Safe and Accountable Federal Evaluations of Artificial Intelligence Act (SAFE AI Act)

Reps. Sara Jacobs, Celeste Maloy, and Don Beyer

Background

Currently, the companies building frontier AI systems create their own safety rules with no clear guidance from the federal government. Additionally, no one outside the company can reliably verify whether the rules are actually being followed. When the federal government does act, there is currently no transparency around the criteria upon which a model is halted. The lack of independent oversight is even more significant before a model is publicly deployed. Frontier AI systems are increasingly deployed inside the companies themselves, including to help build the next generation of AI, where there is little outside visibility even though those uses have already shown the potential to create risks to the public. If a frontier model turns out to be dangerous, the federal government has no reliable way to find out, no ability to check the company's assurances, and no authority to act until public safety is already at risk.

The SAFE AI ACT:

Mandatory safety frameworks and disclosures: Frontier developers must publish and adhere to safety frameworks to ensure transparency and accountability:

- **Developers publish enforceable safety frameworks:** Frameworks contain what each frontier company is testing for, what thresholds will prompt mitigations, and what those mitigations involve. Frameworks must be filed and published to ensure transparency and accountability. These cannot be unilaterally loosened.
- **Minimum government requirements:** The Center for AI Standards and Innovation (CAISI), in collaboration with interagency partners including the intelligence community and the DOE, write minimum requirements for safety frameworks informed by technical experts, standards, public input, and best practices. States can shape regulation above the federal floor.
- **Increased transparency on Automated AI R&D:** Frontier developers must disclose substantially autonomous AI R&D in accordance with prescribed measures.
- **Flexibility to evolve over time:** Ability for scope of coverage, risk domains, and requirements to evolve over time. To avoid stifling innovation, startups and academics are not generally covered.
- **Incident reporting and disclosures:** Any person may confidentially report a safety incident to CAISI, and developers must report safety incidents. Evaluators and developers also must report imminent risk of severe harm.

Truly independent evaluations with government oversight: Enable oversight of AI safety primarily by government, supported by truly independent evaluators:

- **Government teams evaluate each developer:** Government teams examine safety frameworks to ensure they comply with federal requirements.
- **Shared assurance functions with a government backstop:** Third-party verifiers are accredited by the government to conduct compliance evaluations in their accredited domains, while government teams retain ultimate authority.
- **Fee-based funds:** Developers do not pay evaluators directly, and there are limits on the amount of funding that can come from one covered developer and its affiliates.
- **True independence of evaluators:** Evaluators are assigned by CAISI based on qualifications with mandatory rotations to preserve independence. Conflict-of-interest provisions restrict revolving-door talent between developers and accredited independent evaluators.
- **Pre-deployment evaluations:** Evaluations assess framework compliance before commercial release or broad internal use, including automated AI R&D.

Enforcement with due process and transparency: There must be robust accountability to ensure public safety, while providing due process and transparency.

- **Provides narrow emergency authority with judicial review:** Where safeguards are inadequate and an AI system poses imminent risks of severe harm, the government may order mitigation measures subject to prompt judicial review.
- **Civil penalties for violations of safety frameworks,** with opportunities for remediation and judicial review, and enhanced penalties for knowing or willful violations. Includes procedures for criminal referral.
- **Clarifies liability** of developers for reasonably foreseeable injuries to third parties caused by AI systems.
- **Whistleblower protection:** People within covered entities who speak up about safety concerns are protected.