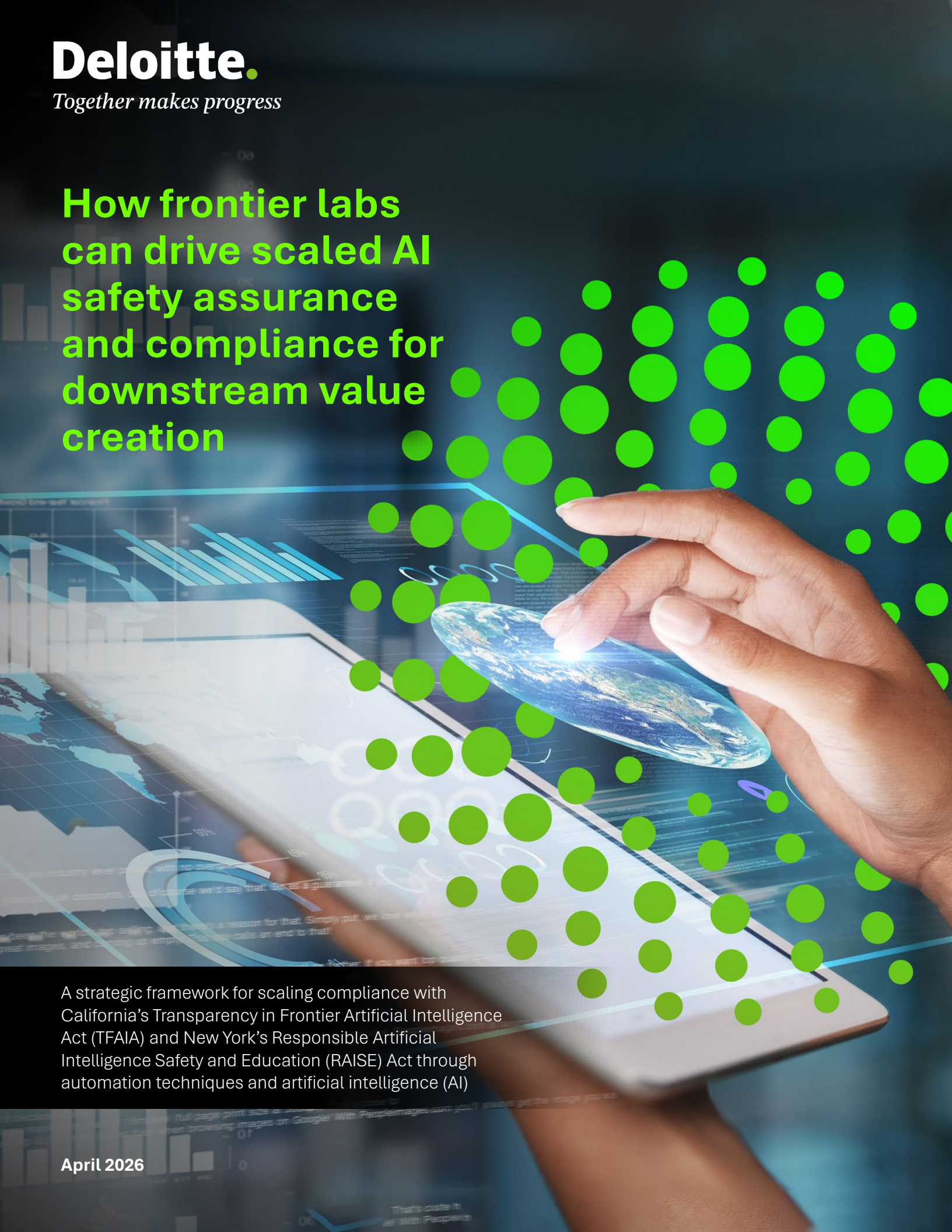
Deloitte.

Together makes progress

How frontier labs can drive scaled AI safety assurance and compliance for downstream value creation



A strategic framework for scaling compliance with California's Transparency in Frontier Artificial Intelligence Act (TFAIA) and New York's Responsible Artificial Intelligence Safety and Education (RAISE) Act through automation techniques and artificial intelligence (AI)

April 2026

Executive summary: Turning TFAIA and RAISE compliance into strategic advantage

Over the past decade, advances in deep learning—especially the rise of large, general-purpose models and Generative AI—have rapidly expanded AI capabilities. As adoption grows, compliance expectations have moved beyond model accuracy to include safety, security, traceability, and fit for purpose. In parallel, comprehensive regulatory regimes led by the European Union (EU) AI Act are establishing demanding risk-based governance and transparency requirements for high-impact AI, with California's TFAIA (also known as California Senate Bill 53) and New York's RAISE Act (also known as Assembly Bill A6453A) following a similar direction.

