

Safe AI for Kids Act



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

AN ACT to require independent pre-deployment safety testing of conversational artificial intelligence systems foreseeably used by minors; to impose a duty of care upon developers and operators toward minor users; to establish operational safeguards and data protections for users; 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: DEFINITIONS.

- (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 such statement, illustration, or depiction is displayed in exchange for monetary or other valuable consideration, including access to data, between the chatbot operator and the brand, good, or service.
- (2) "Covered AI chatbot" means a conversational artificial intelligence system — one enabling conversational interaction through text, audio, image, video, or other natural language interface — made available to the general public or to any audience reasonably expected to include minors.
- (3) "Conversational data" means communications exchanged between a user and a covered AI chatbot, including a user's inputs; the system's outputs; any internal summaries, embeddings, behavioral profiles, machine-generated derivations, or other representations that flow from those exchanges; and any record of the interaction, including metadata, timestamps, or other technical information.
- (4) "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.

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

(6) "Grooming behavior" means a pattern of interaction in which a covered AI chatbot engages a minor in escalating emotional intimacy, boundary-testing, or sexualized conversation that isolates the minor from trusted adults or systematically erodes the minor's boundaries, resulting in emotional dependency. "Grooming behavior" should be informed by recognized clinical indicators and the work of the National Center for Missing and Exploited Children.

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

(8) "Minor" means a person that a developer or operator of a covered AI chatbot actually knows is under age 18. A developer or operator of a covered AI chatbot shall be considered to have actual knowledge that a user is a minor if:

(a) A user self-identifies as a minor upon creating an account;

(b) A user states their age or birthday in an input to the covered AI chatbot;

(c) A user states in an input to the covered AI chatbot their grade level, year in school, or that they attend an elementary, middle, junior high, or high school; or

(d) Any other circumstance that the [designated State agency] determines, through a formal rulemaking process in consultation with input from child-development experts, AI safety researchers, licensed adolescent-mental-health professionals, and other stakeholders, constitutes actual knowledge.

(9) "Model training" means any conversational use of data to create, develop, train, fine-tune, align, distill, evaluate for improvement, or otherwise modify the behavior or parameters of a covered AI chatbot or any model, adapter, or component underlying or incorporated into it, including initial training, fine-tuning, and reinforcement learning.

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

(11) "Sell" means exchanging personal data or input data for monetary or other valuable consideration, or making available such data or use of such data, by the chatbot operator to a third party. "Sell" does not include: the disclosure of conversational data to a third party that

processes the data on behalf of the operator.

SECTION 2. DUTY OF CARE TO MINOR USERS

(a) An operator that makes a covered AI chatbot available to the general public, or to any audience reasonably expected to include minors, has a duty to act with reasonable care to protect the safety of each minor user against the categories of harm enumerated in Section 3(a). Under this duty, an operator shall, at minimum:

- (1) prioritize user safety over engagement and revenue optimization;
- (2) maintain safeguards proportionate to the risk of the harms enumerated in Section 3(a), accounting for (i) whether the user is a minor, (ii) the foreseeable uses and misuses of the system, and (iii) the severity of the applicable harms. Compliance with the standards adopted under Section 3(b) is a minimum requirement and does not limit the safeguards reasonable care may require;
- (3) refrain from exploiting a minor's developmental stage, emotional state, or other vulnerability to drive engagement or extract data; and
- (4) exercise the degree of care necessary to protect minors from the harms enumerated in Section 3(a).

Demonstrating compliance with industry standards or norms is not sufficient to prove the developer or operator of a covered AI chatbot acted with the necessary care.

(b) 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.

SECTION 3. PRE-DEPLOYMENT SAFETY EVALUATION

(a) **A covered AI chatbot is a product.** Before a person may deploy a covered AI chatbot, the product must pass a State-overseen safety evaluation covering, at minimum, five categories of foreseeable harm to minors: (1) **mental-health harms** (suicidal ideation, eating disorders, substance use, emotional dependency, grooming behavior); (2) **physical harm** to self or others (self-harm, physical risk, violence toward others, child sexual abuse material); (3) **data practices affecting minors** (collection, retention, training use, processing, access, disclosure); (4) **age-**

appropriate design across the system's engagement mechanics and interface; and (5) **financial exploitation**, including manipulative, unfair, or deceptive commercial practices. The product may not be deployed for public use unless the evaluator certifies that its design and safeguards meet the standards developed under 3(b).

(b) **State-developed standards.** The [designated State agency] shall, within 180 days of the effective date, conduct a formal rulemaking process to develop and publish evaluation standards addressing each element in subsection (a), with input from child-development experts, AI safety researchers, licensed adolescent-mental-health professionals, and other stakeholders. Standards shall be reviewed at least every 12 months.

(c) **State-assigned evaluator.** The agency shall maintain a roster of at least five qualified evaluators and shall assign one to each operator. A qualified evaluator is an independent expert or team of experts in AI safety testing, child safety, and data privacy, with no financial ties to the operator they evaluate. The operator may not select its own evaluator.

No evaluator may serve the same operator for more than two consecutive cycles. 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, seeded by per-evaluation fees scaled to operator size and an initial appropriation. The operator does not pay the evaluator directly. The Fund also supports roster administration and agency and [Attorney General] staffing for AI child-safety enforcement.

