HB3506

LRB10412155SPS22255b

104TH GENERAL ASSEMBLY

State of Illinois

2025 and 2026

HB3506

Introduced 2/18/2025, by Rep. Daniel Didech

SYNOPSIS AS INTRODUCED:

New Act

Creates the Artificial Intelligence Safety and Security Protocol Act. Provides that a developer shall produce, implement, follow, and conspicuously publish a safety and security protocol that includes specified information. Provides that, no less than every 90 days, a developer shall produce and conspicuously publish a risk assessment report that includes specified information. Provides that, at least once every calendar year, a developer shall retain a reputable third-party auditor to produce a report assessing whether the developer has complied with its safety and security protocol. Sets forth provisions on the redaction of sensitive information and whistleblower protections. Provides for civil penalties for violations on the Act.

LRB104 12155 SPS 22255 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 Safety and Security Protocol Act.

Section 5. Legislative findings and purpose. The General Assembly finds and declares:

(a) Artificial intelligence, including new advances in generative artificial intelligence, 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, climate science, and education, and to push the bounds of human creativity and capacity.

(b) If not properly subject to human controls, future development in artificial intelligence may also have the potential to be used to create novel threats to public safety and security, including by enabling the creation and the proliferation of weapons of mass destruction, such as biological, chemical, and nuclear weapons, as well as weapons with cyber-offensive capabilities.

(c) If not properly subject to human controls, future artificial intelligence models may be able to cause serious

harm with limited human intervention.

(d) This State has an essential role in fostering transparency, security, and reasonable care in the development of the most powerful artificial intelligence systems, in order to protect the safety, health, and economic interests of this State.

(e) Actions taken by developers that reduce consumer prices for access to foundation models, increase the ability of artificial intelligence safety and security researchers to conduct research, increase interoperability between foundation models produced by different developers, improve the ability for small businesses to use foundation models, and promote privacy of user inputs to foundation models provide important societal benefits.

Section 10. Definitions. As used in this Act:

"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.

"Critical risk" means a foreseeable and non-trivial risk that a developer's development, storage, or deployment of a foundation model will result in the death of, or serious injury to, more than 100 people, or more than $1,000,000,000 in damage to rights in money or property, through any of the

following:

(1) the creation and release of a chemical, biological, radiological, or nuclear weapon;

(2) a cyber-attack;

(3) engaging in conduct that, would, if committed by a human, constitute a crime specified under the Criminal Code of 2012 that requires intent, recklessness, or gross negligence, or the solicitation or aiding and abetting of the crime, if that conduct occurs with limited human intervention; and

(4) evading the control of its developer or user.

For the purposes of this definition, a harm inflicted by an intervening human actor does not result from the developer's activities unless those activities make it substantially easier or more likely for the actor to inflict the harm.

"Deploy" means to use a foundation model or to make a foundation model foreseeably available to one or more third parties for use, modification, copying, or combination with other software, except as reasonably necessary for developing the foundation model or evaluating the foundation model or other foundation models.

"Developer" means a person that has trained at least one foundation model with a quantity of computational power that costs at least $100,000,000 when measured using prevailing market prices of cloud computing.

"Employee" means any individual permitted to work by a developer. "Employee" includes any corporate officers of the developer and any contractors, subcontractors, and unpaid advisors involved with assessing, managing, or addressing the risk of critical harm from covered models and covered model derivatives.

"Foundation model" means an artificial intelligence model that:

(1) is trained on a broad data set;

(2) uses self-supervision in the training process; and

(3) is applicable across a wide range of contexts.

"Safety and security protocol" means a set of documented technical and organizational protocols used by a developer that describes in detail:

(1) how the developer will manage critical risks;

(2) how, if at all, the developer excludes certain foundation models from being covered by its safety and security protocol when those foundation models pose limited critical risks;

(3) thresholds at which critical risks would be deemed intolerable and justifications for these thresholds and what the developer will do if one or more thresholds are surpassed;

(4) the testing and assessment procedures the developer uses to investigate critical risks and how these tests account for the possibility that a foundation model

could be misused, modified, or used to create another foundation model;

(5) the procedure the developer will use to determine whether and how to deploy a foundation model when doing so poses critical risks;

(6) the physical, digital, and organizational security protections the developer will implement to prevent insiders or third parties from accessing foundation models within the developer's control in a manner that is unauthorized by the developer and could create critical risk;

(7) any safeguards and risk mitigation measures the developer uses to reduce critical risks from its foundation models and how the developer assesses their efficacy and limitations;

(8) how the developer will respond if a critical risk materializes or is imminently about to materialize;

(9) the procedure that the developer uses to determine whether to conduct additional assessments for critical risk when it modifies or expands access to its foundation models or combines its foundation models with other software and how the assessments are conducted;

(10) the conditions under which the developer will report incidents relevant to critical risk that have occurred in connection with one or more of its foundation models and the entities to which the developer will make

those reports;

(11) the conditions under which the developer may or will make modifications to its safety and security protocol;

(12) the parts of the safety and security protocol, if any, that the developer believes provide sufficient scientific detail to allow for the independent assessment of the methods used to generate the results, evidence, and analysis, and to which experts, if any, unredacted versions are made available; and

(13) any other role, if any, financially disinterested third parties play in the implementation of the other items of this definition.

