HB4705

LRB10416697SPS30101b

104TH GENERAL ASSEMBLY

State of Illinois

2025 and 2026

HB4705

by Rep. Daniel Didech

SYNOPSIS AS INTRODUCED:

New Act

Creates the Artificial Intelligence Public Safety and Child Protection Transparency Act. Provides that a frontier artificial intelligence model developer or large chatbot provider shall write, implement, comply with, and clearly and conspicuously publish on its website a public safety and child protection plan. Provides that the Attorney General shall establish a mechanism to be used by a large frontier developer, a large chatbot provider, or a member of the public to report a safety incident related to specified artificial intelligence models or chatbots. Sets forth provisions concerning the protection of whistleblowers; third party audits of large frontier developers; and civil penalties. Provides for rulemaking by the Attorney General. Effective January 1, 2027.

LRB104 16697 SPS 30101 b

A BILL FOR

AN ACT concerning business.

Be it enacted by the People of the State of Illinois, represented in the General Assembly:

Section 1. Short title. This Act may be cited as the Artificial Intelligence Public Safety and Child Protection Transparency Act.

Section 5. Findings. The General Assembly finds and declares:

(a) Artificial intelligence, including new advances in foundation models, has the potential to catalyze innovation and the rapid development of a wide range of benefits for Illinoisans and the Illinois economy, including advances in medicine, agriculture, and climate science, and to push the bounds of human creativity and capacity.

(b) Targeted interventions to support effective artificial intelligence governance should balance the technology's benefits and the potential for material risks.

(c) In building a robust and transparent evidence environment, policymakers can align incentives to simultaneously protect consumers, leverage industry expertise, and recognize leading safety practices.

(d) As industry actors conduct internal research on their technologies' impacts, public trust in these technologies

would significantly benefit from access to information regarding, and increased awareness of, frontier artificial intelligence capabilities.

(e) Greater transparency can also advance accountability, competition, and public trust.

(f) Whistleblower protections and public-facing information sharing are key instruments to increase transparency.

(g) Incident reporting systems enable monitoring of the post-deployment impacts of artificial intelligence.

(h) Unless they are developed with careful diligence and reasonable precaution, there is concern that advanced artificial intelligence systems could have capabilities that pose catastrophic risks from both malicious uses and malfunctions, including artificial intelligence-enabled hacking, biological attacks, and loss of control.

(i) With the frontier of artificial intelligence rapidly evolving, there is a need for legislation to track the frontier of artificial intelligence research and alert policymakers and the public to serious risks and harms from the most advanced artificial intelligence systems, while avoiding burdening smaller companies behind the frontier.

(j) While the major artificial intelligence developers have already voluntarily established the creation, use, and publication of frontier artificial intelligence frameworks as an industry best practice, not all developers have provided

reporting that is consistent and sufficient to ensure necessary transparency and protection of the public. Mandatory, standardized, and objective reporting by frontier developers is required to provide the government and the public with timely and accurate information.

(k) Timely reporting of critical safety incidents to the government is essential to ensure that public authorities are promptly informed of ongoing and emerging risks to public safety. This reporting enables the government to monitor, assess, and respond effectively if the advanced capabilities emerge in frontier artificial intelligence models that may pose a threat to the public.

(l) In the future, foundation models developed by smaller companies or that are behind the frontier may pose significant catastrophic risk, and additional legislation may be needed at that time.

(m) It is the intent of the General Assembly to create more transparency, but collective safety will depend in part on frontier developers taking due care in their development and deployment of frontier models proportional to the scale of the foreseeable risks.

Section 10. Definitions. As used in this Act:

"Affiliate" means a person controlling, controlled by, or under common control with a specified person, directly or indirectly, through one or more intermediaries.

"Artificial intelligence model" means an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments.

"Catastrophic risk" means a foreseeable and material risk that a frontier developer's development, storage, use, or deployment of a frontier model will materially contribute to the death of, or serious injury to, more than 50 people or more than $1,000,000,000 in damage to, or loss of, property arising from a single incident involving a frontier model doing the following:

(1) providing expert-level assistance in the creation or release of a chemical, biological, radiological, or nuclear weapon;

(2) engaging in conduct with no meaningful human oversight, intervention, or supervision that is either a cyberattack or, if the conduct had been committed by a human, would constitute the crime of murder, assault, extortion, or theft, including theft by false pretense; or

(3) evading the control of its frontier developer or user.

