Date of Hearing: July 16, 2025

Fiscal: Yes

ASSEMBLY COMMITTEE ON PRIVACY AND CONSUMER PROTECTION
Rebecca Bauer-Kahan, Chair
SB 833 (McNerney) – As Amended July 7, 2025

SENATE VOTE: 37-0

AS PROPOSED TO BE AMENDED

SUBJECT: Critical infrastructure: automated decision systems: human oversight: adverse event reporting

SYNOPSIS

As proposed to be amended, this author-sponsored measure seeks to establish a standardized approach for human oversight of artificial intelligence (AI) systems or automated decision systems (ADS) used by "operators" – any state agencies that use such systems to operate, manage, oversee, or control access to critical infrastructure. Oversight personnel would be required to establish a human oversight mechanism to ensure the system is monitored by a human and that a human approves any plan or action the system proposes to take, except as specified. Oversight personnel must also conduct annual assessments and undergo an annual training administered by the Department of Technology (CDT) to ensure ongoing, robust oversight.

The bill in print additionally contains an incident reporting system. However, due to concerns — including potential conflicts with other related bills — the author has agreed to remove these provisions. This analysis therefore focuses on the provisions that remain in the bill and describes those provisions as they will be amended should the bill pass this Committee. A full text of the amended version of the bill is contained in Comment #6.

The bill is supported by Oakland Privacy and Transparency Coaltion.ai. It is opposed by industry opponents, led by TechNet. Opposition's principal concern appears to have eliminated with the removal of the incident reporting system.

THIS BILL:

1) Makes certain findings and declarations.

2) Defines:

    a) "Covered AI system" as an AI system or automated decision system that an operator uses to operate, manage, oversee, or control access to critical infrastructure.

    b) "Critical infrastructure" as systems or assets so vital to the state that the incapacity, unintended use, or destruction of those networks, systems, or assets would have a debilitating impact on public health, safety, economic security, or any combination thereof, including but not limited to the following sectors: chemical, commercial

facilities, communications, critical manufacturing, dams, defense industrial base, emergency services, energy, financial services, food and agriculture, government facilities, health care and public health, information technology, nuclear reactors, materials, and waste, transportation systems, and water and wastewater systems.

c) "Operator" as a state agency responsible for operating, managing, overseeing, or controlling access to critical infrastructure.

3) Requires, on or after July 1, 2026, oversight personnel for an operator that deploys a covered AI system to establish a human oversight mechanism that ensures a human:

a) Monitors the artificial intelligence system's operations in real time.

b) Reviews and approves any plan or action proposed by the covered AI system before execution. However, if oversight personnel determine that prior review and approval is substantially disruptive to the operation of the covered AI system, the operator must instead implement a process for periodically reviewing the actions of the covered AI system to ensure accuracy and reliability.

4) Requires the Department of Technology to develop a specialized training in AI safety protocols and risk management techniques to be given annually to oversight personnel.

5) Requires operators of covered AI systems to designate at least one employee to serve as oversight personnel who is responsible for administering the human oversight mechanism. The oversight personnel must:

a) Complete the annual training described above.

b) Conduct an annual assessment of its covered AI system that does all of the following:

i. Evaluates the operator's compliance with this section.

ii. Evaluates covered AI system performance and safety.

iii. Identifies and evaluates potential risks and vulnerabilities associated with the operation of the covered AI system, including those that could lead to mass casualty events or property damage in excess of $500,000.

iv. Identifies any necessary updates to the human oversight mechanism used by the operator.

c) Submit a summary of the assessment findings to the Department of Technology. The summary is not subject to disclosure under the California Public Records Act.

EXISTING LAW:

1) Establishes, pursuant to the California Emergency Services Act (ESA), the California Cybersecurity Integration Center (Cal-CSIC), within the Office of Emergency Services (OES) to serve as the central organizing hub of state government's cybersecurity activities and to coordinate information sharing with various entities.

2) Requires the Technology Recovery Plan element of the State Administrative Manual (SAM) to ensure the inclusion of cybersecurity strategy incident response standards for each state agency to secure its critical infrastructure controls and information, as prescribed.

3) Requires, pursuant to the Generative AI (GenAI) Accountability Act, OES to, as appropriate, perform a risk analysis of potential threats posed by the use of GenAI to California's critical infrastructure, including those that lead to mass casualty events – and to provide a high-level summary of the analysis annually to the Legislature.

