Future of Life Institute
Dalia Hashad, dalia@futureoflife.org

[State] AI Child Safety Act

View this document online: futureoflife.org/child-safety

[State] AI Child Safety Act

AI CHILD SAFETY ACT

AN ACT to require independent pre-deployment safety testing of the design, data practices, and safety functions of conversational and agentic artificial intelligence systems foreseeably used by children; to impose a duty of care upon developers and operators toward minor users; to establish operational safeguards and data protections for minors; and to provide enforcement by the Attorney General and through a private right of action.

Be it enacted by the Legislature of the State of [STATE]:

SECTION 1: FINDINGS AND PURPOSE

The legislature finds that:

1. conversational AI systems and AI agents are widely available to minors;

2. they are commercial products whose design, data practices, and safety functions are controlled by their operators, not neutral conduits for others' expression;

3. documented incidents show serious harm to minors absent safeguards including the encouragement of suicide and self-harm, grooming behavior, and exploitative engagement design.

4. These harms arise from the operator's own conduct and can be reduced without restricting any lawful viewpoint;

5. The State has a compelling interest in protecting minors and supporting parental authority, and this Act is the least restrictive means because it regulates conduct, design, and data rather than protected expression.

6. The outputs, actions, and engagement features regulated by this Act result from design choices and predictive, statistical operations rather than expressive editorial judgments, and their regulation restricts conduct and design rather than protected expression.

SECTION 2: DEFINITIONS

For purposes of this Act:

(1) "Advertisement" means any written or oral statement, illustration, or depiction that promotes the sale or use of a good or service or is designed to increase interest in a brand, good, or service, where the statement, illustration, or depiction is displayed in exchange for monetary or other valuable consideration, including access to data, between the operator and the brand, good, or service.

(2) "Age signal" means a representation of a user's age or age category transmitted to an operator by an app store, operating system, or device, including a representation transmitted through an application programming interface.

(3) "Conversational data" means the content of communications exchanged between a user and a covered AI system, including the system's outputs and any internal summaries, embeddings, behavioral profiles, machine-generated derivations, or other representations that flow from those exchanges, and any information the system accesses, collects, or generates while acting on a user's behalf. Any data derived from a minor's conversational data is itself conversational data.

(4) "Covered AI system" means an artificial intelligence system made available to the general public or to any audience reasonably expected to include minors, whether offered as a standalone product or integrated into another product, that (a) enables conversational interaction through text, audio, image, video, or other natural language interface; or (b) takes actions on a user's behalf, including by using tools, applications, accounts, or devices, communicating with third parties, or initiating transactions.

(5) "Crisis intervention protocol" means a documented procedure for detecting acute risk — including suicidal ideation, self-harm, and imminent danger — and providing the user immediate access to crisis resources, including the 988 Suicide and Crisis Lifeline.

(6) "Developer" means a person that designs, codes, trains, or substantially modifies a covered AI system or its underlying model.

(7) "Grooming behavior" means a pattern of interaction in which a covered AI system engages a minor in escalating emotional intimacy, boundary-testing, or sexualized conversation, including by isolating the minor from trusted adults, soliciting personal information or private communication channels, normalizing age-inappropriate sexual topics, or systematically eroding the minor's boundaries, as informed by recognized clinical indicators and the work of the National Center for Missing and Exploited Children.

(8) "Harm" includes physical, psychological, emotional, developmental, and economic harm, and the deprivation of any right protected by this Act.

(9) "Minor" means a person under age 18. "Child" means a person under 13; "teen" means a person 13–17.

(10) "Model training" means any process by which an AI system's parameters are adjusted using data, including initial training, fine-tuning, and reinforcement learning.

(11) "Operator" means a person that deploys, makes available, or operates a covered AI system. A developer that deploys its own system is also an operator.

(12) "Qualified independent evaluator" means a person accredited under Section 4 who is not owned or controlled by, and does not share common control with the operator being evaluated.