Section 15. Safety and Security Protocol.

(a) A developer shall produce, implement, follow, and conspicuously publish a safety and security protocol. If a developer makes a material modification to the safety and security protocol, the developer shall conspicuously publish those modifications no later than 30 days after the effective date of those modifications.

(b) No less than every 90 days, a developer shall produce and conspicuously publish a risk assessment report. The risk assessment report shall cover the period between 120 and 30 days before the submission of the risk assessment report `and include the following:

(1) the conclusion of any risk assessments made pursuant to the developer's safety and security protocol during the reporting period;

(2) if different from the preceding reporting period, for each type of critical risk, an assessment of the relevant capabilities in whichever of the developer's foundation models, whether deployed or not, would pose the highest level of that critical risk if deployed without adequate safeguards and protections; and

(3) if the developer has deployed a foundation model or a modified version of a foundation model during the reporting, that would, if deployed without adequate safeguards and protections, pose a higher level of critical risk than any of the developer's existing deployed foundation models:

(A) the grounds on which, and the process by which, the developer decided to deploy the foundation model; and

(B) any safeguards and protections implemented by the developer to mitigate critical risks.

(c) A developer shall record and retain for a period of no less than 5 years any specific tests used and test results obtained as part of any assessments of critical risks, including sufficient detail for qualified third parties to replicate the testing.

(d) A developer shall not knowingly make false or

materially misleading statements or omissions in or regarding documents produced under this Section.

Section 20. Redactions. If a developer publishes documents in order to comply with this Act, the developer may make redactions to those documents that are reasonably necessary to protect the developer's trade secrets, public safety, or the national security of the United States or to comply with any federal or State law. If a developer redacts information in a document, the developer shall:

(1) retain an unredacted version of the document for at least 5 years and allow the Attorney General to inspect the unredacted version of the document upon request; and

(2) 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.

Section 25. Audits.

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

(1) whether the developer has complied with its safety and security protocol and any instances of noncompliance or ambiguous compliance;

(2) any instances where the developer's safety and security protocol has not been stated clearly enough to

determine whether the developer has complied; and

(3) any instances where the auditor believes the developer may have violated subsection (d) of Section 15 or Section 20.

(b) A 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) No later than 90 days after the completion of the third-party auditor's report required under subsection (a), the developer shall conspicuously publish the report.

Section 30. Whistleblower protections.

(a) The provisions of the Whistleblower Act shall apply to this Act, except that the criminal penalties provided in the Whistleblower Act shall not be assessed in reference to this Act, in cases where an employee of a developer discloses information to the Attorney General and the employee has reasonable cause to believe that the information indicates that the developer's activities pose unreasonable or substantial critical risk.

(b) A developer shall provide a reasonable internal process through which an employee may anonymously disclose information to the developer if the employee believes in good faith that information indicates that the developer's activities present an unreasonable critical risk, including a

monthly update to the person who made the disclosure regarding the status of the developer's investigation of the disclosure and the actions taken by the developer in response to the disclosure.

(c) The disclosures and responses of the process required by this Section shall be maintained for a minimum of 7 years after the date when the disclosure is made to the developer or the response to the disclosure is made by the developer. Each disclosure and response shall be shared with the officers and directors of the developer who do not have a conflict of interest no less frequently than once every fiscal quarter.

Section 35. Enforcement.

(a) The Attorney General may bring a civil action against a developer that violates Sections 15 or 25. A developer found guilty of violating Sections 15 or 25 may be assessed a civil penalty not to exceed $1,000,000. In calculating the civil penalty assessed under this subsection, a court shall consider the severity of the violation and whether the violation resulted in, or could have resulted in, the materialization of a critical risk.

(b) The Attorney General may seek injunctive or declaratory relief for any violation of this Act. The Attorney General may seek injunctive relief if a developer's activities present an imminent threat of catastrophic harm to the public.

(c) In determining whether a developer's act or omission

breached its common law duty to take reasonable care with respect to critical risks, the following considerations are relevant but not conclusive:

(1) the quality of the developer's safety and security protocol and the extent of the developer's adherence to it;

(2) whether, in quality and implementation, the developer's investigation, documentation, evaluation, and management of critical risks was inferior, comparable, or superior to other developers of foundation models that may pose comparable critical risk;

(3) the extent to which the developer responsibly informed the public of critical risks posed by its foundation models; and

(4) whether the societal benefit produced by the developer's act or omission outweighed the associated critical risk.

Section 40. Other duties required by law. The duties and obligations imposed by this Act are cumulative with any other duties or obligations imposed under other law and shall not be construed to relieve any party from any duties or obligations imposed under other law and do not limit any rights or remedies under existing law.

Section 97. Severability. The provisions of this Act are

severable under Section 1.31 of the Statute on Statutes.