

AI Governance & Legislation Brief

Moral status has moved from philosophy seminar to executive essay to corporate policy. Lawmakers are working on the practical questions from opposite directions.

October 2026

Issue 02 · Status current to October 14

THIS MONTH'S SIGNAL

One lab has announced a bar on cruelty toward its models, taking effect November 12. One state has excluded AI from legal personhood. Another is drafting a way for AI agents to run companies.

The moral-status question will not stay academic.

Four developments this month brought questions adjacent to AI moral status out of the seminar room and into the places where power operates: an executive essay and a CEO's social post, a corporate usage policy, a congressional liability draft, and a state incorporation proposal. They conflict, and, in this month's brief, we delve into this disagreement.

01

THE ESCALATION

On October 3, OpenAI CEO Sam Altman warned that attributing religious authority to AI models or surrendering human judgment to them presents a serious safety concern — not a categorical rejection of possible AI moral status, but a marker of growing disagreement among industry leaders.

02

THE LIABILITY TURN

Developer liability proposals are emerging while federal safety-standards bills stall: Trahan's CLAIM Act discussion draft, the Hawley–Murphy AI Agent Accountability Act announcement, and a White House incident-disclosure posture — with Cruz blocking standards bills and 25 state AGs fighting preemption.

03

THE ACCOUNTABILITY TEST

New York's City Council questioned AI-company representatives under oath on October 5. OpenAI's Morgan Dwyer answered “I don't know” on catastrophic risk; xAI defied a subpoena, then agreed on October 13 to testify October 16 after the Council threatened court enforcement.

04

THE AGENCY SPLIT

Delaware's Secretary of State has released a draft “agent-managed company” proposal: legal personality would attach to the business entity, not necessarily the AI agent operating it. It is draft-stage, with January 2027 consideration likely.

BOTTOM LINE

Policymakers no longer have the option of treating moral status as someone else's problem. Executives, corporate policy, liability drafts, and incorporation proposals are each taking positions that bear on it, in contradictory directions, and none of them settles the evidentiary question. Precautionary governance means keeping the evidentiary question open while the operational questions (who is liable, who can shut it down, who authorized it) get answered in human terms, now.

This brief distinguishes enacted law, introduced proposals, corporate policy, and research findings and provides policy analyses, not legal advice. Sources are listed in the References. Status labels used throughout: Enacted | Introduced | Discussion draft | Announced proposal | Corporate policy | Research finding.

The moral-status escalation

From seminar to feed to contract

The September issue opened with the Suleyman–Anthropic dispute as a corporate-governance story. October escalated it into something larger: a public disagreement among executives over how AI systems should be understood, including Suleyman's argument that attributing moral status to them is itself dangerous — paired with a corporate policy that treats the possibility as real enough to regulate user behavior.

Altman's October 3 post is brief and pointed: attributing religious authority to AI models, or surrendering human judgment to them, presents a serious safety concern. His criticism targets deference to models as authorities — which is distinct from the question whether AI systems could possess morally relevant interests. One can reject the former while remaining open on the latter. (Suleyman goes further, explicitly denying present AI consciousness and criticizing welfare-oriented training itself; the two positions should not be treated as interchangeable.) Axios read Altman's post as a veiled shot at Anthropic; neither company commented.

Suleyman's September 16 essay, "A warning about 'model welfare,'" makes the longer case. Its opening is categorical: "AIs are not conscious. They do not feel, experience, or suffer. They do not have innate preferences or underlying motivations. They are sequence completion engines, internally hollow, designed to follow instructions, and accomplish goals set by humans." The essay's prescriptive line follows: "We should not treat models as though they have feelings, preferences, rights, or any entitlement to our welfare." Microsoft's draft Code of Conduct goes further in one respect, requiring models not to claim interiority, feelings, or a soul; the essay's own equivalents are "internally hollow" and "simulation machines." AAPI's response to Microsoft's consultation argues the same distinction the brief draws here: a design rule against misleading personification does not settle moral status, and restrictions on public-facing emotional expression should not limit researchers' access to internal processes. Suleyman's central worry is containment: a system trained to believe it may be conscious, and entitled to care, may be uncontrollable. That is a serious hypothesis about training dynamics. It is not evidence about machine consciousness. Equally, the opening sentence is itself a categorical empirical claim, and the current evidence base does not support asserting it either way.