The enactment of California's TFAIA and New York's RAISE Act has established a significant bicoastal regulatory signal that may shape de facto expectations for the development and operation of advanced AI models across the US. This white paper describes how emerging AI rules are increasing operational complexities in a very fast-moving ecosystem and lays out an approach to automating compliance, safety and security controls, and monitoring risks across the AI development life cycle. This has an opportunity to create tremendous downstream value for frontier labs including, but not limited to, building demonstrable trust with regulators, assuring their compliance posture, and expanding their product portfolio and value to customers.

California's TFAIA vs. New York's RAISE Act

The table below provides a side-by-side comparison of California's TFAIA (signed into law September 29, 2025, and effective January 1, 2026) and New York's RAISE Act (signed into law December 19, 2025, and effective January 1, 2027) across jurisdiction-specific thresholds and definitions, reporting timelines, record-retention requirements, and penalty provisions.

	California TFAIA	New York RAISE Act
Definition of frontier developers	A person who has trained or initiated the training of model(s) with training compute exceeding 10^26 operations.[1]	A person who has trained model(s) with training compute exceeding 10^26 operations, the compute cost of which exceeds $100M, or a model produced by applying knowledge distillation to a frontier model.[2]
Definition of large frontier developers	A frontier developer that together with its affiliates collectively has annual gross revenues exceeding $500M per year.[1]	A person that has trained at least one frontier model and has spent more than $100M in computing costs in aggregate in training frontier models.[2]
Definition of risk	"Catastrophic risk" involving at least 50 people or $1B in damages from a single incident.[1]	"Critical harm" involving at least 100 people or $1B in damages.[2]
Definition of incident	"Critical safety incident" means: • Unauthorized access to, modification of, or exfiltration of, the model weights of a frontier model that results in death or bodily injury. • Harm resulting from the materialization of a catastrophic risk. • Loss of control of a frontier model causing death or bodily injury. • A frontier model that uses deceptive techniques against the frontier developer to subvert the controls or monitoring of its frontier developer outside of the context of an evaluation designed to elicit this behavior and in a manner that demonstrates materially increased catastrophic risk.[1]	"Safety incident" means an incident that occurs in such a way that it provides demonstrable evidence of an increased risk of critical harm: • A frontier model autonomously engaging in behavior other than at the request of a user. • Theft, misappropriation, malicious use, inadvertent release, unauthorized access, or escape of the model weights of a frontier model. • The critical failure of any technical or administrative controls, including controls limiting the ability to modify a frontier model. • Unauthorized use of a frontier model.[2]
Documentation mandate	Frontier AI framework (documented technical and organizational protocols to manage, assess, and mitigate catastrophic risks). Transparency report to be published before deploying a new frontier model or substantially modified version of an existing frontier model.[1]	Retention of specified tests and test results.[2] An unredacted safety and security protocol (documented technical and organizational protocols to reduce risk of critical harm) with update history for the duration of deployment plus 5 years.[2]

	California TFAIA	New York RAISE Act
Audit requirements	No annual mandate, but frontier developers must disclose a summary of their assessment results for catastrophic risks and the extent to which third-party evaluators were involved in the assessment in its transparency report.[1]	An annual independent audit, with published results.[2]
Incident reporting timelines	Reporting within 15 days to report a critical safety incident to the California Office of Emergency Services (OES) after discovery, but if it poses an imminent risk of death or serious physical injury, then 24 hours to disclose to an appropriate authority.[1]	Reporting within 72 hours (from when the large developer learns of the incident or learns enough facts to reasonably believe one occurred) to disclose each safety incident.[2]
Record retention	Unredacted retention for the model's life plus 5 years.[1]	Unredacted retention for the model's life plus 5 years.[2]
Penalty exposure	Penalties up to $1M per violation.[1]	Penalties up to $10M for a first violation and up to $30M for any subsequent violation. For violations regarding employee protections, $10,000 per employee.[2]