"Catastrophic risk" does not include a foreseeable and material risk from the following:

(1) information that a frontier model outputs if the information is otherwise publicly accessible in a

substantially similar form from a source other than a foundation model;

(2) lawful activity of the federal government; or

(3) harm caused by a frontier model in combination with other software if the frontier model did not materially contribute to the harm.

"Child safety incident" means death or bodily injury to a minor resulting from the materialization of a child safety risk.

"Child safety risk" means a material and foreseeable risk that a frontier developer's foundation model, when used as part of a covered chatbot operated by the frontier developer, will engage in behavior when conversing with a minor that, if it had been engaged in by a human, would be deemed to intentionally or recklessly do the following:

(1) cause death or bodily injury to that minor, including as a result of self-harm; or

(2) cause damage to mental health that constitutes severe emotional distress.

"Covered chatbot" means a service that:

(1) allows an ordinary person to converse with and have conversations where humanlike responses are generated by a foundation model;

(2) is foreseeably likely to be accessed by minors; and

(3) has at least 1,000,000 monthly active users.

"Covered risk" means a catastrophic risk or a child safety risk.

"Critical safety incident" means the following:

(1) unauthorized access to, modification of, inadvertent release of, or exfiltration of, the model weights of a frontier model;

(2) the death of, or serious injury to, more than 50 people or more than $1,000,000,000 in damage to, or loss of, property resulting from the materialization of a catastrophic risk;

(3) loss of control of a frontier model that causes death, bodily injury, or that demonstrates materially increased catastrophic risk; or

(4) the use of deceptive techniques by a frontier model against its frontier developer to subvert the controls or monitoring of its frontier developer outside of the context of an evaluation designed to elicit this behavior and in a manner that demonstrates materially increased catastrophic risk.

"Deploy" means to make a frontier model available to a third party for use, modification, copying, or combination with other software. "Deploy" does not include making a frontier model available to a third party for the primary purpose of developing or evaluating the frontier model.

"Employee" has the meaning set forth in the Whistleblower Act.

"Foundation model" means an artificial intelligence model that is:

(1) trained on a broad data set;

(2) designed for generality of output; and

(3) adaptable to a wide range of distinctive tasks.

"Frontier developer" means a person who has trained, or initiated the training of, a frontier model, with respect to which the person has used, or intends to use, at least as much computing power to train the frontier model as would meet the technical specifications described in the definition of "frontier model". "Frontier developer" does not include accredited colleges and universities to the extent that the colleges and universities are engaging in academic research. For the purpose of this definition, if a person subsequently transfers full intellectual property rights of a frontier model to another person, including the right to resell the model, and retains none of those rights, then the receiving person shall be considered the frontier developer with respect to that frontier model.

"Frontier model" means a foundation model that was trained using a quantity of computing power greater than 10^26 integer or floating-point operations. The quantity of computing power described in this definition shall include computing for the original training run and for any subsequent fine-tuning, reinforcement learning, or other material modifications the developer applies to a preceding foundation model.

"Large chatbot provider" means a provider who makes a covered chatbot available in this State and who, together with the provider's affiliates, collectively have an annual revenue of at least $25,000,000.

"Large frontier developer" means a frontier developer who, together with the large frontier developer's affiliates, collectively have an annual revenue of at least $500,000,000.

"Minor" means an individual younger than 18 years old.

"Model weight" means a numerical parameter in a frontier model that is adjusted through training and that helps determine how inputs are transformed into outputs.

"Property" means tangible or intangible property.

"Public safety and child protection plan" means a documented technical and organizational protocol to manage, assess, and mitigate covered risks.

"Safety incident" means a child safety incident or a critical safety incident.

Section 15. Public safety and child protection plans.