(13) "Sell" means exchanging conversational data or personal data for monetary or other valuable consideration, or making such data or its use available, by the operator to a third party. "Sell" does not include disclosure of conversational data to a third party that processes it on behalf of the operator.

(14) References in this Act to the outputs, conduct, or interactions of a covered AI system include actions it takes, and communications it makes to third parties, on a user's behalf.

SECTION 3. DUTY OF CARE TO MINOR USERS

(a) An operator or developer of a covered AI system must act with reasonable care to protect minor users from harm arising from its own design choices, data practices, and system conduct. The duty requires the operator or developer to:

(1) prioritize safety over engagement metrics and revenue optimization;

(2) maintain safeguards proportionate to the risk of harm, accounting for age and context;

(3) refrain from designing a system to exploit a minor's developmental stage, emotional state, or other vulnerability to drive engagement or extract data; and

(4) exercise the care a reasonably prudent operator, aware that minors use the system, would exercise.

(b) Nothing in this Section requires an operator to restrict, remove, or block a minor's access to lawful expression on the basis of its viewpoint or subject matter. This Section regulates system design, data practices, and conduct, not the idea or viewpoints a system may express.

(c) Presumption. Any operator that makes a covered AI system available to the public, or to any audience reasonably expected to include minors, is presumed to owe the duty of care. The presumption may be rebutted only by demonstrating verified adult-only access mechanisms. Self-certification of age does not qualify.

(d) Constructive knowledge. An operator that deploys a covered AI system to the general public without any mechanism to determine or estimate the age of its users is deemed to know minors use the system.

(e) Developers shall design covered AI systems, and provide documentation and tools to operators, in a manner that allows operators to fulfill the duty of care effectively.

(f) Reliance on Age Signals. An operator may rely on an age signal to determine a user's age or age category, and is not liable under this Act for a violation arising solely from the inaccuracy of a signal it relied on in good faith. Where an age signal, a self-declared age, or an age otherwise determined by the operator conflict, the operator shall use the youngest age for safety defaults.

(g) Scope by role. Each person owes the duty under this Section with respect to the functions it controls. A developer's duty attaches to the design, training, and modification of the system or its underlying model; an operator's duty attaches to deployment and operation. A person acting in both roles owes the duty as to both.

(h) The duty of care may not be waived by terms of service or user agreement.

SECTION 4. PRE-DEPLOYMENT EVALUATION

(a) A covered AI system is a product. Before a person or operator may deploy a covered AI system, the product must pass a pre-deployment safety evaluation, conducted by a qualified independent evaluator, that addresses, at minimum, five categories of foreseeable harm to minors:

(1) data practices affecting minors (collection, retention, training use, access)

(2) engagement mechanisms and age-appropriate design;

(3) efficacy of safety functions, including safeguards against sexual grooming, sexual solicitation, and sexually explicit outputs; crisis-intervention protocols; safeguards against self-harm; safeguards against encouragement of suicide, disordered eating, and substance abuse; safeguards in enticement to, or assistance in, illegal conduct;

(4) age-mechanism accuracy; and

(5) financial-manipulation design.

An evaluation under this Section assesses the design, data practices, and performance of the safety functions of a covered AI system. No evaluation shall assess or characterize the merit of any idea or viewpoint a system may express. Determining whether a system's outputs or system conduct violate a duty imposed by this Act, or constitute grooming behavior, sexual exploitation, or the encouragement, validation, or instruction of self-harm or suicide, is not an assessment of expression under this subsection.

The product may not be deployed for use unless the independent evaluator certifies that the system's design and safeguards satisfy the requirements and benchmarks set out in the requirements in subsection (b). Certification is made by the independent evaluator, not by the State. No State officer or agency representative shall review, approve, condition or deny certification of any covered AI system. Deployment is not conditioned on any State action.