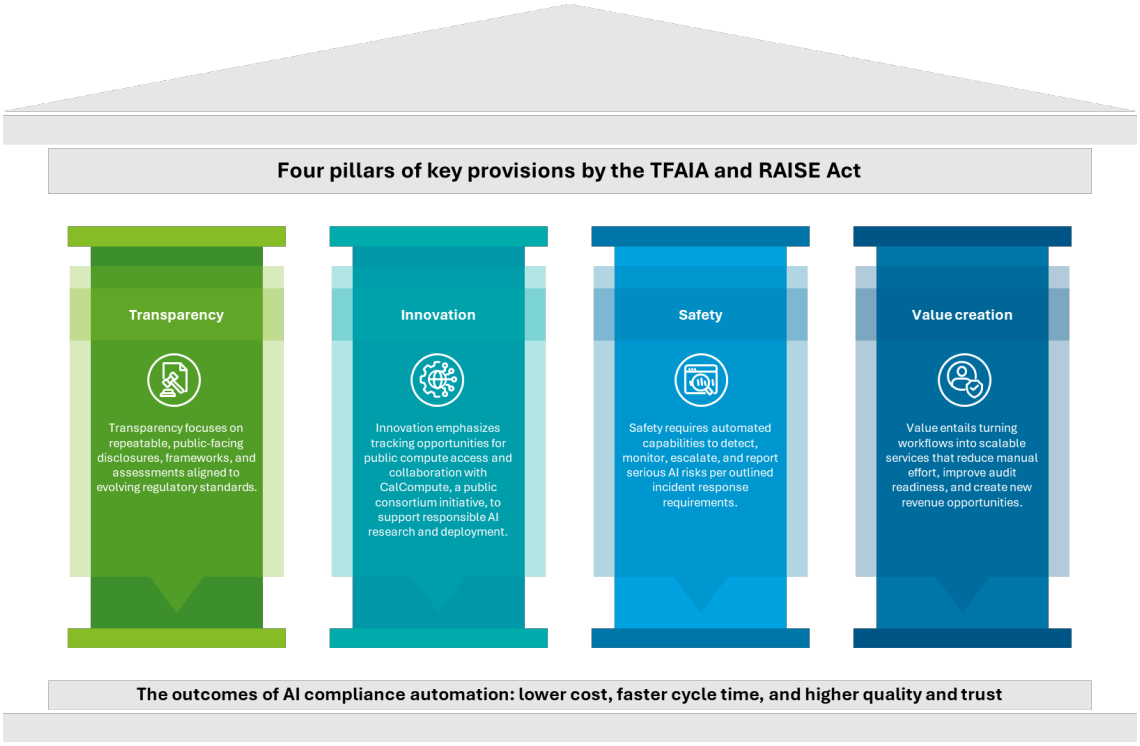
Sources:

1. California SB-53 Artificial intelligence models: large developers.
2. New York State Assembly Bill 2025-A6453A.

Rightsized automation approaches to each regulatory pillar

The TFAIA and RAISE Act are comparatively ambitious proposals and, rather than representing the mainstream of US state activity, they sit closer to the "outlier" end of the spectrum, as most states' near-term AI legislation is concentrating on pragmatic, targeted measures. Even so, the cost of inaction—enforcement actions, remediation, litigation, delayed launches, lost business opportunities, and reputational harm—typically far outweighs the investment in early, proactive compliance. Frontier-model developers that embed controls into the AI development life cycle, standardize documentation, continuously monitor performance, and maintain audit-ready evidence can reduce costs and accelerate timelines while strengthening trust. This approach transforms regulatory readiness into a repeatable competitive advantage.

In this section, we distill TFAIA and RAISE Act requirements into actionable automation strategies. We organize key provisions into four core pillars and outline rightsized automation approaches that help turn regulatory obligations into practical, implementable solutions. These same solutions can seed the development of scalable AI solutions for enterprise customers of frontier developers looking to modernize their risk and compliance functions.



Transparency

Continuous documentation and provenance

AI laws are expected to be updated over the coming years. The TFAIA and RAISE Act require frontier-model developers to continuously interpret regulatory changes and update compliance protocols, safeguards, and audit documentation in near real time. Both California's SB 53 and New York's RAISE Act establish extensive transparency requirements for developers of advanced AI models. These requirements center on the publication of safety protocols, the disclosure of risk assessments, and the reporting of safety incidents.