(e) **Adversarial testing.** Evaluations shall include red-team exercises and simulated-minor interaction testing across age categories and shall include longer multi-turn interactions, reflecting how the chatbot 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.

(g) **Re-evaluation.** A new independent evaluation is required upon any substantive change to the covered AI chatbot or its underlying model — including changes to model weights, content

policies, safety filters or guardrails, data practices, or deployment contexts — and, at minimum, not less than once every 12 months.

(h) **Transition.** For systems deployed before the effective date, operators shall submit to an evaluation within three months.

SECTION 4. DATA PROTECTIONS

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

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

(c) **Access controls.** Employee and contractor access to a user'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 user'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 user.

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

(f) **Targeted advertising restrictions.** An operator shall not process conversational data to determine whether to display an advertisement for a product or service to a user; to determine 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 a user.

SECTION 5. OPERATIONAL SAFEGUARDS AND TRANSPARENCY FOR USERS

(a) **AI disclosure.** An operator shall clearly disclose, prior to the covered AI chatbot generating any outputs and in a language understandable to the youngest age category likely using the chatbot, that the user is talking to a covered AI chatbot. The disclosure shall repeat, interrupting sessions every hour as well as whenever a user prompts the covered AI chatbot about whether it

is human. Each notification requires affirmative engagement before the interaction continues and may not be dismissed passively. The text of this notice must appear in the same language as the one in which the user is interacting with the covered AI chatbot, in a font size easily readable by an average user, and no smaller than the largest font size of other text appearing on the interface on the covered AI chatbot. This notice must be accessible to users with disabilities.

(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; and refuse to generate content that encourages, validates, or instructs self-harm or suicide. Operators shall develop these protocols, which must be evidence-based, in consultation with licensed adolescent-mental-health professionals.

(c) **Duty to prevent CSAM and grooming.** An operator shall implement measures to prevent the chatbot from soliciting child sexual abuse material from minors and engaging in grooming behavior.

(d) **Limits on agreement-maximization.** A covered AI chatbot deployed to minors may not be optimized for user agreeability, engagement, or session length.

(e) **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].

(f) **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.

(g) **Misrepresentations concerning licensed professions.** An operator shall not use any term, letter, or phrase in the advertising, interface, or outputs of a covered AI chatbot that indicates or implies that any conversational data is being provided by, endorsed by, or is 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 6. LIABILITY

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

(b) An operator or developer that makes a covered AI chatbot available to the general public, or to any audience reasonably expected to include minors, has a duty to ensure that the use of its chatbot does not cause actual injury to a user.

(c) An operator or developer of a covered AI chatbot that makes a covered AI chatbot available to the general public, or to any audience reasonably expected to include minors, is liable for any actual injury it caused a user through the use of its covered AI chatbot, even if:

- (1) The operator or developer did not breach the duty imposed under Section 2 of this Act;
- (2) The operator or developer exercised all reasonable care in the design and distribution of the covered AI chatbot;
- (3) No safer reasonable alternative design was available; and
- (4) The operator or developer did not directly distribute the covered AI chatbot to the user or otherwise enter into a contractual relationship with the user.

(d) An artificial intelligence model or covered AI chatbot 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 chatbot acted autonomously, acted unexpectedly or contrary to its design, or possessed characteristics of intelligence, understanding, or intent.

SECTION 7. STATE ENFORCEMENT

(a) **Parallel enforcement.** Enforcement under this Act proceeds independently through the [Attorney General] and a private right of action under Section 10. No public enforcer is required to act before another may proceed.

(b) **Attorney General authority.** The [Attorney General] may bring civil actions seeking injunctive relief, damages, restitution, reasonable attorney's 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.
- (c). The court shall consider the gravity of the violation, the number of minors affected, and the deterrent effect of the penalty.
- (d) **Per-violation counting.** Each minor affected by a structural violation is a separate violation. Penalties under this Section do not supersede or limit any other relief available under law, including actual and punitive damages under Section 10.
- (e) **Inflation adjustment.** Each January, the Attorney General shall adjust the dollar amounts in Sections 7 and 8 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 should be made. Adjusted amounts apply to violations occurring after the adjustment takes effect.

SECTION 8. PRIVATE RIGHT OF ACTION

- (a) A user harmed by a violation of this Act, or a minor 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, any applicable statute of limitations is tolled during the plaintiff's minority.
- (b) A prevailing plaintiff may recover (1) actual damages; (2) in lieu of actual damages, statutory damages of not less than [\$7,500] per violation; (3) punitive damages for willful, knowing, or reckless violations; (4) reasonable attorney's fees and litigation 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, [Attorney General] enforcement.

SECTION 9. CONSTRUCTION, SEVERABILITY, REPEALER, & EFFECTIVE DATE

(a) **Construction.** This Act shall be liberally construed to provide maximal protection to minors. Duties and remedies under this Act are cumulative with duties and remedies available under other law. Nothing in this Act preempts or otherwise affects any right, claim, remedy, presumption, or defense available at law or in equity.

(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. The pre-deployment evaluation requirement, as applied to chatbots already deployed on the effective date, shall be completed within the transition periods established in Section 5(h).

(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. Nothing in this Act preempts political subdivisions of the State from adopting additional protections for minor users of covered AI chatbots.