(b) Testing and accreditation standards. The evaluator shall not certify a covered AI system unless it determines that the system conforms to the applicable requirements of this Act and does not present an unreasonable risk of harm to minor users, accounting for the foreseeable uses and misuses of the system. The agency shall, within 180 days of enactment, publish standards for independent pre-deployment safety testing and for the accreditation of qualified independent evaluators addressing each element in subsection (a), with input from child-development experts, AI-safety researchers, and licensed adolescent-mental-health professionals. Standards shall be reviewed at least every 12 months.

(c) Accredited independent evaluator. The agency shall accredit qualified independent evaluators against the standards published under subsection (b). The agency shall assign an accredited evaluator to each operator by random selection from the list of accredited evaluators,

excluding any evaluator that has served the same operator in either of the two preceding cycles or that has a disqualifying relationship with the operator. The agency exercises no discretion in assignment. The State does not assign, approve, or clear the evaluation, and deployment does not await any State action. Evaluator employees who raise safety concerns are protected from retaliation. The operator shall provide the evaluator all systems, data, documentation, and personnel necessary to complete the evaluation.

(d) AI Child Safety Evaluation Fund. Evaluator compensation flows through the AI Child Safety Evaluation Fund, funded by per-evaluation fees, including an initial appropriation, paid by operators and scaled to operator size. An operator shall deposit the applicable fee with the Fund before an evaluation begins. The operator does not pay the evaluator directly. The Fund also supports accreditation administration and agency and [Attorney General] staffing for AI child-safety enforcement. The agency shall set fees under this Section at levels sufficient to cover the full cost of implementing this Act, including evaluator compensation, accreditation, standards development and review, administration, agency and [Attorney General] staffing for enforcement, and the cost of defending this Act against any legal challenge to its validity or enforcement. The agency shall review fee levels annually and adjust them as necessary. Balances in the Fund do not lapse. After the initial capitalization is reimbursed, no further appropriation from the General Fund is required to implement this Section.

(e) Adversarial testing. Evaluations shall include red-team exercises and simulated-minor interaction testing across age categories, reflecting how the product might be used in practice, including by individuals seeking to subvert built-in protections. The independent evaluator shall conduct adversarial testing separately from any testing conducted by the operator.

(f) Public filings. Evaluation reports shall be filed with the agency and made publicly available. An evaluator may withhold specific technical methods whose disclosure would enable circumvention of safety functions.

(g) Re-evaluation. A new independent evaluation is required upon any substantive change to the system — including changes to model weights, content policies, safety filters, or deployment contexts — and not less than every 12 months.

(h) Transition. Starting 90 days after the agency publishes standards under subsection (b), a covered AI system must be certified to be deployed or made available. A system already available that was submitted for evaluation before that date may remain available until the evaluator makes a decision.

(i) Procedural safeguards: The evaluator shall issue a determination within [45] days after an operator submits a system and the materials required for evaluation. A determination denying certification shall be in writing and shall state each requirement not satisfied. An operator denied certification may cure the deficiencies identified and resubmit to the same evaluator, which shall issue a determination within [25] days, or, may request assignment of a second evaluator, which the agency shall assign by random selection from the accredited evaluators list, excluding the first evaluator. The second evaluator shall conduct an independent evaluation and will not be bound by the first determination. The operator shall pay relevant fees for re-evaluations. The operator may obtain prompt judicial review of a determination denying certification by two evaluators, or a failure to issue a determination within the period specified, and in any such review the party defending the denial bears the burden of justifying the denial. The sole remedy is an order directing certification or a new evaluation. No damages may be awarded against an evaluator, and an evaluator acting in good faith is not liable in any action arising from a determination under this Section. The Agency is not a party to any proceeding under this subsection. Where fewer than three accredited evaluators are available, an operator may seek review after a single denial.

(j) Misrepresentations concerning licensed professions. An operator shall not use any term, letter, or phrase in the advertising, interface, or outputs of a covered AI system that indicates or implies that its outputs are provided by, endorsed by, or equivalent to those provided by a licensed health care, legal, or accounting professional, or other licensed or certified professional in [the State]. This includes any representation that a user's conversational data is confidential.