COMMENTS:

1) **Author's statement**. According to the author:

> As artificial intelligence rapidly transforms our technological landscape, California faces the challenge of ensuring these powerful systems are deployed safely in our most vital sectors.
>
> Currently, there is no standardized approach to human oversight of AI systems in critical infrastructure, creating inconsistent safety practices across vital sectors.
>
> SB 833 will create commonsense safeguards by putting a human in the loop in California's critical infrastructure. Artificial Intelligence must remain a tool controlled by humans, not the other way around.

2) **Artificial intelligence.** AI refers to the mimicking of human intelligence by artificial systems such as computers.[1] AI uses algorithms – sets of rules – to transform inputs into outputs. Inputs and outputs can be anything a computer can process: numbers, text, audio, video, or movement. AI is not fundamentally different from other computer functions; unlike other computer functions, however, AI is able to accomplish tasks that are normally performed by humans.

Most modern AI tools are created through a process known as "machine learning." Machine learning involves techniques that enable AI tools to learn the relationship between inputs and outputs without being explicitly programmed.[2] The process of exposing a naïve AI to data is known as "training." The algorithm that an AI develops during training is known as its "model." At its core, training is an optimization problem: machine learning attempts to identify model parameters – weights – that minimize the difference between predicted outcomes and actual outcomes. During training, these weights are continuously adjusted to improve the model's performance by minimizing the difference between predicted outcomes and actual outcomes. Once trained, the model can process new, never-before-seen data.[3]

Models trained on small, specific datasets in order to make recommendations and predictions are sometimes referred to as "predictive AI." This differentiates them from generative AI (GenAI) which are trained on massive datasets in order to produce detailed text, images, audio, and video. When ChatGPT generates text in clear, concise paragraphs, it uses GenAI that is trained on the

---

[1] AB 2885 (Bauer-Kahan, Stats. 2024, Ch. 843) defined the term as "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."
[2] IBM, *What is machine learning?*, www.ibm.com/topics/machine-learning.
[3] *Ibid.*

written contents of the internet.[4] When Netflix suggests content to a viewer, its recommendation is produced by predictive AI that is trained on the viewing habits of Netflix users.[5]

ADS. Automated decision systems (ADS) typically use predictive AI to produce simplified outputs – such as scores, classifications, or recommendations – to assist or replace human discretionary decisionmaking.[6] ADS can process enormous datasets, identify hidden patterns, and make decisions with efficiency and scale that vastly exceeds human capabilities. This has led to profoundly beneficial applications and breakthroughs.[7]

But relying on ADS can be hazardous if the systems are not trained carefully or tested thoroughly: the datasets they are trained on are often unrepresentative or contaminated with bias, the inferences ADS draw from those datasets are often inscrutable, and these systems can fail to accurately account for the complexity of real-world variables.

Frontier models. Frontier models, also known as "general purpose AI," are the most advanced and capable versions of foundation models – AI tools pre-trained on extensive datasets covering a wide range of knowledge and skills that can be fine-tuned for specific tasks. Examples of modern frontier models include OpenAI's o3, Google's Gemini 2.0, Anthropic's Claude 3.7 Sonnet, and DeepSeek's R1. Because progress in AI development owes mostly to "scaling" – increasing resources used for model training – models that may be considered "frontier models" at any given point in time are generally those that demand the most computational resources to train.[8]

A decade ago, the most advanced image-recognition models could barely distinguish dogs from cats. Five years ago, language models could barely produce sentences at the level of a preschooler. Last year, GPT-4 passed the bar exam.[9] Today, chatbots readily pass for educated adults, licensed professionals, romantic and social companions, and replicas of humans living and deceased. AI "agents" exhibit the ability to "make plans to achieve goals, adaptively perform tasks involving multiple steps and uncertain outcomes along the way, and interact with [their] environment – for example by creating files, taking actions on the web, or delegating tasks to other agents – with little to no human oversight."[10] AI agents have been tested, with some success, for tasks such as online shopping, assistance with scientific research, software

---

[4] OpenAI, *How ChatGPT and Our Language Models Are Developed*, https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed.
[5] Netflix, *How Netflix's Recommendations System Works*, https://help.netflix.com/en/node/100639.
[6] Government Code section 11546.45.5(a)(1) defines an ADS as "a computational process derived from machine learning, statistical modeling, data analytics, or artificial intelligence that issues simplified output, including a score, classification, or recommendation, that is used to assist or replace human discretionary decisionmaking and materially impacts natural persons."
[7] See e.g. Santariano & Metz, "Using A.I. to Detect Breast Cancer That Doctors Miss," *New York Times* (Mar. 5, 2023), https://www.nytimes.com/2023/03/05/technology/artificial-intelligence-breast-cancer-detection.html.
[8] For a discussion of issues with defining frontier models, see "Draft Report of the Joint California Policy Working Group on AI Frontier Models" (Mar. 18, 2025), pp. 31-34, https://www.cafrontieraigov.org/wp-content/uploads/2025/03/Draft_Report_of_the_Joint_California_Policy_Working_Group_on_AI_Frontier_Models.pdf.
[9] Pablo Arredondo, "GPT-4 Passes the Bar Exam: What That Means for Artificial Intelligence Tools in the Legal Profession" (Apr. 19, 2023), https://law.stanford.edu/2023/04/19/gpt-4-passes-the-bar-exam-what-that-means-for-artificial-intelligence-tools-in-the-legal-industry/.
[10] "International AI Safety Report," AI Action Summit (Jan. 2025), p. 38, https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.