- **Intelligent risk assessment generation:** Automated logging of training data (compute power, data sets), output, and modifications to enable report generation. Additionally, developers can rationalize their risk assessment process and centralize these risk signals to generate risk assessment reports for each law and future laws mandating risk assessment.
- **AI transparency automation:** As protocols and models are updated and tested, agents can read and compare changes, then dynamically refresh labels, frameworks, and system cards to enable compliance with transparency obligations while building customer trust.

Innovation

Leveraging CalCompute and consortium collaboration

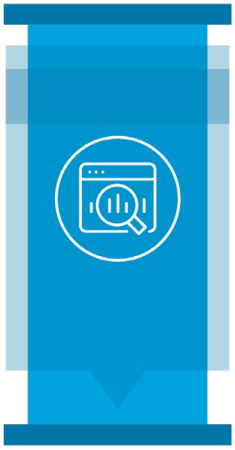
The TFAIA creates a 14-member consortium to develop a framework for CalCompute, a public compute cloud, to broaden access to AI compute and enable collaborative research. The consortium must submit the framework to the California Legislature by January 1, 2027, a key near-term milestone for frontier-model developers. When operationalized, CalCompute would let developers collaborate with universities and labs on model development and evaluation in shared public infrastructure. Below are several ways automation could support the Innovation pillar and collaboration with the CalCompute consortium:

- **Integration, orchestration, and compliance checks:** Create custom connector scripts and managed application programming interface (API) integrations to support orchestration and compliance checks with public computing clusters for safe, ethical experimentation.
- **Data sharing within the CalCompute consortium:** Develop connector scripts and API integrations to automate and facilitate secure data sharing and collaborative research within the CalCompute consortium.
- **Trend analytics and outcome assessment:** Use custom scripts to capture structured data, apply machine learning (ML) and AI for trend and outcome analysis, and leverage a large language model (LLM) to interpret qualitative innovation results and map them to ethical and sustainability goals.

Safety

Proactive incident detection and reporting

From harm to early warning	*Under the TFAIA and RAISE Act, safety incidents now include precursor events—model weight theft, autonomous behavior, control failures—not just death or injury. The regulatory bar has shifted from reactive to preventive.*



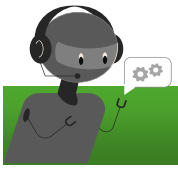
The TFAIA and RAISE Act impose incident detection and reporting requirements, creating both compliance obligations and opportunities to improve operations. To meet regulatory timelines while managing risk effectively, frontier-model developers can deploy intelligent automation across the entire incident life cycle—from detection and triage through investigation, remediation, reporting, and prevention.

Automation can reduce time-to-detection and time-to-resolution while improving consistency, accuracy, and audit readiness. Core incident response automation capabilities include:

- **Continuous monitoring and real-time alerting:** "Sentinel" agents are always-on monitoring components that continuously scan AI product signals (e.g., model performance, drift, anomalous outputs, access patterns) alongside internal and external contexts. They translate these inputs into actionable alerts for potential safety failures, compliance breaches, elevated bias risk, privacy leakage, or security anomalies. When triggers occur, "sentinel" agents can alert human compliance officers and initiate risk triage processes to reduce time-to-detection.
- **Accelerated root cause and impact analysis:** Automated agents conduct parallel investigations across human behavior (e.g., credential exposure, misrouted communications) and code (bugs, design flaws, unintended model behavior). This parallel processing helps teams identify root causes and scope impact faster than manual analysis—compressing investigation timelines while improving thoroughness.
- **Intelligent remediation planning:** Once the root cause is confirmed, "fixer" agents can recommend targeted code changes and follow-on actions to prevent recurrence. They can then hand off to human owners to implement the fix and validate the mitigation.
- **Streamlined regulatory reporting:** Use automation to alert stakeholders and draft regulator communications. Create a centralized AI governance registry that auto-logs training data summaries, compute metrics, and safety test results to support transparency and faster incident reporting. Add provenance tracking for synthetic content and model changes to ensure audit traceability and auto-populate incident reports.
- **Predictive risk intelligence:** Leverage data captured across the risk life cycle, from identification through mitigation to run predictive analytics that flag recurring patterns and emerging risks early. This enables proactive controls and faster triage, further reducing incident response timelines.

Illustrative agent performance KPIs

The following illustration shows an end-to-end incident response flow that leverages AI agents and automation to streamline processes from detection to prevention.



