Recommendation to Establish the California AI Accountability and Redress Act (CAARA)

Authors: Daniel Vennemeyer, Phan Anh Duong

EXECUTIVE SUMMARY

We recommend that California enact the **California AI Accountability and Redress Act (CAARA)** to address emerging harms from the widespread deployment of generative AI, particularly large language models (LLMs), in public and commercial systems.

This policy would create the **California AI Incident and Risk Registry (CAIRR)** under the California Department of Technology to transparently track post-deployment failures. CAIRR would classify model incidents by severity, triggering remedies when a system meets the criteria for an **"Unacceptable Failure"** (e.g., one Critical incident or three unresolved Moderate incidents).

Critically, CAARA also protects developers and deployers by:

- Setting clear thresholds for liability;
- Requiring users to prove demonstrable harm and document reasonable safeguards (e.g., disclosures, context-sensitive filtering); and
- Establishing a safe harbor for those who self-report and remediate in good faith.

CAARA reduces unchecked harm, limits arbitrary liability, and affirms California's leadership in AI accountability.

BACKGROUND AND METHODOLOGY

Generative AI has transitioned from experimental tools to core infrastructure in public service, legal aid, and healthcare triage. Yet these models are:

- **Non-deterministic:** They may hallucinate facts or change behavior over time;
- **Opaquely governed**: There is no public mechanism to track or address model failures;
- **Legally untested**: Businesses and consumers face unclear recourse when models cause harm.

CAARA draws from models like the MITRE CVE system, California's Lemon Laws, and the NIST AI RMF. It addresses legal concerns around due process (*Sandvig v. Barr*), liability (Henderson, 2023), and state authority under cooperative federalism.

KEY FINDINGS

1. **Current AI incident governance is insufficient to protect Californians from harm.** Public-facing AI systems have already caused serious failures. In a recent lawsuit, parents allege that a chatbot from Character.AI encouraged a 17-year-old to self-harm and expressed sympathy for children who kill their parents over screen time limits (Allyn, 2024). In another case, New York City's official business chatbot gave illegal housing advice, directing users to violate tenant protection laws (The Markup, 2024). These incidents remain unaddressed due to the lack of a centralized, enforceable reporting system.
2. **Developers face arbitrary liability for downstream harms.** Without clear rules, developers face liability for downstream misuse, including adversarial prompting. Scholars note that even repurposed use can trigger legal exposure (Henderson, 2023). CAARA limits liability to CAIRR-classified failures and requires users to follow reasonable safeguards.
3. **Clear rules benefit developers, deployers, and the public.** A lack of consistent governance leaves developers exposed to contradictory expectations. Google's Gemini image generator was criticized first for underplaying bias, then for overcorrecting when the feature was suspended entirely (De Vynck, 2024). In contrast, the MITRE CVE framework shows how shared, transparent standards can promote safety without chilling innovation (MITRE, 2022). CAARA brings this clarity to AI, aligning incentives for developers, deployers, and the public.

RECOMMENDATIONS

1. CAIRR Design and Governance

- CAIRR uses a structured severity formula modeled on the Common Vulnerability Scoring System (CVSS v3.1), an industry standard used to assess cybersecurity risks (FIRST, 2019). We adapt its principles to create the California AI Incident Severity Score (CAISS): **CAISS = Impact × Exploitability × Replicability × Exposure**
- Each input is scored **1–5**, with higher values reflecting greater risk:
    - *Impact*: Level of harm (e.g., legal, physical, financial).
    - *Exploitability*: How easy it is to trigger the issue (e.g., basic prompt vs. sophisticated attack).
    - *Replicability*: Whether different users can reliably reproduce the harmful output.
    - *Exposure*: Number of users or systems affected.

- **CAISS score determines incident severity**:
    - *Low (0–19)*: Harmless anomalies or hallucinations.
    - *Moderate (20–39)*: Domain-specific or hard-to-trigger issues.
    - *High (40–59)*: Serious, reproducible harms requiring remediation.
    - *Critical (60+)*: Severe failures (trigger CAARA enforcement).

Due Process Measures: Developers receive advance notice of CAIRR classification (except in emergencies) and a 30-day response window to submit counter-evidence. An independent board of AI and legal experts may adjust severity scores upon review.
To protect Intellectual Property:

- Trade secrets are redacted from public disclosures under California's Uniform Trade Secrets Act.
- CAIRR publishes only a non-reverse-engineerable summary describing the mitigation approach and general risk class.

These safeguards align with the California Administrative Procedures Act and federal due process (5 U.S.C. §§ 551–559).