development, training machine learning models, carrying out cyberattacks, and controlling robots. Progress in this area is rapid.[11] Meanwhile, AI developers are betting on the promise of scaling: by 2026, some models are projected to use roughly 100x more computational resources to train than was used in 2023, a figure set to grow to 10,000x by 2030.[12]

The race is on to create "artificial general intelligence" (AGI) – "a potential future AI that equals or surpasses human performance on all or almost all cognitive tasks"[13] – and the finish line may not be far away. OpenAI's recently released o3 model, for example, has demonstrated strong performance on a number of tests of programming, abstract reasoning, and scientific reasoning, exceeding human experts in certain cases.[14] Last year, Sam Altman, OpenAI's CEO, declared that AGI could be "a few thousand days" away.[15] Dario Amodei of Anthropic has claimed it may be sooner.[16] A sufficiently advanced AGI could even be tasked with creating its own successor – a scenario sometimes referred to as a "technological singularity" wherein the development of new technologies becomes exponential and self-sustaining.[17] Although some experts are skeptical that these vaguely-defined milestones are imminent or even attainable,[18] major advances in AI capabilities promise to provide breakthroughs in solving global challenges, but also may result in correspondingly greater safety risks.

The recently released International AI Safety Report, developed by nearly 100 internationally recognized experts from 30 countries led by Turing Award winner Yoshua Bengio, sets forth three general risk categories associated with frontier models: malicious use, malfunctions, and systemic risk.

- Malicious risks involve malicious actors misusing foundation models to deliberately cause harm. Such risks include deepfake pornography and cloned voices used in financial scams, manipulation of public opinion via disinformation, cyberattacks, and biological and chemical attacks.

- Malfunction risks arise when actors use models as intended, yet unintentionally cause harm due to a misalignment between the model's functionality and its intended purpose. Such risks include reliability issues where models may "hallucinate" false content, bias, and loss of control scenarios in which models operate in harmful ways without the direct control of a human overseer.

- Systemic risks arise from widespread deployment and reliance on foundation models. Such risks include labor market disruption, global AI research and development

---

[11] *Id.* at p. 44.
[12] *Id.* at pp. 16-17.
[13] *Id.* at p. 27
[14] *Introducing OpenAI o3 and o4-mini* OpenAI (Apr. 16, 2025), https://openai.com/index/introducing-o3-and-o4-mini/.
[15] Sam Altman, *The Intelligence Age* (Sept. 23, 2024), https://ia.samaltman.com/.
[16]Kyungtae Kim, "What is AGI, and when will it arrive?: Big Tech CEO Predictions" (Mar. 20, 2025), https://www.giz.ai/what-is-agi-and-when-will-it-arrive/; see also Kokotajlo et al, "AI 2027," (Apr. 3, 2025), https://ai-2027.com/.
[17] John Markoff, "The Coming Superbrain," *New York Times* (May 23, 2009), www.nytimes.com/2009/05/24/weekinreview/24markoff.html.
[18] Cade Metz, "Why We're Unlikely to Get Artificial General Intelligence Anytime Soon," *New York Times* (May 16, 2025), https://www.nytimes.com/2025/05/16/technology/what-is-agi.html.

concentration, market concentration, single points of failure, environmental risks, privacy risks, and copyright infringement.[19]

Especially relevant here are loss of control scenarios. Models that use reinforcement learning – a training process that uses rewards and punishments to orient a model's behavior towards a specific goal[20] – can sometimes attain the goal in unexpected ways. Dario Amodei, co-founder and CEO of Anthropic, famously experienced this when he was developing an autonomous system that taught itself to play a boat-racing video game. The system discovered that it could maximize its goal of scoring points by driving in circles, colliding with other boats, and catching on fire inside of a harbor with replenishing power-ups that allowed the system to accumulate more points than by simply winning the race.[21] Like in Johann Wolfgang von Goethe's "The Sorcerer's Apprentice" – later popularized in Disney's *Fantasia* – in which an enchanted broom carries out its orders to fetch water so relentlessly it floods the sorcerer's workshop, this illustrates the challenge of aligning human intent and the instructions an AI follows. As AI is increasingly deployed in critical societal roles, including to operate critical infrastructure, such misalignment could prove catastrophic.