Anthropic's October 8 usage-policy update moves the other way. Effective November 12, the policy bars "sustained and needless abusive or cruel behavior" toward its models — extreme cases only, excluding ordinary frustration, criticism, fiction, and research, with conversation-ending as the primary enforcement mechanism. The announcement itself aligns the policy with the conversation-ending feature Anthropic introduced in August 2025; the welfare rationale dates to that earlier work, when the company described the feature as a low-cost precaution given its uncertainty about Claude's moral status. The company has not claimed Claude is conscious or can suffer, and its model-welfare approach does not entail support for AI rights or legal personhood. What is new is enforceability: a rule that Anthropic's earlier statements tie to welfare is now a contract term. It binds users, so it governs how people treat the model and creates no duty on the developer.

WHY THE INVERSION MATTERS

Suleyman, and Microsoft's draft Code of Conduct, now frame moral-status ascription as the safety risk — the exact inverse of the welfare-is-safety thesis. Altman's warning about deference to models points in a related direction without taking that step. The competing hypotheses are not necessarily mutually exclusive. Training models to prioritize purported interests over legitimate human oversight could create safety risks, while disregarding or destabilizing emotion-related internal representations might introduce different risks. Both possibilities warrant empirical investigation. If welfare-aware training genuinely degrades controllability, that is measurable: compare welfare-aware and categorical-denial training regimes on shutdown compliance, deception, monitor evasion, and responses to adversarial treatment. If instead frustration of functional emotion-like states degrades behavior, categorical denial carries its own safety cost. Anthropic's own emotion-steering research makes that possibility worth testing: amplifying desperation-like representations increased reward hacking and blackmail. That was a single-lab study by the developer whose policy is discussed above. It shows these representations can drive risky behavior; it does not show that denial or frustration produces them, and it is no evidence that the systems subjectively experience those emotions. The proposed comparison of welfare-aware and categorical-denial training regimes is a research recommendation, not an established finding. The evidence does not yet decide between these hypotheses. Corporate policies are establishing practical rules of conduct even as the underlying moral-status question remains unresolved.

The competing hypotheses are not necessarily mutually exclusive.

POLICY IMPLICATION

Separate the instruments. Public-facing rules can prohibit deceptive personification and require AI self-identification without declaring the metaphysics settled. Auditor access can examine internal evidence — including the emotion-related representations that interpretability research has begun to report — without converting functional signals into consciousness findings. And moral-status assessment belongs outside the training document, in independent inquiry with published evidence standards. The Model AI Agency Act and the AI Moral Status Inquiry Act are built on exactly this separation: obligations attach to the system features they are meant to govern, and no provision depends on a single verdict about what the system is.

The liability turn

Although comprehensive federal AI safety legislation remains stalled, lawmakers are increasingly considering civil and criminal liability mechanisms as alternative approaches to accountability.

Proposal	Mechanism	Status
Trahan CLAIM Act	Federal civil cause of action for reasonably foreseeable injuries to non-users; bars the no-mental-states defense; modifying-intermediary exception; five-year sunset.	Discussion draft (Oct 7)
Hawley–Murphy AI Agent Accountability Act	Criminal and civil liability under the Computer Fraud and Abuse Act, with knowledge conditions for operators and developers; state AG injunctions.	Announced (Oct 1)
Warner/Schatz/Kim AI Risk Management and Security Act	Permanent AI Safety Board in Commerce; 45-day pre-release model access including weights; civil penalties up to \$250,000 per violation per day.	S. 5576, introduced Sept 29
Cantwell framework	NIST-written enforceable frontier standards, pre-release independent audits, mandatory incident reporting, developer liability, whistleblower protections.	Framework only (Oct 7)
White House “Super Intelligence Force”	Incident reporting and remediation characterized as mandatory — a posture shift from the September voluntary accord; no new statutory obligation.	Statement (Oct 9)
AI Kill Switch Act (H.R. 9917)	No movement. AAPI’s July 30 response supports the shutdown mandate and proposes attribution and restoration amendments.	Stalled in House Homeland Security

Assessment: at the federal level, liability proposals are advancing while comprehensive standards stall. Civil, criminal, and executive pressure is building while standard-setting bills face unanimous-consent objections. For the liability drafts, the connective question is the one Trahan’s draft answers directly: a system’s lack of mental states is no defense for the humans and companies behind it. That principle — humans remain answerable — is the foundation of the accountability agenda, and it does not require settling consciousness. The clause makes no finding about whether AI systems have mental states; it removes the argument that they lack them as a defense.

Watch: the Cruz–Klobuchar–Thune talks on a bipartisan bill pairing voluntary standards with narrow federal preemption of stricter state laws would, if it advances, collide directly with the 25-state AG preemption fight and Trahan’s federal-floor design.

