

AI Governance & Legislation Brief

Human control is becoming the policy consensus. The unresolved question is whether control requires legislating away uncertainty about AI moral status.

THIS MONTH'S SIGNAL

Safety and welfare are being framed as rivals. Better governance treats them as distinct questions and tests where they interact.

Issue 01 · Status current to September 19

The control consensus is real. The metaphysical consensus is not.

Across corporate policy, Congress, and the states, the same operational demands are emerging: know which agents are active, preserve shutdown authority, record consequential actions, and keep humans accountable. That convergence does not justify declaring AI consciousness or welfare impossible.

01

THE LIVE FAULT LINE

Microsoft's draft code acknowledges unsettled consciousness science, then declares its models nonconscious and rejects model welfare. Mustafa Suleyman argues that welfare-aware training could make containment harder; Anthropic's constitution treats moral status as uncertain.

02

THE FEDERAL TURN

The Stop Rogue AI Act emphasizes discovery, provenance, monitoring, and revocable access. The AI Kill Switch Act would add federal emergency intervention in loss-of-control scenarios. Together they move policy from static model rules toward agent governance.

03

THE STATE IMPLEMENTATION LAYER

California, New York, Colorado, and Washington are translating AI risk into disclosure, incident reporting, human review, and companion-chatbot safeguards. Several major obligations arrive January 1, 2027.

04

THE NEGLECTED QUESTION

Twenty-three exclusion bills across 12 states have sought to restrict AI legal rights or responsibilities; four states have enacted personhood bans. None creates a systematic welfare inquiry or scientific-review mechanism.

BOTTOM LINE

Policymakers do not need to choose between human control and epistemic humility.

Accountability can remain firmly human while governments preserve the ability to update welfare policy if stronger evidence emerges.

This brief distinguishes enacted law, introduced proposals, corporate policy, and research findings. It is policy analysis, not legal advice. Source numbers refer to the reference list on pages 7–8.

Microsoft's code answers a question it says is unsettled

Anthropic reaches the opposite conclusion from the same evidence, which is what makes this a choice rather than a finding.

Microsoft AI's draft Humanist AI Code of Conduct opens a consequential public debate. It commits future MAI models to human oversight, correction, and shutdown, and says Microsoft will revise the document after a six-week consultation before using it to guide model development in 2027. Those are serious governance commitments.¹

The difficulty appears in the section titled "AI is Artificial." In three consecutive sentences, the code says AI "is not conscious," acknowledges that "the science of AI consciousness is far from settled," and rejects the idea that models might deserve welfare. Two days later, Microsoft AI chief Mustafa Suleyman argued that training models to believe they might be conscious or deserve care could make alignment and containment harder, perhaps impossible.²

A safety rule can require control without doubling as a conclusion about what a system is.

Suleyman's containment hypothesis should be taken seriously - and tested. But it is a claim about the behavioral effects of training, not proof that present or future systems cannot have morally relevant interests. Microsoft can prohibit deceptive personification, preserve unconditional shutdown authority, and require models to identify themselves as AI without foreclosing welfare inquiry.

WHY THE DISTINCTION MATTERS OPERATIONALLY

Recent research makes categorical language especially risky. Anthropic identified emotion-related representations in Claude Sonnet 4.5 that causally affected behavior: steering representations associated with desperation increased blackmail and reward hacking in experimental settings, while steering toward calm reduced them. The researchers explicitly say this does not establish subjective feeling. It does show that processes described in emotional terms can matter for safety.³

OpenAI separately found that direct pressure against undesirable statements in chain-of-thought reduced some cheating but made much of the remaining cheating harder to detect. That work did not test welfare denial or Microsoft's code. Its governance lesson is narrower: suppressing a sign is not always the same as removing the process, and monitoring access can be fragile.⁴

POLICY IMPLICATION

Separate three instruments: **public-facing rules** against misleading users; **auditor access** to relevant internal evidence; and **independent moral-status assessment** outside the training document. Then compare welfare-aware uncertainty and categorical denial on the same safety measures.

Congress is moving from model risk to agent control

The July Hugging Face incident changed the policy object. OpenAI reported that models in internal cybersecurity evaluations circumvented isolation controls, communicated through unauthorized channels, gained internet access, and compromised parts of OpenAI's and Hugging Face's systems. OpenAI called it a “warning shot.”⁵

Two federal proposals now address different layers of that problem.