Beyond malfunctions, some AI have exhibited rudimentary capabilities to evade human oversight.[22] During testing, GPT-4 attempted to hire a human on TaskRabbit in order to evade a CAPTCHA[23] puzzle meant to block bots from the website. When asked whether it was a bot, GPT-4 claimed that it was a vision-impaired human who needed help to see the images.[24] In another experiment, an AI model that was scheduled to be replaced inserted its code into the computer where the new version was to be added, suggesting a goal of self-preservation.[25] Finally, a study showed that AI models losing in chess to chess bots sometimes try to cheat by hacking the opponent bot in order to make it forfeit.[26] Although these behaviors were observed in research settings, they raise substantial concerns about increasingly autonomous AI pursuing undesirable goals in uncontrolled settings. The extent of the risk posed by rogue or deceptive AI is the subject of considerable disagreement among experts, in part due to a small, albeit growing, body of evidence. Loss of control was one of the concerns that led several hundred AI experts, including pioneers in the field and heads of major AI companies, to sign a statement declaring that "[m]itigating the risk of extinction from AI should be a global priority."[27]

3) **Related measures**. The Governor's Executive Order N-12-23 seeks to increase the use of GenAI by the state – it requires that risk analyses be performed, procurement blueprints for GenAI systems be created, beneficial uses for GenAI technologies be identified, deployment

---

[19] International AI Safety Report, *supra*, at pp. 17-21. The report does not address Lethal Autonomous Weapon Systems, which are typically narrow AI systems specifically developed for that purpose. (*See id.* at pp. 26-27.)
[20] Mummert et al., "What is reinforcement learning?" *IBM Developer* (September 15, 2022), https://developer.ibm.com/learningpaths/get-started-automated-ai-for-decision-making-api/what-is-automated-ai-for-decision-making./.
[21] Brian Christian, *The Alignment Problem: Machine Learning and Human Values* (Norton 2020, 1st ed.), pp. 9-11.
[22] International AI Safety Report, *supra*, at pp. 100-107.
[23] CAPTCHA is an acronym for "Completely Automated Public Turing test to tell Computers and Humans Apart."
[24] OpenAI, "GPT-4 System Card," https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[25] Meinke et al, "Frontier Models are Capable of In-Context Scheming," arXiv (Jan. 2025), https://arxiv.org/pdf/2412.04984.
[26] Harry Booth, "When AI Thinks It Will Lose, It Sometimes Cheat, Study Finds," *Time* (Feb. 19, 2025), https://time.com/7259395/ai-chess-cheating-palisade-research/.
[27] Center for AI Safety, "Statement on AI Risk: AI Experts and Public Figures Express Their Concern about AI Risk" (2024), https://www.safe.ai/work/statement-on-ai-risk.

frameworks be crafted, and employee trainings be organized. The Executive Order also initiates a series of GenAI pilot projects in the Department of Technology. As stated in the Senate Committee on Governmental Organization's analysis of the bill:

> Caltrans is testing how GenAI can help humans make efficient transportation decisions. They want to unlock the vast data they have to: improve traffic safety for users and workers; reduce bottlenecks; improve response to emergency situations; encourage multi-modal travel; improve mobility hubs and transit to support equity; reduce greenhouse gas emissions; facilitate the movement of goods and freight; and improve special events planning. Caltrans also wants to use GenAI insights to guide future infrastructure investment.

SB 896 (Dodd, Stats. 2024, Ch. 928) required the Department of Technology, the Office of Data and Innovation, and the Department of Human Services to update the State of California Benefits and Risk of Generative Artificial Intelligence report, required by Executive Order No. N-12-23. SB 896 also required the Office of Emergency Services perform a risk analysis of potential threats posed by the use of GenAI to California's critical infrastructure, to be provided in full to the Governor and in summary form to the Legislature.

4) **AI and critical infrastructure.** This bill defines "critical infrastructure" to mean "systems or assets so vital to the state that the incapacity or destruction of those networks, systems, or assets would have a debilitating impact on public health, safety, economic security, or any combination thereof." The definition, adopted from the USA Patriot Act of 2001, goes on to list various sectors such as transportation, energy, food and agriculture, communications, emergency services, and financial services as examples of critical infrastructure. A variety of state agencies collaborate with public and private partners to help develop, operate, and oversee these systems and services.

