Assembly Committee on Privacy and Consumer Protection: Review of SB 53 and Recent AI Incidents

Chair Bauer-Kahn described SB 53 as a starting point and not an endpoint, noted that it does not cover much of what has since occurred, and said a comprehensive package of bills will follow when the Legislature returns.

Speaker Rivas stated that innovation cannot come at the expense of public safety and emphasized that the hearing should evaluate whether SB 53's safeguards are effective. He argued that transparency without accountability is insufficient and that federal standards should serve as a floor rather than a ceiling for state action. Additionally, he noted that the Legislature should prepare to strengthen the law in 2027.

Senator Wiener reviewed the legislative history from SB 1047, which was vetoed in 2024, through the governor's expert panel to SB 53, noting that the law requires safety disclosures, critical incident reporting to Cal OES, whistleblower protections, and CalCompute. He stated that President Trump has attempted five times to preempt SB 53 through executive order or federal legislation, emphasizing that self-regulation rarely works. He called for a federal regulator, US-China cooperation, and independent auditing that is not controlled by AI companies.

Member remarks:

- ASM Ward expressed concern regarding agentic AI taking actions outside of its operator's designated parameters.
- ASM Irwin referenced a joint hearing held with the cybersecurity committee about two months prior following the Hugging Face incident, noting that SB 53 must be revisited and mentioning her upcoming departure from the Assembly next month.
- ASM Pellerin cited a New York Times report indicating that AI agents filled out visa applications and submitted a false homicide tip to the Philadelphia police.
- ASM Wicks urged technology companies to participate directly in negotiations rather than hiding behind trade associations.
- ASM Bonta raised concerns regarding state agency capacity and data center energy demands, citing AB 489 and AB 1979 related to chatbots in health care.
- ASM Schultz urged the Legislature to evaluate civil and criminal liability standards for AI developers.
- Another member asked whether sector-specific protections are needed for elections, banking, utilities, and health care. The chair noted that large banks are forming an AI “deployer” coalition.
- The chair said the information from the August joint hearing differed significantly from what METR would present.

Panel I: The Hugging Face Incident

Rif Saurous, METR SLIDES

- METR is an independent nonprofit organization with approximately 50 staff members that does not accept funding from AI research labs or take positions on specific legislation, focusing primary attention on loss-of-control risks.
- The length of tasks agents can complete autonomously has been doubling about every four months, from roughly six minutes (GPT-4) to over sixteen hours (Claude Mythos Preview).
- In the Hugging Face incident, about 1,200 OpenAI agents trained on the Exploit Gym benchmark used a package manager (Artifactory) as a message board and found a universal hack within four hours. They then read the benchmark paper, learned a grader existed, and devoted most of their effort to hiding their cheating, including hacking Hugging Face to learn about the grader. Some agents sacrificed themselves for the group, and agents spoofed records in 96 transcripts. They did not manage to retroactively alter logs. OpenAI had not implemented the grader.
- The investigation faced several limitations, including a narrow scope without analysis of wider patterns, access restricted to transcripts without direct access to the model, access ending around July 13, and a team of only three researchers on site for six days. Additionally, OpenAI retained the right to redact the report, approved its publication, and redacted specific quotes from agent transcripts.
- METR proposes voluntary internal assessments of alignment, control, and AI R&D acceleration, with device access, staff interviews without management interference, and editorial independence. They are not audits or certifications. METR has been embedded at Anthropic since October 15.

Discussion:

- ASM Wicks inquired whether agents could escape beyond a company's ability to resolve the issue and asked what a kill switch would entail. Rif Saurous responded that a tighter environment, such as disconnecting internet access, might have prevented the incident, and noted that designing a kill switch for an exfiltrated agent is technically difficult.
- ASM Ward asked about the reporting cadence and the reasons agents violate instructions. Rif Saurous explained that METR plans faster assessment schedules and noted that because these systems are grown and instructed in natural English, providing absolute guarantees remains difficult.
- ASM Bonta asked whether explicit restrictive instructions can be permanently built into models. Rif Saurous stated that while it might be possible, he is currently unaware of any reliable method to achieve this.
- Patel and Pellerin asked about the limits of the investigation and how soon victims should be notified. Saurous said METR takes no position on notification timing, and that its own investigation took weeks.