(a) A large frontier developer or large chatbot provider shall write, implement, comply with, and clearly and conspicuously publish on its website a public safety and child protection plan that describes in detail:

(1) For a large frontier developer only, how the large frontier developer:

(A) defines and assesses thresholds used by the

large frontier developer to identify and assess whether a frontier model has capabilities that could pose a catastrophic risk, which may include multiple-tiered thresholds;

(B) applies mitigations to address the potential for catastrophic risks based on the results of the assessments undertaken in accordance with subparagraph (A);

(C) reviews assessments of catastrophic risk and adequacy of mitigations of catastrophic risk as part of the decision to deploy a frontier model or use it extensively internally;

(D) uses third parties to assess the potential for catastrophic risks and the effectiveness of mitigations of catastrophic risks;

(E) implements cybersecurity practices to secure unreleased frontier model weights from unauthorized modification or transfer by internal or external parties; and

(F) assesses and manages catastrophic risk resulting from the internal use of its frontier models, including risks resulting from a frontier model circumventing oversight mechanisms or being used for artificial intelligence research and development in a manner that could materially increase catastrophic risk.

(2) For a large chatbot provider only, how the large chatbot provider:

(A) assesses potential for child safety risks;

(B) applies mitigations to address the potential for child safety risks based on the results of the assessments undertaken in accordance with subparagraph (A); and

(C) uses third parties to assess the potential for child safety risks and the effectiveness of mitigations of child safety risks.

(3) For both large frontier developers and large chatbot providers, how the large frontier developer or large chatbot provider:

(A) incorporates national standards, international standards, and industry-consensus best practices into its public safety and child protection plan;

(B) revisits and updates the public safety and child protection plan, including any criteria that trigger updates and how the large frontier developer determines when its foundation models or frontier models are substantially modified enough to require disclosures in accordance with subsection (d) or subsection (e);

(C) identifies and responds to safety incidents; and

(D) institutes internal governance practices to

ensure implementation of these processes.

(b) A large frontier developer shall write its public safety and child protection plan so that, if successfully implemented, it would prevent unreasonable catastrophic risk.

(c) If a large frontier developer or large chatbot provider makes a material modification to its public safety and child protection plan, the large frontier developer or large chatbot provider shall clearly and conspicuously publish the modified public safety and child protection plan and a justification for that modification within 30 days after the modification is made.

(d) Before, or concurrently with, integrating a new foundation model, or a version of an existing foundation model that has been substantially modified, into a covered chatbot operated by a large chatbot provider, a large chatbot provider shall conspicuously publish on its website summaries of the following:

(1) all assessments of child safety risks conducted in accordance with the large chatbot provider's public safety and child protection plan;

(2) the results of those assessments;

(3) the extent to which third-party evaluators were involved; and

(4) other steps taken to fulfill the requirements of the public safety and child protection plan with respect to child safety risks.

(e) Before, or concurrently with, deploying a new frontier model or a version of an existing frontier model that a large frontier developer has substantially modified, a large frontier developer shall implement appropriate safeguards to prevent unreasonable catastrophic risk and conspicuously publish on its website summaries of the following:

(1) all assessments of catastrophic risks from the frontier model conducted in accordance with the large frontier developer's public safety and child protection plan;

(2) the results of those assessments;

(3) the extent to which third-party evaluators were involved; and

(4) other steps taken to fulfill the requirements of the public safety and child protection plan with respect to catastrophic risks from the frontier model.

A large frontier developer that publishes the information described in this subsection as part of a larger document, including a system card or model card, shall be deemed in compliance with this subsection.

(f) A large frontier developer shall not use or deploy a frontier model if doing so would pose unreasonable catastrophic risk.

(g) A large frontier developer or large chatbot provider shall not make a materially false or misleading statement or omission about covered risks from its activities or its

management of covered risks.

A large frontier developer or large chatbot provider shall not make a materially false or misleading statement or omission about its implementation of, or compliance with, its public safety and child protection plan.

This subsection does not apply to a statement that was made in good faith and was reasonable under the circumstances.