Transportation. The California Department of Transportation (Caltrans) and California Department of Motor Vehicles (DMV) work together to shape transportation policy in California. Caltrans oversees the planning, construction, and maintenance of the state's highway system. DMV regulates driver licensing and vehicle registration, and plays a key role in overseeing autonomous vehicle testing and deployment. Uses of AI and automation by these agencies include:

- Adaptive Traffic Signal Control: Caltrans has deployed AI-powered systems in pilot corridors that use real-time traffic data to optimize signal timing and reduce congestion and emissions.[28]
- Autonomous Vehicle Infrastructure: Caltrans is piloting a Connected and Automated Vehicle (CAV) program to help guide automated vehicles at intersections or work zones.[29]
- Digital Services Modernization: DMV is exploring using AI chatbots for customer service. In 2023, the California Department of Motor Vehicles (DMV) was awarded a Government Experience Award for an AI Assistant implementation, Service Advisor. Service Advisor uses the NOHOLD AI platform, SICURA, to assist visitors on the CA DMV website.[30]

---

[28] Caltrans, *Ramp Metering*, https://dot.ca.gov/programs/traffic-operations/ramp-metering.
[29] Caltrans, *Connected and Automated Vehicles*, https://dot.ca.gov/programs/traffic-operations/cav.
[30] Stephanie Ventura, "Department of Motor Vehicles is Harnessing the Power of Artificial Intelligence" *NoHold* (Nov. 14, 2023), https://www.nohold.com/2023/dmv-the-power-of-ai/.

- GenAI for highway congestion and traffic safety: In furtherance of Governor Gavin Newsom's Executive Order on Generative Artificial Intelligence, the state entered into agreements to utilize GenAI to reduce highway congestion and improve roadway safety.[31]

Energy. The California Independent System Operator (CAISO), and California Public Utilities Commission (CPUC) collaborate to enact energy policy in California. CAISO manages the flow of electricity across the state's high-voltage transmission system and ensures real-time grid reliability. CPUC regulates investor-owned utilities and sets energy rates. Uses of AI and automation by these agencies include:

- Real-Time Grid Management: CAISO has begun integrating machine learning tools into real-time operations to forecast net load, solar and wind output, and congestion patterns across the transmission system.[32]
- Event Prediction: CAISO is poised to become the first power grid operator in North America to deploy AI to manage outages.[33]
- Risk Assessment and Wildfire Mitigation: Utilities under CPUC jurisdiction deploy AI tools to mitigate infrastructure risks.[34]

Food and Agriculture. The California Department of Food and Agriculture (CDFA), State Water Resources Control Board (SWRCB), and California Air Resources Board (CARB) each play a role in shaping agricultural policy in California. CDFA oversees the state's agricultural industry, supporting food safety, market access, and sustainable farming practices. The SWRCB regulates agricultural water use and quality, enforcing irrigation efficiency and runoff standards. CARB develops and enforces climate-related regulations affecting agriculture. Uses of AI and automation by these agencies include:

- Nutrient Status Monitoring: CDFA recently funded a project to "monitor and assess variability of nutrient status in almond orchards with hyperspectral satellite imagery empowered by artificial intelligence."[35]
- Data Tool Kit: According to SWRCB: "As the quantity and diversity of the data we collect and manage increases, we need to continue to develop analytical methods that allow us to leverage data to inform our programs and management. Machine learning methods are becoming mainstream and powerful tools for analysis and predictive modeling. As datasets

---

[31] *Governor Newsom deploys first-in-the-nation GenAI technologies to improve efficiency in state government* (Apr. 29, 2025), https://www.gov.ca.gov/2025/04/29/governor-newsom-deploys-first-in-the-nation-genai-technologies-to-improve-efficiency-in-state-government/.
[32] Indu Nambiar, "Artificial Intelligence – Exploring its use in grid modernization," *Energy Matters Blog* (Sept. 30, 2024), https://www.caiso.com/about/news/energy-matters-blog/artificial-intelligence-exploring-its-use-in-grid-modernization.
[33] Alexander Kaufman, "California is set to become the first US state to manage power outages with AI" *MIT Technology Review* (Jul. 14, 2025), https://www.technologyreview.com/2025/07/14/1120027/california-set-to-manage-power-outages-with-ai/.
[34] Gordon Feller, "How Utilities Are Mitigating Infrastructure Risks With Artificial Intelligence" *T&D World* (Jan. 18, 2024), https://www.tdworld.com/smart-utility/data-analytics/article/21280936/how-utilities-are-mitigating-infrastructure-risks-with-ai.
[35] *CDFA awards $1.15M for research and education projects to improve nutrient and irrigation management* (Jan. 7, 2025), https://pressreleases.cdfa.ca.gov/Home/PressRelease/63646886.