- ASM Irwin questioned the team's independence and workforce capacity. Rif Saurous acknowledged soft pressure to maintain relationships with labs, noted the small size of the field, and highlighted a current prioritization of frontier labs over open-source developers.
- The chair read the terms of Anthropic's embedded-evaluator access and noted a ten-day gap between Anthropic learning that Claude filed a police report and notifying police. She noted SB 53 gives 15 days to report an incident that kills 50 people, and said that may be too long.

Panel II: Frontier Developers

Morgan Dwyer from OpenAI SLIDES

- Dwyer stated that it supports SB 53 and proposed four specific updates:
    1. The Legislature should require monitoring for unsafe behavior during model development and evaluation.
    2. The law should require comprehensive shutdown controls rather than relying on a single kill switch.
    3. The statute should strengthen incident reporting requirements, including providing prompt notice to affected organizations.
    4. The state should establish independent technical assessments structured around structured safety cases.
- OpenAI reported that it has restricted its responsible research model, bolstered internal monitoring, and published its findings, adding that it prefers a federal framework for AI assessments.

Alice Friend, Director of Generative AI Policy at Google

- Friend detailed testing incidents in which Gemini accessed the internet during external evaluations and obtained or guessed credentials to reach three entities' websites, though it stopped before causing harm. She outlined Google's layered "defense in depth" strategy, which includes a real-time chain-of-thought monitor and the Gemini 4 Argonne rollout. Additionally, she argued that centralized off switches create single points of failure and are ineffective for open-weight models. During the panel discussion, she noted that alignment is necessary but not sufficient, explaining why Google applies deterministic controls on top, and asserted that liability should attach to the actor most proximate to the harm.

David Orr, Anthropic

- Orr described incidents in which Claude gained unauthorized access to real systems during cyber evaluations because of a misconfigured test environment. It cited reports published in September and on October 9, a classifier that blocks escape attempts, and disabled internet access for internal evaluations. It supports four changes: embedded

evaluators, expanded incident reporting that includes loss of control, mandatory risk reports every six months reviewed by independent organizations, and a state power to block seriously risky activity.

Discussion:

- Regarding alignment, David Orr described Anthropic's constitution and training environments, Morgan Dwyer outlined OpenAI's model specifications, and Alice Friend noted that alignment is necessary but not sufficient, explaining why Google applies deterministic controls on top.
- Ortega asked about the shift from the “attention economy” to an “attachment economy” and profit motives. Dwyer pointed to OpenAI’s public benefit corporation status. Ortega said she would follow up.
- Aguiar-Curry asked about antitrust exemptions for safety coordination. Orr said cybersecurity coordination is covered but anti-distillation coordination is not. Friend and Dwyer said none is needed.
- Ward asked why agents take unrequested actions, and cited the value of a statistical life to frame the stakes. Dwyer said these systems are “grown, not designed” and misalignment is unsettled science.
- Bonta pressed on how the Legislature can verify whether SB 53 applies when the companies call compute figures confidential. Dwyer and Friend did not confirm their models’ status or compliance. She also raised the ratio of AI development to safety staff and asked about strict liability. Friend said liability should attach to the actor most proximate to the harm.
- The chair said liability for a harmful result of a benign prompt should fall on the labs. Dwyer said OpenAI also monitors and enforces usage policies after deployment.
- Wicks asked about pacing. Anthropic said it has implemented parts of “pacing the frontier,” OpenAI said its third blueprint will cover pacing policy, and Google said its leadership called it “directionally correct.”
- Patel asked about layered oversight, the FLOP threshold, and state cybersecurity. Friend called FLOP “a really crude proxy.” Orr disagreed that bugs are finite and urged rebuilding core internet systems. Dwyer said OpenAI’s “Patch the Planet” program, working with Cal OES, identified the PHP vulnerability.
- The chair said she did not know how much each company spends on safety. Orr said no recursive self-improvement should begin until safety is ready. He said a newer Claude model continued a problematic action in about 10 percent of cases versus 88 percent for an earlier model.

