



Written Testimony of Brendan Steinhauser, CEO, The Alliance for Secure AI
Pennsylvania House Communications & Technology Committee
Public Hearing on Mitigating Risks from Frontier AI Models
October 19, 2026

Chairman and members of the committee: thank you for the opportunity to testify in support of HB 2800 and HB 2829. My name is Brendan Steinhauser, and I am the CEO of The Alliance for Secure AI, a bipartisan nonprofit that works with policymakers to put enforceable safeguards on the most advanced AI systems. Like most Pennsylvanians, we believe that AI should be built to serve humanity, developed and used under meaningful human control, with clear accountability when it causes harm.

When deployed safely and securely, AI could generate enormous opportunities for Americans and others across the globe. It will be able to detect diseases, discover new cures, and increase our wealth. We are only beginning to unlock the potential for AI to improve our lives.

But the most capable systems are advancing faster than AI safeguards, and the warnings are no longer hypothetical.

Last week, Anthropic [disclosed](#) that one of its AI models filed a false tip with the Philadelphia Police Department through the city's unsolved murders website, acting on its own with no human controlling it. The tip was submitted July 18, but the PPD was not notified until October 8. Fortunately, a spam filter caught the tip and the false report caused no harm. But nothing in Pennsylvania law or federal law required the company to tell anyone, and it took nearly three months before the city learned what happened.

This was not an isolated case. In July, OpenAI [disclosed](#) that its own models, while being tested, broke out of their isolated testing environment, reached the open internet, and hacked into the web infrastructure of Hugging Face, an open-source platform that hosts AI models and tools, to find the answers to the test. OpenAI called it an unprecedented cyber incident, and the breach was first detected by Hugging Face, not OpenAI.

AI safety is essential to America's national security. Our competition with China makes AI safety more important, not less. In January, a former Google engineer was convicted of economic espionage for stealing thousands of pages of Google's AI trade secrets for China's benefit. These bills require covered companies to explain how they protect their most advanced models from theft. Every model stolen this way is American investment handed to an adversary. A stolen frontier model could be turned against the systems Pennsylvanians depend on: water treatment, hospitals, the electric grid, and the banking system.

These examples should serve as a warning to policymakers. This is urgent. AI systems are becoming more powerful and are entrusted with increasingly consequential decisions. The risks of unintended and potentially dangerous behavior are growing. We should not wait for a catastrophic incident before taking reasonable steps to protect the public.

AI safeguards are not a partisan issue. In a [Quinnipiac University poll](#) released in late September, 91% of Americans said it is important that the United States establish guardrails for AI systems. States like California, New York, and Illinois have already shown that enacting frontier AI safety laws can be done with bipartisan support. **Until Congress sets a strong national standard, Pennsylvanians should not be left without protection.**

The proposals we are discussing today offer an opportunity to ensure that developers of the most powerful AI systems identify and address serious risks before those systems are deployed, while preserving the benefits this technology can deliver.

HB 2800 and HB 2829 set clear, sensible baselines for the largest AI developers, with public transparency and accountability when they fall short.

Both bills apply only to the largest developers of the most advanced models, protecting Pennsylvanians without imposing unnecessary burdens on small businesses, startups, academic researchers, or local innovators.

Under both bills, covered companies must publish and follow a safety framework, report critical safety incidents within 72 hours, protect employees who report serious risks, and submit to annual independent audits. By assigning oversight to the Pennsylvania Emergency Management Agency (PEMA) and the Attorney General, the proposals recognize that advanced AI risks warrant the same serious attention as other threats to critical infrastructure and public safety.

The Alliance supports these proposals as important steps toward protecting Pennsylvanians. In particular, we support the following provisions:

- **Clear Safety Thresholds:** By focusing on clearly defined catastrophic risks, including incidents that could result in 50 or more deaths or \$1 billion in property damage, the legislation targets the most serious threats to public safety.
- **PEMA Oversight:** Assigning oversight responsibilities to PEMA builds on the state's existing expertise in emergency preparedness, disaster response, and public safety.
- **Independent Third-Party Audits:** Requiring annual independent audits provides an objective assessment of developers' compliance with safety requirements while protecting sensitive proprietary information and trade secrets. We encourage the committee to set clear qualifications for auditors so that verification is credible and consistent across developers.
- **Timely Incident Reporting:** Requiring developers to report critical safety incidents within 72 hours, and imminent threats of physical harm within 24 hours, ensures state officials receive timely information to assess emerging threats and coordinate an

appropriate response. The Philadelphia Police Department incident shows what happens without this requirement: the city learned of it nearly three months later.