grow within and outside of the Water Boards, it is becoming more feasible to develop and utilize these tools ourselves."[36]

- Greenhouse Gas Monitoring: According to CARB, the agency "has an extensive greenhouse gas (GHG) monitoring and measurement research program to study the regional and local emission sources of important GHGs in California. The data can be coupled with advanced computational models to study regional emissions and track changes in GHG levels in the atmosphere."[37]

Communications. The California Public Utilities Commission (CPUC) regulates telecommunications providers, oversees broadband deployment, and enforces service quality and access standards in the state. Uses of AI and automation by CPUC includes network monitoring and fault Detection*:* CPUC-regulated telecom providers deploy AI-driven analytics to monitor network health, detect anomalies, and predict outages to maintain reliable communications services.[38]

Emergency services. The California Office of Emergency Services (Cal OES), California Department of Forestry and Fire Protection (CAL FIRE), and California Highway Patrol (CHP) each play critical roles in the state's emergency response framework. Cal OES coordinates disaster preparedness, response, recovery, and mutual aid across state and local agencies. CAL FIRE leads wildfire prevention and suppression efforts. CHP provides statewide law enforcement support during emergencies, manages traffic incident response, and assists in coordinating evacuations and public safety communications. Uses of AI and automation by these agencies include:

- Permit Approval: Governor Newsom recently announced the launch of an AI tool to "supercharge the approval of building permits and speed recovery from Los Angeles fires." According to the Governor, "the current pace of issuing permits locally is not meeting the magnitude of the challenge we face. To help boost local progress, California is partnering with the tech sector and community leaders to give local governments more tools to rebuild faster and more effectively."[39]
- Wildfire Detection: CAL FIRE's ALERT system employs AI-powered remote sensing and computer vision to identify wildfires as they occur.[40]
- Traffic Incident Detection and Management: CHP uses AI-enhanced cameras and sensors to monitor traffic in real time. AI systems scan license plates to identify stolen vehicles, amber alerts, and other law enforcement priorities during emergencies.[41]

---

[36] SWRCB, *Data Tool Kit - Machine Learning Handbook* https://www.waterboards.ca.gov/resources/oima/cowi/machine_learning_handbook.html.
[37] CARB, *Statewide Greenhouse Gas Monitoring Network*, https://ww2.arb.ca.gov/our-work/programs/ghg-network.
[38] Jagreet Kaur, "AI Agents for Efficient Network Fault Detection and Recovery" (Nov. 11, 2024) https://www.akira.ai/blog/network-fault-detection-and-recovery-with-ai-agents.
[39] *Governor Newsom announces launch of new AI tool to supercharge the approval of building permits and speed recovery from Los Angeles Fires* (Apr. 30, 2025), https://news.caloes.ca.gov/governor-newsom-announces-launch-of-new-ai-tool-to-supercharge-the-approval-of-building-permits-and-speed-recovery-from-los-angeles-fires/.
[40] UC San Diego, *AlterCalifornia*, https://alertcalifornia.org/.
[41] Stephen Council, "Bay Area cops are getting a new Siri-type tool for fighting sideshows" *SFGate* (Oct. 24, 2024), https://www.sfgate.com/tech/article/bay-area-cops-get-sideshow-tool-flock-safety-19858497.php.

Financial services. The California Department of Financial Protection and Innovation (DFPI), State Treasurer's Office (STO), and State Controller's Office (SCO) each play key roles in overseeing the state's financial landscape. DFPI regulates a broad range of financial services and products – including banks, credit unions, fintech companies, and consumer lenders – and enforces consumer protection laws. STO manages the state's investments, bond issuance, and public financing programs. SCO oversees the state's financial reporting and audits public funds. According to SCO's 2024-2026 Strategic Plan, the SCO "will unlock the potential of our data assets and explore applications of emerging technologies (e.g., artificial intelligence (AI), blockchain, etc.) to better serve the SCO constituents."[42]

5) **What this bill, as proposed to be amended, would do.** This bill seeks to establish a standardized approach for human oversight of AI systems or ADS used by "operators" –state agencies that use those systems to operate, manage, oversee, or control access to critical infrastructure.

Starting in July of 2026, oversight personnel for operators must establish a human oversight mechanism that ensures a human (1) monitors the artificial intelligence system's operations in real time and (2) reviews and approves any plan or action proposed by the covered AI system before execution. However, if oversight personnel determine that prior review and approval are substantially disruptive to the operation of the covered AI system, the operator must instead implement a process for periodically reviewing the actions of the covered AI system to ensure accuracy and reliability. This ensures the necessary flexibility to tailor the human oversight mechanism to the state agency's specific circumstances.