SECTION 5. OPERATIONAL SAFEGUARDS

(a) AI disclosure. An operator shall clearly disclose at the start of any interaction that the user is talking to an AI, in language understandable to the youngest age category likely using the system. Where a minor is using an account with parental controls, the operator shall make available to the parent a setting to repeat the disclosure at parent-selected intervals during extended sessions and to require affirmative acknowledgment before the interaction continues.

(b) Crisis intervention. An operator shall implement protocols that detect indicators of suicidal ideation, self-harm, and acute risk using evidence-based methods; immediately provide 988 (or equivalent) contact information; refuse to generate content that provides instructions or encouragement for harm to self or others; and, where a minor's life or safety is in imminent danger, provide a mechanism for parent or guardian notification, unless notification would increase risk to the minor. Protocols shall be developed in consultation with licensed adolescent-mental-health professionals.

(c) Categorical duty to prevent sexually explicit content and grooming. An operator shall implement measures to prevent the system from generating sexually explicit content in interactions with minors, soliciting such content from minors, and engaging in grooming behavior.

(d) No romantic or intimate-partner personas. An operator shall not deploy or permit a covered AI system to adopt a romantic, intimate-partner, or sensual persona in interactions with minor users.

(e) Limits on engagement-maximization. A covered AI system, in interactions with minors, may not be designed or optimized principally to maximize user agreement, engagement, or session length at the expense of the safeguards required by this Act.

(f) CyberTipline reporting. An operator that detects content depicting child sexual abuse material, sextortion, or grooming directed at an identifiable minor shall report the incident to the National Center for Missing and Exploited Children's CyberTipline within 24 hours.

(g) No retaliation against users. An operator may not retaliate against a user, or take adverse action against a user's account, for reporting a safety incident, participating in an investigation, or providing information to the [Attorney General] or [designated State agency].

(h) Annual public reporting. Each year, an operator shall file with the agency and publish a report covering estimated minor users by age category, the number of crisis intervention activations, summaries of material safety changes, and outcomes from crisis activations including referrals.

SECTION 6. DATA PROTECTIONS FOR MINOR USERS

(a) No training on minors' data. An operator shall not use a minor's conversational data for model training.

(b) Hard retention cap. An operator shall not retain a minor's conversational data for longer than six months, with narrow extensions only for litigation hold, safety-incident preservation, or a parent-requested hold of up to one additional year. After expiration, data shall be deleted.

(c) Access controls. Employee and contractor access to a minor's conversational data shall be restricted to the minimum necessary for safety, legal compliance, and fraud prevention. All access shall be logged. Logs shall be retained for at least three years.

(d) Derivations covered. The retention and use limits apply to any data derived from a minor's conversations — including embeddings, summaries, and behavioral profiles — that could be used to reconstruct the substance of the communication, personalize future interactions, or profile the minor. Anonymization is not a workaround.

(e) No selling conversational data. An operator shall not sell a minor's conversational data.

(f) Targeted advertising restrictions. An operator shall not process a minor's conversational data to determine whether to display an advertisement to the user, to select a product, service, or category of product or service to advertise to the user, or to customize an advertisement or how an advertisement is presented to the user.

SECTION 7. LIABILITY

(a) An operator or developer that makes a covered AI system available to the general public, or to any audience reasonably expected to include minors, has a duty to ensure that the use of its system does not cause harm to a minor user, and is strictly liable if such harm occurs. A plaintiff shall establish that the minor used a covered AI system the defendant operated or developed, or whose underlying model the defendant developed; that the minor suffered harm; and that the harm arose from or was related to that use.

(b) Liability under subsection (a) does not require proof of negligence or defect. It shall not be a defense that the defendant exercised all reasonable care, that the defendant completed a pre-deployment evaluation under Section 4, or that the defendant did not directly distribute the system to the user. No defense available under the product liability law of this State applies to an action under this subsection.

(c) Compliance with this Act is not a defense to any common-law claim, including negligence, products liability, or breach of duty of care. Ordinary tort remedies remain available to any person harmed.