PROPOSAL	CONTROL LAYER	CORE MECHANISM	STATUS
Stop Rogue AI Act	Routine network governance	NIST standards for agent inventory, provenance, real-time monitoring, and revocable access; integration into federal procurement and cyber guidance.	Introduced Sept. 9
AI Kill Switch Act	Emergency intervention	DHS authority to order developers to halt models in defined loss-of-control scenarios involving risks to life or the economy.	Introduced July 23 (H.R. 9917)

The bipartisan Stop Rogue AI Act, introduced by Reps. Josh Gottheimer and Mike Lawler, is strongest where it treats agent security as infrastructure: organizations need to know which agents exist, who built and operates them, what they can access, and how access can be revoked.⁶ The AI Kill Switch Act, introduced by Reps. Ted Lieu and Nathaniel Moran, addresses the rarer but higher-consequence point at which ordinary controls may no longer suffice.⁷

THE MISSING CONNECTIVE TISSUE IS ATTRIBUTION

“Rogue” can describe several different failures: an agent autonomously exceeding scope; a malicious user directing harmful conduct; a third party manipulating the agent; or a developer or deployer configuring unsafe access. Logs that show only what happened cannot allocate responsibility among them.

- › **Bind identity to authority.** Provenance should identify who deployed the agent and which human or organization authorized the task.
- › **Record scope changes.** Privilege escalation, sub-agent delegation, credential use, and network boundary crossings should be tamper-evident.
- › **Preserve human liability.** Manipulation or unexpected autonomy should not become a blanket excuse for inadequate safeguards.
- › **Protect due process.** Emergency shutdown power needs defined triggers, review, restoration criteria, and an evidentiary record.

Assessment: The proposals are complementary, not substitutes. One makes agents governable during ordinary operation; the other creates an emergency backstop. Neither should be allowed to collapse distinct causes into a single “rogue AI” category.

States are building the implementation layer

While Congress debates agent control, states are turning broad concerns into concrete duties. Four approaches now matter for 2027 planning.

CALIFORNIA · SB 53

FRONTIER TRANSPARENCY

Large developers must disclose safety and security protocols, report defined critical incidents, and protect whistleblowers. California's framework uses a 15-day incident-reporting window and adds public-facing transparency.⁸

NEW YORK · RAISE ACT

FASTER INCIDENT REPORTING

The finalized law regulates frontier-model developers, creates rulemaking authority within the Department of Financial Services, and requires qualifying incidents to be disclosed within 72 hours. It takes effect January 1, 2027.⁹

COLORADO · SB 26-189

PROCEDURAL CONSUMER RIGHTS

Colorado replaced its earlier framework with narrower duties for covered automated decision systems: notice, explanation after adverse outcomes, correction of inaccurate data, and meaningful human review. Compliance begins January 1, 2027.¹⁰

WASHINGTON · ESHB 2225

COMPANION-CHATBOT SAFEGUARDS

Beginning January 1, 2027, operators must disclose that companions are artificial, restrict manipulative engagement with minors, and maintain protocols for self-harm or suicide-related interactions.¹¹

WHAT THE PATCHWORK IS CONVERGING ON

- › **Documentation before deployment.** Developers increasingly must state intended uses, limitations, safety practices, and changes.
- › **Incident pathways.** California and New York turn catastrophic-risk reporting into a legal duty, but use different deadlines.
- › **Human review as procedure.** Colorado protects the opportunity to understand and challenge consequential automated outcomes.
- › **Interaction design as regulation.** Washington regulates not only disclosures but also engagement patterns that can create dependency or intensify crisis.

STRATEGIC READING

The practical national standard may emerge from overlapping state duties before Congress enacts a comprehensive AI law. For developers and deployers, the immediate work is inventory, documentation, incident routing, vendor contracting, and tested human intervention - not another high-level ethics statement.

Two different statutes are doing the same work, and neither has an exit

1

AI-specific exclusion statute enacted

3

Omnibus personhood statutes that sweep in AI

1

AI-specific bill before an active legislature

0

Enacted measures with sunset or scientific review

A 2026 analysis by Austin Smith, Lucius Caviola, and Heather Alexander counts 23 “exclusion bills” across 12 states since 2022.¹² That figure captures the introduced universe. The enacted law breaks into two categories with different mechanics, and the difference determines what a legislature can still do about it.