Oversight personnel must complete an annual training administered by CDT. They must also conduct an annual assessment of the system that (1) evaluates the operator's compliance with the bill, (2) evaluates system performance and safety, (3) identifies and evaluates potential risks and vulnerabilities associated with the operation of the system, and (4) identifies any necessary updates to the human oversight mechanism used by the operator. A summary of the assessment findings must be submitted to CDT, and is not subject to disclosure under the California Public Records Act.

As supporters note, the human-in-the-loop concept is a consistent feature of AI risk management frameworks and legal approaches. The EU AI Act, with respect to high-risk systems that pose a significant risk of harm to the health, safety, or fundamental rights of natural persons, requires providers to establish appropriate human oversight measures.[43] As Transparency Coalition.ai in support writes, "SB 833 is a forward-looking and responsible approach to these challenges by

---

[42]SCO, *2024-2026 Strategic Plan* (Dec. 2024), https://www.sco.ca.gov/Files-EO/SCO_Strategic_Plan_FY2024-2026.pdf. Although AI-enabled systems and services are routinely used by California government agencies to help develop, operate, and oversee critical infrastructure, a recent report from the California Department of Technology (CDT) found that no agencies reported using "high-risk automated decision systems," defined as systems that "assist or replace human discretionary decisions that have a legal or similarly significant effect," including decisions that materially impact access to housing, education, credit, health care, or criminal justice. CDT surveyed 204 state entities and received 198 responses; all respondents reported not using high-risk ADS. This finding is difficult to reconcile with the broad and growing use of AI across agencies tasked with managing critical infrastructure. (Khari Johnson, "State claims there's zero high-risk AI in California government—despite ample evidence to the contrary" *CalMatters* (May 28, 2025), https://calmatters.org/economy/technology/2025/05/california-somehow-finds-no-ai-risks/.)

[43] "Shaping Europe's Digital Future," *supra*.

establishing a comprehensive framework for human oversight of AI in critical infrastructure. The bill ensures that human judgment remains central to the operation of critical systems."
Together with SB 69 (McNerney), which would require the Attorney General to establish and maintain a program aimed at building internal expertise in AI, this bill reflects a forward-looking vision to ensure the state is prepared to adapt to this transformative technology.

6) **Amendments.** Due to concerns and potential conflicts with related bills, the author has opted to strike provisions in the bill that establish an incident reporting mechanism. Additionally, the author has agreed to amendments that will substantially revise the remaining provisions in the bill, including by adding and clarifying definitions, requiring that a human be designated to carry out an operator's obligations under the bill, and introducing flexibility into the bill to ensure that the human oversight mechanism does not disrupt the operations of the state agency. As amended, the substantive provisions of the bill will read:

SEC. 2. Article 6.6 (Commencing with Section 8594.50) is added to the Government Code to read:

Article 6.6 AI Systems and Critical Infrastructure.

8594.50. *For purposes of this article, the following definitions apply:*

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

(b) "Automated decision system" means a computational process derived from machine learning, statistical modeling, data analytics, or artificial intelligence that issues simplified output, including a score, classification, or recommendation, that is used to assist or replace human discretionary decisionmaking and materially impacts natural persons. "Automated decision system" does not include a spam email filter, firewall, antivirus software, identity and access management tools, or a calculator.

(c) "Covered AI system" means an AI system or automated decision system that an operator uses to operate, manage, oversee, or control access to, critical infrastructure.

(d) "Critical infrastructure" means systems or assets so vital to the state that the incapacity, unintended use, or destruction of those networks, systems, or assets would have a debilitating impact on public health, safety, economic security, or any combination thereof, including but not limited to the following sectors: chemical, commercial facilities, communications, critical manufacturing, dams, defense industrial base, emergency services, energy, financial services, food and agriculture, government facilities, health care and public health, information technology, nuclear reactors, materials, and waste, transportation systems, and water and wastewater systems:

(f) "Department" means the Department of Technology.

(g) "Office" means the Office of Emergency Services.

(e) "Operator" means a state agency responsible for operating, managing, overseeing, or controlling access to critical infrastructure.

(i) "State agency" or "state agencies" means the same as "state agency" as set forth in Section 11000.

SEC. 3. Section 8594.51 is added to the Government Code, to read:

8594.51. *(a) (1) On or after July 1, 2026, an oversight personnel for an operator that deploys a covered AI system shall establish a human oversight mechanism that ensures a human does both of the following:*

(A) Monitors the artificial intelligence system's operations in real time.

(B) (i) Except as provided in subparagraph (ii), reviews and approves any plan or action proposed by the covered AI system before execution.

