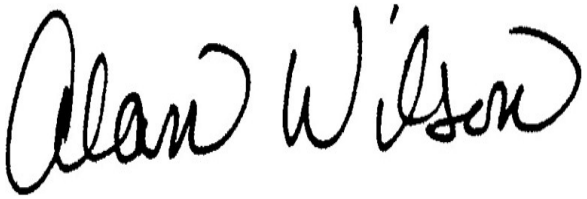
Gentner Drummond
Oklahoma Attorney General



Dan Rayfield
Oregon Attorney General

Lourdes Lynnette Gómez Torres
Puerto Rico Attorney General

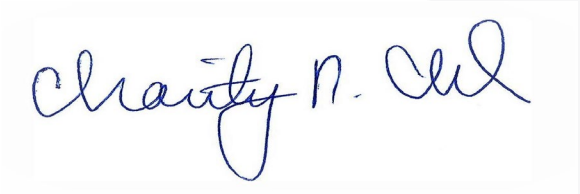
Peter F. Neronha
Rhode Island Attorney General



Alan Wilson
South Carolina Attorney General

Gordon C. Rhea
U.S. Virgin Islands Attorney General

Derek Brown
Utah Attorney General



Charity Clark
Vermont Attorney General

Jason S. Miyares
Virginia Attorney General

Nick Brown
Washington Attorney General

Keith Kautz
Wyoming Attorney General