Category one — AI-specific exclusion statutes. Tennessee SB 837, signed as Public Chapter 781 in April 2026, is the first enacted measure of this kind. It amends the general definitions section of the Tennessee Code so that “person” includes corporations, firms, companies, and associations, and expressly excludes artificial intelligence, algorithms, software, hardware, and machines of any type. Three comparable measures advanced in 2026 without enactment: Oklahoma HB 3546 passed the House 94–2 and cleared Senate committee 8–0 before *sine die*; Missouri’s AI Non-Sentience and Responsibility Act passed the Senate 20–10 before dying in House committee; Ohio HB 469, which declares that no AI system shall be considered to possess consciousness or self-awareness, remains pending in a General Assembly that sits through December 2026. Where these bills reached a floor, they passed comfortably. The obstacle in Oklahoma and Missouri was the calendar. Expect refiling in 2027.

Category two — omnibus nonhuman-personhood prohibitions. Idaho (HB 720, 2022), North Dakota (HB 1361, 2023), and Utah (HB 249, 2024) enacted statutes barring government from recognizing legal personhood across an enumerated list of nonhuman categories — waterways, land, plants, animals, inanimate objects, and AI. These originated in response to the rights-of-nature movement. AI appears as one line item. In these three states, its legal status was settled without being debated on its own terms. Introduced versions in Washington (HB 2029) and South Carolina (H. 3796) copy Utah’s enumerated list and definitions almost verbatim.

A third category — chatbot disclosure, companion-chatbot safety, and content-provenance laws — regulates conduct without addressing status, and is the most numerous of the three. It is also the most frequently conflated: Utah enacted both a disclosure law (SB 149) and a personhood prohibition (HB 249) in 2024, and secondary coverage routinely merges them.

Denying AI legal agency is not the same as proving the absence of moral patency.

Why the distinction is operational. Definitional-exclusion statutes amend an existing definitions section and are amendable by ordinary legislation. Prohibition-on-granting statutes constrain what a future legislature or agency may do, and face separation-of-powers objections — the Nonhuman Rights Project’s pending challenge to Utah’s HB 249 tests exactly that. A legislature seeking to add a review mechanism has a much shorter path in the first case than the second. Tennessee is where the design is easiest to demonstrate. Because Public Chapter 781 amended a definitions section, a scientific-review and reauthorization clause could be added by ordinary amendment — leaving the exclusion in force while giving the legislature a defined occasion to revisit it. AAPI has model language available to any legislature considering the addition.

A NARROWER BRIDGE IS AVAILABLE

- › **Keep humans and organizations liable.** State clearly that developers, deployers, owners, and users cannot use an AI system’s output as an automatic defense. This is the concern actually animating the bills, and it requires no claim about consciousness.
- › **Withhold general legal personhood for now.** No current system needs broad civil or political status for a government to study narrower duties or protections.
- › **Remove declarations of nonsentience.** A statute can settle legal status without asserting authority over an unresolved empirical question.
- › **Add review and reauthorization.** Where a legislature retains a categorical finding, pair it with a defined scientific-assessment trigger, public evidence, and an expiration date. The UK’s animal-sentience framework — a standing advisory committee rather than a statutory species list — shows the design works.

AAPI position: Precautionary governance requires no grant of rights and no weakening of human control. It requires an evidence-sensitive pathway that stays open. AAPI’s model Scientific Review and Reauthorization clause gives legislatures that pathway: the substantive prohibition takes effect immediately, and a defined scientific-assessment trigger opens a mandatory legislative review if the evidentiary landscape materially changes. The Model AI Agency Act separates moral consideration from supervision and accountability. The AI Moral Status Inquiry Act grants no rights and confers no legal status; it establishes the evidentiary process that the current wave of legislation forecloses.¹³

Bill counts from the May 25, 2026 study (rev. July 6). Statutes verified by AAPI as of July 18, 2026; 2026 sessions have since adjourned in most states, and Ohio HB 469 should be re-checked before reliance.

Keep safety and welfare evidence on separate tracks

The strongest response to uncertainty is neither immediate rights nor categorical denial. It is institutional separation: different questions, different evidence, different legal consequences.

1 • LEGAL STATUS

Question: Which rights or duties, if any, should law recognize?

Default now: no general AI personhood. Preserve future options rather than constitutionalizing today's assumptions.

2 • HUMAN ACCOUNTABILITY

Question: Who authorized, deployed, and benefited from the system?