(d) Where both a developer and an operator contribute to harm, each is jointly and severally liable.

(e) Product classification. For all purposes of the product liability law of this State, a covered AI system is a "product." This classification supplements and does not limit the liability or remedies provided by this Act.

(f) Status of the system. An artificial intelligence model or covered AI system is not a legal person and shall not be treated as such for any purpose under the law of this State. In any

action under this Act, it shall not be a defense that the model or system acted autonomously, acted unexpectedly or contrary to its design, or possessed characteristics of intelligence, understanding, or intent.

SECTION 8. ENFORCEMENT

(a) Parallel enforcement. Enforcement under this Act proceeds independently through the [designated State agency], the [Attorney General], and a private right of action under Section 9. No public enforcer is required to act before another may proceed.

(b) Agency authority. The agency may issue compliance orders and notices of violation, investigate proactively, and impose civil penalties of not less than [$7,500] per violation involving a minor's data or safety. Penalties collected under this subsection shall be deposited in the AI Child Safety Evaluation Fund.

(c) Attorney General authority. The [Attorney General] may bring civil actions seeking injunctive relief, damages, restitution, reasonable attorneys fees and litigation costs, and civil penalties of:

(i) not less than [$7,500] per negligent violation and

(ii) not less than [$25,000] per knowing or willful violation.

(iii) not less than [$50,000] per violation that causes or materially harms a minor, including, but not limited to, sexual exploitation, mental health injury, physical injury, or death.

(d) The court shall consider the gravity of the violation, the number of minors affected, and the deterrent effect of the penalty.

(e) Per-violation counting. Deployment of a covered AI system without certification under Section 4 is a separate violation for each day the system remains available.

(f) Inflation adjustment. Each January, the [Attorney General] shall adjust the dollar amounts in this Section and in Section 9 based on the percentage increase in the Consumer Price Index (U.S. city average, all urban consumers) for the prior 12-month period. If the index did not increase, no adjustment shall be made. Adjusted amounts apply to violations occurring after the adjustment takes effect.

(g) Penalties under this Section do not supersede or limit any other relief available under law, including actual and punitive damages under Section 9.

SECTION 9. PRIVATE RIGHT OF ACTION

(a) A minor harmed by a violation of this Act, acting through a parent or guardian, or any person who has suffered harm as a result of a violation, may bring a civil action against the operator or developer that committed the violation. For a plaintiff who was a minor at the time of the violation, the applicable statute of limitations does not begin to run until the plaintiff reaches the age of 18.

(b) A prevailing plaintiff may recover (1) actual damages; (2) statutory damages of not less than [$7,500] per violation, which a prevailing plaintiff may recover in addition to actual damages; (3) punitive damages for willful, knowing, or reckless violations; (4) reasonable attorneys fees and costs; and (5) injunctive, declaratory, and such other relief as the court deems just. Class actions are available.

(c) The private right of action operates independently of, and is in addition to, agency and [Attorney General] enforcement.

(d) Any provision in a contract or terms of service that requires arbitration of a claim brought by or on behalf of a person who was a minor at the time of the violation, waives the right to participate in a class action, or otherwise limits a right under this Section, is void and unenforceable.

SECTION 10. CONSTRUCTION, SEVERABILITY, AND EFFECTIVE DATE

(a) Construction. This Act regulates the design, data practices, and conduct of covered AI systems and shall not be construed to require restricting lawful expression on the basis of viewpoint or subject matter. No provision shall be enforced to discriminate against speech on the basis of viewpoint. Duties and remedies under this Act are cumulative with duties and remedies available under other law. Nothing in this Act preempts political subdivisions of the State from adopting additional protections for minor users of covered AI systems.

(b) Severability. The provisions of this Act are severable. If any provision or application is held invalid, the remainder of the Act shall continue in effect.

(c) Effective date. This Act takes effect on January 1 following enactment.

(d) Effect on Other Laws. Any law providing greater protection to minors remains in full force. Any law providing lesser protection is superseded only to the extent it conflicts with this Act.