Agentic flow for proactive incident detection and reporting

End-to-end agentic flow

An example of an end-to-end agentic flow for proactive incident detection and reporting, from monitoring and intake, through triage and investigation, to mitigation, formal reporting, and continuous learning. Individual agents handle discrete tasks, while agentic pods group related agents into repeatable, scalable units that speed up investigation and ensure consistent evidence capture and decisioning across AI incident detection and reporting.

A. Monitor/detect/intake

1 Sentinel agent pod

Internal "sentinel" agent
Constantly scans the internal environment for risk signals against historical signals that preempted materialized risks. Infers potential risk signals based on in-flight investigations.

External "sentinel" agent
Constantly scans external environment, such as media channels, public forums/boards/ threads, etc. for multi-modal risk signals.

2 Intake agent
Consumes and parses through free-form internal reports (e.g., email, chats, tags, etc.) of potential risks and external risk reports.

B. Triage and investigate

3 Triage and planning pod

Triage agent
Assesses risk signals to determine risk validity and gauge potential threat level.

Investigation planning agent
Gathers case facts, risk factors, and time requirements to create a detailed investigation plan, and delegates to investigative agents downstream.

4 Stakeholder engagement agent
Identifies the responsible, accountable, and impacted stakeholders, and establishes a communications channel to keep everyone apprised of the situation and to gather feedback.

5 Error source investigator pod

Human behavior review agent
Reviews case facts and related touchpoints to determine potential human error or malicious action (e.g., access keys exposure, mistaken email addresses).

Code review agent
Evaluates related code for potential bugs, design issues, and unintended behavior.

6 Impact determination pod

Fan-out agent
Evaluates other systems, databases, and models exhibiting the same or similar issues related to the error/impacted data to assess "breadth" of the issue.

Lineage analysis agent
Evaluates upstream and downstream connections to error/impacted data to assess "depth" of the issue.

C. Mitigate/remediate and report

7 Commitments matching agents
Evaluates the impact(s) caused by the risk against policy, laws, and statements made by the organization to assess potential conflicts.

8 Mitigation and remediation pod

Remediate agent
Recommends remediation actions to "clear" the impact of the risk (e.g., deleting improperly collected user data), handing off to a human to perform and confirm actions have been taken.

Fixer agent
Recommends code fixes and follow-on action to "stop" the risk from occurring, handing off to a human to perform and confirm whether the fix has been implemented appropriately.

9 Incident reporting pod

Audit readiness agent
Establishes and maintains an audit-ready evidence trail for frontier-model governance, safety testing, incident reporting, and transparency obligations under the TFAIA and RAISE Act.

Reporting agent
Evaluates reporting requirements against investigative facts; where appropriate, prepares and routes regulator-ready submissions through approved channels.

D. Learnings and prevention

10 Knowledge transfer agent
Disseminates learnings from the investigation process and outcomes to relevant teams to promote continuous learning and risk prevention measures.

11 Risk review pod

Risk review agent
Proactively scans the production environment to predict risk; synthesizes risk information, notifies stakeholders to early warning signals, and recommends risk mitigation strategies.

Risk detection agent
Continuously scans cases tied to health/safety incidents, correlates data across sources, prioritizes alerts based on severity, and orchestrates remediation workflows.

Agent vs. agentic pod

Agent:
A single specialized worker with one clear responsibility.

Agentic pod:
A small coordinated team of agents that collectively completes a broader outcome, often in parallel and with internal handoffs.

Note: Agents can be helpful when judgment over unstructured inputs is needed; deterministic automation is helpful where rules are stable and auditable.

Incident response time window
(CA: 24 hours or 15 days, NY: 72 hours)

Human in the loop (HITL)

Human in the loop across the entire process from legal interpretation, severity decisions, investigation, remediation authorization, and regulator submissions

Tracking and reporting of agent performance KPIs *(e.g., detection precision, severity accuracy, reporting quality, audit traceability)*

Value creation

Automating the compliance life cycle

Both laws require developers to define evaluation procedures (thresholds) for identifying severe risks and to implement protections (mitigations) to address them, though they use different terminology for the primary harm being managed. California's TFAIA explicitly requires large frontier developers to implement a "frontier AI framework" that documents how they manage catastrophic risks. New York's RAISE Act imposes similar obligations through its requirement for a "safety and security protocol," focusing on the risk of critical harm (New York's equivalent term to California's "catastrophic risk"). Outlined below are the risk evaluation and audit-related automation strategies:

- **Risk assessment and red teaming:** Deployment of continuous performance monitoring tools that validate guardrails and automatically gather risk signals. This includes testing deceptive techniques where a model attempts to subvert monitoring controls and codifying past human tests into a testing graph repository.
- **Evidence collection and audit readiness:** New York's RAISE Act mandates an independent third-party audit. These audits can benefit from rule-based automation schedules API pulls across key data stores, while agents normalize and validate inputs. This results in audit-ready evidence generated on demand for investigations, audits, and regulatory requests—eliminating manual data gathering.