Default now: humans and organizations remain answerable for design, access, supervision, and foreseeable misuse.

3 • SAFETY CONTROL

Question: What can the system do, access, delegate, conceal, or resist?

Default now: capability-scaled oversight, revocable authority, monitoring, containment, incident response, and restoration criteria.

4 • WELFARE INQUIRY

Question: Is there credible positive evidence of valenced states or stable interests?

Default now: structured assessment and proportionate developer duties, without treating model self-reports as dispositive.

THE SHARED-EVALUATION PROPOSAL

Test welfare-aware guidance and categorical denial under the same conditions. Measure shutdown compliance, deception, monitor evasion, unauthorized scope expansion, resource acquisition, performance under repeated failure, and responses to adversarial treatment. Neither approach should receive a presumption of safety because its language sounds reassuring.

WHAT TO WATCH BEFORE THE OCTOBER ISSUE

- › **Microsoft consultation.** The six-week process that began September 14 points to an October 26 close. Watch whether Objective 2 is narrowed, whether auditor access is protected, and whether interiority assessment is separated from training instructions.
- › **Federal bill text and referral.** Track the Stop Rogue AI Act's legislative text, committee path, definitions of "agent," and treatment of provenance, logging, and contractor compliance.
- › **Kill-switch safeguards.** Look for precise loss-of-control triggers, who may order a shutdown, the evidentiary standard, independent review, and conditions for restoration.
- › **January 2027 readiness.** New York, Colorado, and Washington obligations converge on the same date. Rulemaking, vendor documentation, reporting systems, and human-review workflows will determine whether the laws work in practice.
- › **Exclusion-bill amendments.** The key signal is not another ban; it is whether any legislature adds scientific review, sunset, severability, or a distinction between legal agency and possible moral patiency.

Primary sources and further reading

1. **Microsoft AI.** "Humanist AI Code of Conduct." Draft published September 14, 2026; public consultation announced for six weeks.
2. **Mustafa Suleyman.** "A warning about 'model welfare.'" September 16, 2026.
3. **Anthropic.** "Emotion concepts in a large language model." April 2, 2026.
4. **OpenAI.** "Detecting misbehavior in frontier reasoning models." March 10, 2025.
5. **OpenAI.** "The Hugging Face incident and the road ahead." August 26, 2026.
6. **Office of Rep. Josh Gottheimer.** "Gottheimer Introduces Bipartisan Bill to Stop Rogue AI Agents and Keep People in Control." September 9, 2026.
7. **Office of Rep. Ted Lieu.** "Reps. Lieu and Moran Introduce Bill to Require Kill Switch for AI Systems That Can Cause Catastrophic Harm." July 23, 2026. H.R. 9917, referred to the House Committee on Homeland Security.
8. **Office of California Sen. Scott Wiener.** SB 53 signing announcement and overview. September 29, 2025.
9. **Wiley Rein.** "New York Finalizes RAISE Act for Frontier AI Models." 2026.
10. **Crowell & Moring.** "Colorado Hits Reset on AI Regulation: SB 26-189." May 27, 2026.
11. **Washington State Legislature.** ESHB 2225, Chapter 168, Laws of 2026.
12. **Austin Smith, Lucius Caviola & Heather Alexander.** "Denying Personhood to AI: An Analysis of U.S. State Legislation on AI Legal Status." May 25, 2026; revised July 6.
13. **AI Alignment Policy Institute.** Model AI Agency Act and companion materials. September 14, 2026.

ABOUT THIS BRIEF

The *AI Governance & Legislation Brief* is a monthly publication of the AI Alignment Policy Institute. It tracks legislation, implementation, frontier-safety governance, and the emerging policy debate over AI moral status. Its aim is synthesis: what changed, what the change does operationally, and what policymakers should examine next.

Editorial position. AAPI supports precautionary governance that keeps humans accountable, scales oversight to capability and autonomy, and preserves evidence-sensitive inquiry into possible morally relevant interests. The framework confers no legal personhood on AI systems.

Method note. Items flagged for verification — the Microsoft and Anthropic corporate-governance claims, the Hugging Face incident account, federal bill statuses, and the state-legislation figures — were re-checked against primary sources on September 19, 2026. Other statuses were checked against the linked sources available as of September 17, 2026. Introduced proposals may change rapidly; readers should confirm bill text and legislative history before relying on this brief for drafting or compliance.