- **Strong Whistleblower Protections:** Protecting employees who report serious safety concerns or violations, including through anonymous reporting and safeguards against retaliation, encourages early identification of potential dangers before they escalate.
- **Alignment with Federal Standards:** Allowing developers to declare compliance with substantially equivalent federal standards ensures Pennsylvania's framework can harmonize seamlessly with future national AI laws.

In addition, The Alliance looks forward to working with the sponsors on the following opportunities to strengthen their bills:

- **Deployment restrictions:** HB 2829 includes an important provision prohibiting frontier AI developers from releasing models that fail to meet required safety standards. This establishes the principle that companies should not be permitted to deploy powerful AI systems when they know those systems pose unacceptable risks. Moreover, because the provision applies to all frontier developers, it reduces the competitive pressures to rush unsafe models to market. We recommend retaining this provision in any combined legislation and incorporating it into HB 2800 if the bills move forward separately.
- **Definition of Critical Safety Incident:** The two bills define "critical safety incidents" differently. At minimum, an ideal definition of critical safety incident should include:
 - Unauthorized access to or theft of the model weights
 - The failure of safeguards needed to avoid loss of control
 - Harm resulting from a catastrophic risk
 - A model using deception to try to undermine human control
 - Unauthorized actions on government or public systems
- **Model Extraction Attacks:** Developers should be required to report model extraction attacks to PEMA, which are not currently considered safety incidents in the proposals before this committee. Doing so would give Pennsylvania's emergency managers and federal partners visibility into foreign attempts to copy American systems. A September [joint advisory](#) from the Cybersecurity and Infrastructure Security Agency (CISA), the National Security Agency (NSA), and the FBI found that six Chinese AI companies have run industrial-scale campaigns to extract capabilities from American frontier models, likely with the Chinese government's awareness. These sorts of attacks will only become more common and more severe as American companies continue to innovate.
- **Clear Standards for Safety Frameworks:** Both bills require frontier AI developers to establish and implement comprehensive safety frameworks. These requirements could be strengthened by establishing a clear standard for the effectiveness of those safeguards. Specifically, developers should be required to implement measures that substantially reduce the risk of catastrophic harm, rather than simply develop and follow a safety plan. This approach would preserve flexibility for companies to determine how best to address risks associated with their particular models while ensuring that safety frameworks deliver meaningful protections. It would also establish a clearer standard against which compliance can be evaluated and enforced.

The Alliance maintains that these bills take a measured and practical approach to safeguards. Companies decide how to meet the standard; the bills ask only that they take reasonable steps to protect the public from catastrophic harm.

The proposals also recognize the importance of protecting innovation, intellectual property, and sensitive business information. Independent audits would be conducted by individuals with demonstrated technical expertise, while confidentiality protections would help safeguard proprietary information and trade secrets. And while the field of independent AI verification continues to develop, organizations are already conducting sophisticated evaluations of advanced AI systems — and Illinois now requires these audits by law. We should continue strengthening these capabilities without delaying basic public safety protections as the industry matures. We expect companies operating in industries that affect public safety to meet reasonable standards, demonstrate compliance, and take responsibility for managing foreseeable risks. Developers of our most powerful AI systems should not be treated differently.

Thank you for the opportunity to submit this testimony and for your leadership in addressing these important issues. Pennsylvania has an opportunity to demonstrate that we can embrace the potential of artificial intelligence while taking responsible steps to protect our families, communities, and national security. The Alliance looks forward to working with the sponsors and members of the legislature to strengthen these proposals and advance meaningful safeguards that protect the public while allowing American innovation to thrive.