(h) When a large frontier developer or large chatbot provider publishes documents to comply with this Section, the large frontier developer or large chatbot provider may make redactions to those documents that are necessary to protect the large frontier developer's or large chatbot provider's trade secrets, the large frontier developer's or large chatbot provider's cybersecurity, public safety, or the national security of the United States or to comply with any State or federal law. If a large frontier developer or large chatbot provider redacts information in a document under this subsection, the large frontier developer or large chatbot provider shall describe the character and justification of the redaction in any published version of the document to the extent permitted by the concerns that justify redaction and shall retain the unredacted information for 5 years.

Section 20. Reporting of safety incidents.

(a) The Attorney General shall establish a mechanism to be used by a frontier developer, a large chatbot provider, or a

member of the public to report a safety incident that includes the following:

(1) the date of the safety incident;

(2) the reasons the incident qualifies as a safety incident; and

(3) a short and plain statement describing the safety incident.

(b) A frontier developer shall report any critical safety incident pertaining to one of its frontier models to the Attorney General within 15 days after discovering the critical safety incident.

(c) If a frontier developer discovers that a critical safety incident poses an imminent risk of death or serious physical injury, the frontier developer shall disclose that incident within 24 hours after discovering the critical safety incident to an authority, including any law enforcement agency or public safety agency with jurisdiction, that is appropriate based on the nature of that incident and as required by law.

(d) A large chatbot provider shall report any child safety incident pertaining to one of its covered chatbots to the Attorney General within 15 days after discovering the child safety incident.

(e) The Attorney General shall establish a mechanism to be used by a large frontier developer to confidentially submit summaries of any assessments of the potential for catastrophic risk resulting from internal use of its frontier models.

(f) A large frontier developer shall transmit to the Attorney General a summary of any assessment of catastrophic risk resulting from internal use of its frontier models no less frequently than every 3 months.

(g) The Attorney General may transmit reports of safety incidents, summaries of assessments of the potential for catastrophic risk from internal use, and reports from employees to the General Assembly, the Governor, the federal government, or an appropriate State agency. The Attorney General shall consider any risks related to trade secrets, public safety, cybersecurity, or national security when transmitting reports.

Section 25. Rulemaking; definitions.

(a) On or before January 1, 2028, and annually thereafter, the Attorney General shall assess recent evidence and developments relevant to the purposes of this Act and may adopt rules to update the following definitions for the purposes of this Act to ensure that they accurately reflect technological developments, scientific literature, and widely accepted national and international standards:

(1) "Frontier model" so that it applies to foundation models at the frontier of artificial intelligence development.

(2) "Frontier developer" so that it applies to developers of frontier models who are themselves at the

frontier of artificial intelligence development.

(3) "Large frontier developer" so that it applies to well-resourced frontier developers.

(4) "Large chatbot provider" so that it applies to well-resourced companies developing covered chatbots that may pose child safety risks.

(b) In adopting rules under this Section, the Attorney General shall take into account the following:

(1) similar thresholds used in international standards or federal law, regulations, or guidance documents for the management of catastrophic risks or child safety risks;

(2) input from stakeholders, including academics, industry, the open-source community, and governmental entities;

(3) the extent to which a person will be able to determine, before beginning to train or deploy a foundation model, whether that person will be subject to this Act as a frontier developer or as a large frontier developer with a focus toward allowing earlier determinations if possible;

(4) the complexity of determining whether a person or foundation model is covered, with a focus toward allowing simpler determinations if possible;

(5) the external verifiability of determining whether a person or foundation model is covered, with a focus toward definitions that are verifiable by parties other

than the frontier developer; and

(6) thresholds used by other states in similar law.

(c) The Attorney General shall align any rules adopted under this Section with a definition adopted in a federal law or regulation, to the extent that it is consistent with the purposes of this Act.

Section 30. Protection of whistleblowers.

(a) A frontier developer or large chatbot provider shall not make, adopt, enforce, or enter into a rule, regulation, policy, or contract that prevents an employee from making a disclosure protected under the Whistleblower Act.