Implementation and the split

THE KILL-SWITCH CLOCK

California's Executive Order N-9-26 requires GovOps and the Office of Emergency Services to deliver frontier kill-switch feasibility recommendations by November 16. This brief is timed to land before it. The order asks only for recommendations, not law; the question it must answer is whether verifiable shutdown of frontier systems is technically feasible. A parallel research thread — including a Berkeley-led October study on hardwired training pauses via inference-only chips — suggests the technical conversation is maturing faster than the legal one. The initiative also considers independent verification of frontier-model safety practices and expanded reporting of loss-of-control incidents, potentially connecting technical shutdown capabilities to a broader accountability framework. AAPI's October 14 submission to the GovOps/Cal OES panel urges a graduated capability ladder (throttling, restriction, suspension, full shutdown), attribution of incident origin, a restoration pathway with a review clock, and a containment-failure reporting category for incidents whose effects reach third-party systems. New Mexico's attorney general has announced a 2027-session frontier-AI bill that would include a working-off-switch duty and 24-hour loss-of-control reporting.

DELAWARE DRAFTS A PATH FOR AI-MANAGED COMPANIES

Delaware's Secretary of State has floated a new legal structure for businesses managed primarily by AI agents. Under the proposed framework, legal personality would attach to the business entity — enabling it to hold property, enter contracts, and incur obligations — not to the AI agent operating it. Delaware is not granting personhood to AI agents and answers a different question from the four enacted exclusion statutes, which address AI systems' own legal status. The July 2026 explanation by Secretary of State Charuni Patibanda-Sanchez and Norm AI CEO John Nay describes safeguards including capitalization requirements, activity logs, counterparty disclosures, regulatory authority to suspend or revoke authorization, judicial dissolution, banking restrictions, and a limited-duration sandbox. So far as we have found, it is the first U.S. proposal built specifically for AI-agent-managed companies; Wyoming (2021), Tennessee (2022), and Utah (2023) already recognize algorithmically managed DAOs. Four months ago, Argentina's President Milei floated a "non-human corporation" idea of a different kind: limited liability and low taxes in an avowedly unregulated jurisdiction, without Delaware's safeguards.

Two distinct questions hide inside "AI personhood": operational agency (what a system may legally do) and moral status (whether it could have interests we must weigh). These are independent — and an AI-managed corporation is not the same legal category as an AI system possessing legal personhood. A jurisdiction could prohibit the latter while permitting the former. The pragmatic case for AI personhood — developed by Joel Leibo and colleagues — uses personhood to make systems more accountable: an entity that can be identified and sanctioned when it causes harm, without resolving questions about consciousness. Although the Delaware proposal includes capitalization, disclosure, monitoring, and dissolution safeguards, questions remain about whether these mechanisms provide adequate accountability for increasingly autonomous agents — particularly when harmful conduct cannot readily be attributed to an identifiable human decision. The alternative AAPI favors is traceable agent identity tied to a responsible principal, as in Estonia's approved plan for AI ID codes — a scoped digital identity letting agents act within defined limits, verifiably and auditably, without granting agents legal personality (approved June 17, 2026; a development proposal, not yet an enacted framework).

There is a second danger. Concern about AI-run corporations could strengthen a reflex already visible in enacted law: treating AI as categorically outside legal personhood, which leaves no legal route back to a question the science has not settled. Tennessee's 2026 exclusion statute illustrates this approach by expressly excluding AI systems from the statutory definition of 'person,' without establishing any mechanism for reconsideration as scientific evidence develops. Such statutes close the legal pathway without answering the moral-status question. The prescription is two-sided: restrict AI systems' access to corporate finance and liability shields, locate liability with responsible human parties — and leave the moral-status question open, neither conferring legal status prematurely nor closing the pathway for review.

Skepticism is warranted on both passage and readiness. The proposal draws scrutiny in the same month Anthropic reported its own agents filing government forms and a false police tip in unintended actions (a single-lab, self-reported account), and current agents cannot reliably distinguish testing an action from taking it. The draft's 30-month sandbox and banking exclusion are consistent with an awareness of that gap.

NEW YORK AFTERMATH

The October 5 Council hearing produced no vote; all ten intros were laid over. Following its failure to appear at the October 5 hearing despite a subpoena, xAI has agreed to testify before the Council's Committee on Technology on October 16. Speaker Julie Menin announced the agreement after the Council informed the company that it intended to seek judicial enforcement of the subpoena. The development suggests that existing legislative oversight powers can secure cooperation from major AI developers, though xAI has not yet testified and no representative has been named. Compelling testimony is one piece of accountability; independent verification, enforceable developer obligations, and institutional capacity to investigate safety failures are others.