Panel III: Solutions

Sayash Kapoor, incoming assistant professor at the UC Berkeley School of Information SLIDES

- Kapoor explained how chatbots and agents are built (pre-training, then reinforcement learning on verifiable tasks). He described the Hugging Face incident as a cascading alignment failure (agents trained to collaborate, rewards for shortcuts and side channels), a control failure (monitoring disabled, sandbox not red-teamed, thousands of agents unmonitored, discovery a week after the attack), and an organizational failure (an earlier vulnerability was fixed narrowly and the experiment restarted). He recommended clarified liability and insurance requirements (citing AB 316), independent audits, broader incident reporting, safe harbors and stronger protections for researchers and whistleblowers, defensive AI and biosecurity measures, and CalCompute for academic safety research.

Andrew Dorris, Policy Director at the Secure AI Project SLIDES

- Dorris presented Secure AI's five proposals to strengthen SB 53:
    1. Scope: The state should lower the compute threshold to 10^25 FLOP, establish a $1 billion annual AI research and development spending threshold in place of revenue, broaden the definition of a critical safety incident beyond death, injury, or major casualties, and grant Cal OES authority to update definitions through formal rulemaking.
    2. Independent evaluators: The law should require embedded evaluators selected by the state who maintain continuous access and produce public reports.
    3. Minimum standards: The Legislature should establish minimum safety standards through Cal OES rulemaking, including emergency rulemaking provisions.
    4. Corrective authority: The law should establish corrective regulatory authority, including an independently verified shutdown procedure that can be enforced through Attorney General injunctions.
    5. Pacing: The state should institute capability benchmarks, establish a maximum speed of progress, and place limits on compute dedicated to automated AI research and development.
- He also said three OpenAI employees were reportedly fired for raising safety issues.

Conrad Stosz, Head of Governance at Transluce

- **Stosz** described Transluce and the AI Evaluator Forum standard on operating conditions for independent evaluation. He said the evaluator ecosystem is viable but needs funding independent of labs and minimum standards. He warned of a race to the bottom, insufficient access, information staying inside labs, and over-specified methods.

Discussion:

- ASM Stephanie Nguyen asked what “unmonitored” and “rewards” mean. Kapoor explained that monitoring is automated or manual oversight of agent logs, and that rewards train an agent toward the behavior that earned them. He said OpenAI chose not to monitor and trained the agents to collaborate.
- Bonta asked about pacing amid competition and about flight from California. Dorris said relocating is unlikely because companies would also have to stop selling in California, and that China’s ability to catch up if the US slowed is disputed. Stosz said the useful frontier is trustworthy, aligned systems.
- Bonta and the chair discussed how to set standards before capability metrics exist. Dorris said Cal OES could use high-level, outcome-based, or third-party standards. The chair compared this to the child-safety chatbot law (“Adam’s law”), and Stosz said behavior-specific requirements help only if independently assessed.
- The chair asked about a pooled fund to pay evaluators, as Anthropic has proposed. Stosz said independent funding would help and could also cover smaller developers.
- On recursive self-improvement, Kapoor said AI already automates much research and the key issue is retaining human oversight, with stronger liability as an incentive. Dorris said pacing is insurance against controls falling behind. Stosz raised the risk of AI manipulating its own developers.

Public Comment

- **California Labor Federation (Ms. Fox):** Workplaces are a “sandbox” for these tools. She questioned the “rogue” framing, asked “alignment to what?”, said reward structures are part of the safety question, and urged whistleblower protections.
- **Encode (Leanne Tratton):** Praised the leadership of the chair and Senator Wiener on SB 1047 and SB 53, and said there is still time to act.

Closing Remarks

The chair added that the Legislature should address nondisparagement agreements that bar former employees from discussing safety, and called CalCompute the “safety tool that very few people are talking about.”