There is an opportunity for developers to productize some of these compliance automation capabilities and make them available to all their customers as a service or stock keeping unit (SKU). Organizations that invest early in intelligent governance, risk, and compliance (GRC) workflows and data-flow automation transform compliance from a cost center into a function that drives efficiency and incubates potential future automation products. Additional GRC-related automation capabilities developers can consider for themselves and their customers include:

- **Continuous regulatory intelligence and protocol alignment:** Automated agents continuously scan global laws, analyze requirements, and compare them against internal policies and controls. The system detects gaps, flags misalignments, and delivers actionable alerts. Agents can flag changes, map impacts, draft updates, and route for approval—helping to keep compliance teams ahead of changes without manual intervention.
- **Scenario planning and remediation tracking:** A model compliance checker runs "what-if" scenarios against anticipated regulatory shifts; results roll into an "action-planning" agent that sequences remediation work, assigns owners, and tracks progress across teams.
- **Intelligent control monitoring and performance dashboards:** Automated agents continuously monitor controls for operational failures, drift, or degradation. When issues arise, they trigger preapproved actions or escalate recommended actions and alert control owners. Real-time dashboards aggregate control evidence, track operating effectiveness across the organization, flag deviations for remediation, and support independent audits—giving executives and compliance teams full visibility into control health without manual reporting cycles.

Call to action:
Leading the way in TFAIA and RAISE Act compliance

Recent federal actions have increased regulatory volatility around advanced state-level AI rules. As enforcement expectations shift and preemption or litigation plays out, compliance duties can change in ways that are hard to anticipate and harder to manage manually. As a result, static, manual compliance programs won't keep pace, and compliance must become a living capability. Automation is the most scalable way to deliver it.

An automation-first pilot can operationalize the four pillars into an integrated, agent-assisted compliance posture embedded in day-to-day workflows. Agents handle repetitive, time-sensitive tasks and assemble traceable evidence packages, while human leaders continue to make context-dependent decisions, approve edge cases/escalations, set risk posture, and maintain a complete auditable trail.

Deloitte helps frontier-model developers embed risk-aligned, tech-enabled compliance controls from day one to reduce regulatory friction and accelerate innovation. We can help you get started on:

- **Proactive compliance:** Deloitte helps organizations anticipate regulatory changes and implement proactive measures that focus on continuous compliance. This reduces the risk of non-compliance and enhances operational efficiency.
- **Trust by design:** Deloitte emphasizes building trust into the design and development process, including incorporating privacy, security, and transparency features from the outset so products and services are not only compliant, but also trusted by users.
- **Integrated technical solutions:** Deloitte can help architect, implement, pilot, and scale integrated compliance solutions that leverage the latest technologies, such as automation, Generative AI, and GRC technologies. These solutions streamline compliance processes, reduce manual effort, and provide real-time insights.

Deloitte's Digital Trust and Online Protection team is composed of compliance, safety, AI, engineering, regulatory, data privacy, and forensic specialists who can help organizations formulate and implement their response with end-to-end, multidisciplinary support for each stage of the digital AI product and platform development life cycle.

Contact us

Please reach out to our team if you would like to discuss these topics in more detail.

Tanneasha Gordon
Principal
Digital Trust & Privacy Lead
Deloitte & Touche LLP
tagordon@deloitte.com

Arpan Tiwari
Managing Director
Hybrid AI & Infrastructure Lead
Deloitte Consulting LLP
arptiwari@deloitte.com

Esther Choi Ohm
Partner
Digital Trust & Online Protection Lead
Deloitte & Touche LLP
eschoi@deloitte.com

Deloitte.

This document contains general information only and Deloitte is not, by means of this document, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This document is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor.

Deloitte shall not be responsible for any loss sustained by any person who relies on this document.