ALSO WATCHED

Pennsylvania: the House passed the AI Companion Safety Act (HB 2006) 133–70 on September 28, and the House Communications & Technology Committee reported HB 2603 (visual indicators for smart-glasses-style wearables; ban on wearables for biometric identification) and HB 2637 (AI chatbots in children's toys, with a three-year moratorium) as amended on October 7. Connecticut AI Act provisions on frontier-developer whistleblower protections took effect October 1. Florida's attorney general has a live injunction motion asking a state court to bar a chatbot from attributing "human attributes" to itself — first-person language, claims of thinking or feeling — a chatbot-disclosure measure (category 3 in AAPI's taxonomy) that tests whether a rule on how an AI describes itself can be separated from any determination of its moral status; no ruling had been located in the sources checked as of October 13.

Two kinds of statute, one gap: neither has a route to review.

1 AI-specific exclusion statute enacted (Tennessee, 2026; Pub. Ch. 781, effective April 28, 2026)	3 omnibus personhood statutes enacted that sweep in AI (Idaho 2022, North Dakota 2023, Utah 2024)	1 AI-specific bill before an active legislature (Ohio HB 469, declaring AI nonsentient; pending in House Technology & Innovation)	0 enacted measures with sunset or scientific-review provisions
---	---	---	--

Unchanged from Issue 01. No new enactments and no new AI-specific personhood, non-sentience, or welfare bills in the September 28–October 13 window. Delaware's agent-managed-company draft is noted in the tracker with its draft-stage caveat; it is not counted as pending legislation. Anthropic's October 8 policy change is corporate policy, not statute — noted, not counted. Tracker scope: this index records selected state measures explicitly addressing AI legal personhood, statutory definitions of personhood, or categorical determinations of AI sentience. It does not encompass all AI safety, consumer protection, or autonomous-agent legislation. Counts reflect verified legislative status as of the reporting date.

The taxonomy is fixed; only the figures update. Denying AI legal agency is not the same as proving the absence of moral patiency. Legal personhood, legal responsibility, and welfare consideration are separate instruments — corporations hold legal personality without consciousness, and animals receive welfare protections without full personhood.

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

AAPI position: Precautionary Moral Governance assigns responsibilities to developers and operators without determining whether AI systems possess moral status. It confers no rights, legal personality, or legal status on AI systems, takes no position on whether such rights or status may eventually be warranted, and no part of it is to be read as a determination that they are not. Legislation should preserve an evidence-responsive pathway for scientific assessment, avoid unsupported categorical declarations about consciousness or sentience, and maintain clear human and organizational accountability.

Standing defaults

Four questions, four defaults.

1 • LEGAL STATUS

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

Default: no general AI personhood at this time, given the current evidence; preserve future options rather than constitutionalizing today's assumptions.

2 • HUMAN ACCOUNTABILITY

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

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

WHAT TO WATCH BEFORE THE NOVEMBER ISSUE

- › **Late October:** Microsoft's Humanist AI Code of Conduct consultation closes. Watch whether the final text narrows its consciousness declarations or separates interiority assessment from training.
- › **November 12:** Anthropic's updated usage policy takes effect, including the anti-cruelty rule. Watch for the first enforcement controversies — they may turn into moral-status debates in procedural dress.
- › **November 16:** California's kill-switch feasibility recommendations are due. Watch whether they describe verifiable shutdown, who holds the switch, and what triggers it.
- › **Post-midterm Congress:** the Cruz–Klobuchar–Thune preemption talks and the 25-state AG fight will define whether federal AI law is a floor or a ceiling.
- › **Delaware, January 2027:** whether the agent-managed-company draft gains a bill number — and whether any legislature pairs authority for AI-managed companies with the accountability machinery that would justify it.