(b) A large frontier developer shall provide a reasonable internal process through which an employee may anonymously disclose information to the large frontier developer if the employee has a good faith belief that the information discloses a substantial and specific danger to employees, public health, or safety or a violation of this Act, including a monthly update to the person who made the disclosure regarding the status of the large frontier developer's investigation of the disclosure and the actions taken by the large frontier developer in response to the disclosure.

(c) Except as provided in subsection (d), the disclosures and responses of the process required by this Section shall be shared with officers and directors of the large frontier developer at least once each quarter.

(d) If an employee has alleged wrongdoing by an officer or director of the large frontier developer in a disclosure or response, subsection (c) shall not apply with respect to that officer or director.

Section 35. Third-party audits.

(a) At least once every calendar year, a large frontier developer shall retain a reputable third-party auditor to produce a report assessing the following:

(1) whether the large frontier developer has complied with its public safety plan and any instances of noncompliance;

(2) any instances where the large frontier developer's public safety plan has not been stated clearly enough to determine whether the large frontier developer has complied; and

(3) whether redactions made by the large frontier developer in documents published in accordance with this Act are reasonable and whether any statements made by the large frontier developer may be false or misleading.

(b) A large frontier developer shall allow the third-party auditor access to all materials produced to comply with this Act and any other materials reasonably necessary to perform the assessment required under subsection (a).

(c) The large frontier developer shall retain the auditor's report for 5 years and allow the Attorney General to

inspect the unredacted version of the report upon request.

(d) In conducting the audit, the auditor shall employ or contract one or more individuals with expertise in corporate compliance and one or more individuals with technical expertise in the safety of foundation models.

Section 40. Civil penalties.

(a) A large frontier developer that violates this Act shall be subject to a civil penalty in an amount dependent upon the severity of the violation that does not exceed $1,000,000 per violation.

(b) A large chatbot provider that violates this Act shall be subject to a civil penalty in an amount dependent upon the severity of the violation that does not exceed $50,000 per violation.

(c) A civil penalty described in this Section shall be recovered in a civil action brought by the Attorney General.

Section 45. Loss of equity. The loss of value of equity does not count as damage to or loss of property for the purposes of this Act.

Section 50. Compliance with other laws.

(a) The Attorney General may adopt rules creating alternative compliance pathways for frontier developers or large chatbot providers that comply with a federal law,

regulation, or guidance document or a law of another state of the United States.

(b) A rule adopted under this Section shall:

(1) Specify the provisions of this Act for which the alternative compliance pathway is being established.

(2) Specify the federal law, regulation, or guidance document, or the law of another state, compliance with which shall serve as the alternative compliance pathway for the provisions specified under paragraph (1). The federal law, regulation or guidance document or the law of another state shall be substantially equivalent to, or more protective against catastrophic risk than, the provisions of this Act specified under paragraph (1).

(c) If a rule adopted under this Section identifies, as described in paragraph (1) of subsection (b), a provision of this Act that requires reporting to the State and if the alternative compliance pathway requires reporting to the federal government, the rule may, but need not, continue to require reporting to the State. The rule shall not consider reporting to another state to be sufficient for compliance with the relevant provision of this Act.

(d) A rule adopted under this Section may establish steps frontier developers or large chatbot providers must take to demonstrate their compliance with the alternative compliance pathway if it would otherwise be challenging for the State to verify compliance, such as the submission of documentation to

the State.

(e) A frontier developer or large chatbot provider that intends to make use of an alternative compliance pathway created by rule under this Section shall declare its intent to do so to the Attorney General.

After declaring its intent, a frontier developer or large chatbot provider shall be deemed in compliance with the provision of this Act identified by the rule under paragraph (1) of subsection (b) to the extent that the frontier developer or large chatbot provider complies with the requirements of the rule and meets the standards of, or complies with the requirements imposed or stated by, the federal law, regulation, or guidance document or law of another state identified by the rule under paragraph (2) of subsection (b) until the frontier developer or large chatbot provider declares the revocation of that intent to the Attorney General or the Attorney General revokes the rule in accordance with subsection (f).

(f) The Attorney General shall revoke a rule adopted under this Section if the conditions specified by this Section no longer apply.

Section 99. Effective date. This Act takes effect January 1, 2027.