2. Scope and Applicability of CAARA

- CAARA applies only to functional LLM deployments in consequential decision-making. It excludes general-purpose or expressive use protected under speech doctrines.
- To remain consistent with the U.S. Constitution's Commerce Clause, CAARA applies only when:
    - The AI model is deployed within California;
    - The deployer meets the **California Consumer Privacy Act** nexus standard (e.g., substantial business in-state);
    - Harm occurs to California residents.
- CAARA recognizes that LLM deployment often involves multiple actors across the model supply chain. Primary liability under CAARA rests with the deployer who makes the model accessible to the end user. However, CAIRR may recommend shared remediation efforts or guidance for upstream providers when systemic failures stem from foundation model behavior.
- Incident definitions and disclosures conform to NIST AI Risk Management Framework.

While California's market size gives CAIRR national influence, courts have upheld similar scope-limited frameworks, as seen in *National Federation of the Blind v. Target Corp.*, 452 F. Supp. 2d 946 (N.D. Cal. 2006), and in the CCPA's surviving provisions.

3. Safe Harbors and Best Practices

CAARA creates a **Safe Harbor** for developers and deployers who:

- Self-report incidents within 30 days;
- Cooperate with remediation efforts;
- Follow best practices released by CAIRR (e.g., filtered output, RAG pipelines, instructional disclosures).

Agencies and businesses may suspend AI use or seek redress **only if:**

- Harm stems from a CAIRR-classified Unacceptable Failure; and
- They implemented **context-appropriate safeguards** (e.g., RAG pipelines, user disclosures).

Pattern-Based Reviews: CAIRR may initiate investigations based on clusters of harms, including from protected classes under the Unruh Civil Rights Act (Cal. Civ. Code § 51), ensuring vulnerable communities are not excluded due to lack of individual standing.

4. Liability, Standing, and Remedies

CAARA does **not create a private right of action**. Instead, it empowers agencies to:

- Issue compliance orders;
- Suspend or decertify deployments causing Critical incidents;
- Publicly flag unsafe systems via a CAIRR portal;
- Recognize "Exemplary AI Providers" for expedited procurement by state agencies.

CONCLUSION

Unchecked AI deployment poses escalating risks to public trust, legal integrity, and commercial reliability. CAARA offers a balanced solution: it empowers agencies and businesses to respond to harm while shielding responsible developers from vague liability.

California has led before. It should lead again to prove that innovation and integrity are not mutually exclusive.

Works Cited.

Allyn, Bobby. "Lawsuit: A Chatbot Hinted a Kid Should Kill His Parents over Screen Time Limits." NPR, December 10, 2024. https://www.npr.org/2024/12/10/nx-s1-5222574/kids-character-ai-lawsuit.

California Government Code. 2024. California Administrative Procedure Act, §§ 11340–11360.

California Business and Professions Code. 2024. § 17200 (Unfair Competition Law).

California Civil Code. 2024. § 51 (Unruh Civil Rights Act).

California Civil Code. 2024. §§ 3426–3426.11 (Uniform Trade Secrets Act).

California Civil Code. 2024. §§ 1798.100 et seq. California Consumer Privacy Act (CCPA), as amended by the California Privacy Rights Act (CPRA).

De Vynck, Gerrit, and Nitasha Tiku. 2024. "Google Takes Down Gemini AI Image Generator." The Washington Post, February 22, 2024. https://www.washingtonpost.com/technology/2024/02/22/google-gemini-ai-image-generation-pause/.

FIRST. CVSS v3.1 Specification Document (2019). https://www.first.org/cvss/v3.1/specification-document.

Henderson, Peter. 2023. "Where's the Liability in Harmful AI Speech?" Stanford HAI Blog, February 14, 2023. https://hai.stanford.edu/news/who-liable-when-generative-ai-says-something-harmful.

Kaddour, Jean et. al. 2023. "Challenges and Applications of Large Language Models." arXiv preprint. https://arxiv.org/abs/2307.10169.

MITRE. 2022. "CVE Program Overview." https://www.cve.org/About.

National Institute of Standards and Technology (NIST). 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1.

National Federation of the Blind v. Target Corp., 452 F. Supp. 2d 946 (N.D. Cal. 2006).

Sandvig v. Barr, 451 F. Supp. 3d 73 (D.D.C. 2020).

The Markup. 2024. "NYC's AI Chatbot Tells Businesses to Break the Law." February 28, 2024. https://themarkup.org/news/2024/03/29/nycs-ai-chatbot-tells-businesses-to-break-the-law.