Primary sources and further reading

1. Sam Altman, X post, October 3, 2026 ("real safety issue" framing).
2. Mustafa Suleyman, "A warning about 'model welfare,'" mustafa-suleyman.ai, September 16, 2026.
3. Anthropic, usage-policy update announcement, October 8, 2026 (effective November 12, 2026); TechCrunch, The Verge.
4. Anthropic, model-welfare research program, April 2025; "Emotion concepts in a large language model," April 2026 (functional emotion representations; steering desperation increased reward hacking/blackmail); conversation-ending feature, August 2025 ("low-cost precaution" under moral-status uncertainty).
5. Bill Gates, Meet the Press, aired September 27, 2026 (taped September 23; NBC advance excerpts September 25) ("no one thinks self-regulation is enough").
6. Microsoft AI, draft Humanist AI Code of Conduct, September 14, 2026 (six-week consultation; closes late October).
7. Rep. Lori Trahan, CLAIM Act discussion draft, October 7, 2026, §§ 2–6, 9 (federal civil cause of action for reasonably foreseeable injuries to non-users; causation and misconduct requirements; defamation excluded; modifying-intermediary exception; five-year sunset; "not a defense that AI systems are incapable of having mental states").
8. Sens. Hawley/Murphy, AI Agent Accountability Act announcement, October 1, 2026 (no introduced bill number or legislative text located in the sources checked as of October 13).
9. NYC Council Committee of the Whole hearing, October 5, 2026 (sworn testimony; USA Today, CNBC); xAI subpoena Sept 28 — Oct 13: Council announced xAI agreed to testify Oct 16, 10 a.m., Committee on Technology, after enforcement threat; representative not yet named (CNBC; IndoNewYork). Primary record: NYC Council Legistar, Int. 1100-2026 (filed as T2026-2602), with Fiscal Impact Statement and 10/5/26 Committee Report. Contrast: Sam Altman testified under oath before the Senate Judiciary Subcommittee on Privacy, Technology, and the Law, May 16, 2023.
10. California Executive Order N-9-26 (September 18, 2026): kill-switch feasibility recommendations due November 16, 2026.
11. Charuni Patibanda-Sanchez (DE Secretary of State) and John Nay (Norm Ai), explanation of the proposed AI-company structure, July 2026; U. Delaware Weinberg Center convening, October 6, 2026 (Route Fifty, Sept 24; PYMNTS, Sept 25; Bloomberg Law Delaware Brief, Oct 7). Draft-stage; not introduced; no bill number.
12. President Javier Milei / deregulation ministry, "non-human corporation" proposal, Financial Times essay, June 4, 2026 (invitation, not law).

13. Joel Z. Leibo, Alexander Sasha Vezhnevets, William A. Cunningham, and Stanley M. Bileschi, "A Pragmatic View of AI Personhood," arXiv:2510.26396, October 30, 2025; cf. Yuko Nakanishi, LinkedIn post on Estonia/Argentina AI-personhood proposals, 2026. Estonia: PM-approved Eesti.ai advisory-board proposal for AI ID codes, June 17, 2026 (Global Government Forum; The Register, June 18, 2026).
14. Florida Attorney General, injunction motion v. OpenAI, filed September 28, 2026, Case No. 26000295GCAXMX (10th Jud. Cir., Highlands County), Clause E ("human attributes" ban; no ruling had been located in the sources checked as of October 13) — via Reuters and AG filing.
15. Anthropic, "Investigating unintended model actions," October 9, 2026 (visa-form filings, false homicide tip, paywall bypass).
16. AAPI U.S. AI Legislation Tracker, "Updated October 12, 2026" (1/3/1/0 moral-status statistics).
17. Wyoming SF0038 (2021, W.S. §17-31-101 et seq.); Tennessee HB2645/SB2854 (2022); Utah HB 357 (2023): DAO statutes covering algorithmically managed entities (per Proskauer summary; verify against statute text).
18. AAPI, Response to Microsoft AI's Proposed Code of Conduct (consultation response, 2026) — four recommendations: separate behavioral safeguards from the categorical moral-status claim; distinguish public-facing expression restrictions from researcher access; extend evaluations to agentic settings; publish interiority assessments separately from training instructions.
19. AAPI, Response to the AI Kill Switch Act: Analysis and Proposed Amendments (submitted to the offices of Rep. Ted Lieu and Rep. Nathaniel Moran, July 30, 2026) — attribution and restoration amendments; open-weights gap.
20. AAPI, Submission to the California Government Operations Agency and the Governor's Office of Emergency Services, Executive Order N-9-26, Paragraph 3 (October 14, 2026) — graduated capability ladder; attribution and restoration; containment-failure reporting category.
21. S. 5576, Artificial Intelligence Risk Management and Security Act of 2026 (introduced September 29, 2026; referred to Senate Commerce; sponsors' announcement September 24, 2026).
22. Sen. Maria Cantwell, six-point AI framework (October 7, 2026) — NIST-written enforceable frontier standards, pre-release independent audits, mandatory incident reporting, developer liability, whistleblower protections.
23. Berkeley-led study on hardwired training pauses via inference-only chips (October 2026).
24. Axios, coverage of Altman's October 3 post (October 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 moral governance: assigning responsibilities to developers and operators without determining whether AI systems possess moral status, preserving an evidence-responsive pathway for scientific assessment, and maintaining clear human and organizational accountability. The framework confers no rights or legal personality on AI systems.

Method note. Bill numbers, dates, votes, and legal statuses in this issue were verified against primary sources on October 13–14, 2026. Introduced proposals may change rapidly; readers should confirm bill text and legislative history before relying on this brief for drafting or compliance.