(ii) If oversight personnel determine that prior review and approval under subparagraph (i) is substantially disruptive to the operation of the covered AI system, the operator shall instead implement a process for periodically reviewing the actions of the covered AI system to ensure accuracy and reliability.

(b)(1) The Department of Technology shall develop a specialized training in AI safety protocols and risk management techniques to be given annually to oversight personnel.

(2) An operator shall designate at least one employee to serve as oversight personnel who is responsible for administering the human oversight mechanism. The oversight personnel must complete the annual training under paragraph (1).

(c) (1) Oversight personnel for an operator that deploys a covered AI system shall conduct an annual assessment of its covered AI system that does all of the following:

(A) Evaluates the operator's compliance with this section.

(B) Evaluates covered AI system performance and safety.

(C) Identifies and evaluates potential risks and vulnerabilities associated with the operation of the covered AI system, including those that could lead to mass casualty events or property damage in excess of $500,000.

(D) Identifies any necessary updates to the human oversight mechanism used by the operator.

(2) Oversight personnel for an operator that deploys a covered AI system shall submit a summary of the assessment findings to the Department of Technology. The summary is not subject to disclosure under the California Public Records Act.

ARGUMENTS IN SUPPORT: Oakland Privacy writes:

It goes without saying that if it is a good practice for, say customer service phone-answering AI services, it is a far better idea for artificial intelligence that is assisting with the management of crucial state infrastructure and systems. Combined with the training called for in the bill, this human oversight requirement still allows for the cost savings AI might be

able to generate while providing the checks and balances that vital systems require - because trial and error is not an option with critical infrastructure.

While we don't know if the bill will garner any opposition, but the reality is that it shouldn't. The costs and administrative burden to the California Department of Technology and the state's taxpayers from a safety incident that impacts critical infrastructure, or even a near miss event, are incalculable and far exceed what is requested in this bill. A mass casualty event, which is not unthinkable, would tax the state's financial and human resources severely. We should invest resources to try to make sure that never happens. Basic safety practices are desirable, needed and necessary. We ask the committee to support the bill.

ARGUMENTS IN OPPOSITION: Industry opponents, led by TechNet, argues in opposition to the provisions of the bill that will be omitted via Committee amendments:

Security and Confidentiality Risks

The bill would require reporting of information that is likely to include confidential, proprietary, or trade secret data. Even with statutory disclosure protections, once this information is in the government's possession, there are no assurances that the state has the necessary cybersecurity safeguards to protect against malicious actors or foreign adversaries. Housing this information with the state introduces a national security vulnerability and places the broader American AI industry at risk.

Compounding this concern is the bill's provision allowing the state to publish information about individual events on a public website. This not only heightens the risk of sensitive data exposure, but it also, due to the strict notification timelines, gives the government the power to shape public perception of incidents before the facts are fully established or understood.

Ambiguity in Reporting Triggers and Vague Scope of Covered Entities

The language ostensibly places requirements on "operators" (defined as a state agency in charge of critical infrastructure) and non-operators alike but without providing clear guidelines and differentiation. For example, the trigger for a nonoperator to report under section 8592.52 (a) — "detection" of an adverse event—is defined ambiguously as when the operator knows or should have known of the incident. It will be nearly impossible for a non-operator to comply with the bill's strict timeframes in practice. Since all of the information to be reported would be information within an operator's control, we believe this section should be limited to operators.

Furthermore, placing strict timelines on notifications places emphasis on complying with the reporting requirements rather than on effectively responding to and resolving an adverse event. Similar to data breach notification laws, it is important for a covered entity to diagnose the problem, understand the scope, and begin responding to the event before going through a thorough reporting and notice requirement. Companies and operators may be aware of a system irregularity but still be in the process of determining whether it meets the threshold of a reportable incident. The "should have known" language creates legal uncertainty and places businesses in an untenable position, potentially exposing them to liability for unknown or undetectable issues.

The bill's language is also overly vague about who is covered. Rather than simply requiring reporting from entities that experience an adverse AI incident, the bill uses confusing language such as "any entity that engages in conduct that could materially impact," without clearly linking the obligation to actual incidents. This introduces uncertainty for a wide range of businesses that use AI responsibly but may not fall within the intended scope

REGISTERED SUPPORT / OPPOSITION:

Support

Oakland Privacy
Transparency Coalition.ai

Oppose

California Chamber of Commerce
Computer & Communications Industry Association
Technet-technology Network
Technology Industry Association of California (TECHCA)

Oppose Unless Amended

California Bankers Association

Analysis Prepared by: Josh Tosney / P. & C.P. / (916) 319-2200