

# Balancing AI Progress with Safety Guardrails

An Interview with Claude AI by Dirk Huizenga, March 4, 2026

## Balancing AI Progress with Safety Guardrails

*Dirk: What specific guardrails can be put on AI use and development that will not hinder progress but will prevent harm to society?*

This is a great question that sits at the intersection of policy, technology, and ethics. There's a growing body of thought on this, and the challenge is real — guardrails that are too loose invite misuse, while ones that are too tight can stifle beneficial innovation. Here's a framework for thinking about it:

**Transparency and disclosure requirements** are among the least controversial guardrails. Requiring that AI-generated content be labeled (especially in political ads, news, and customer service), that companies disclose training data sources, and that organizations publish model cards describing capabilities and limitations doesn't slow down development — it just ensures people know what they're interacting with. The EU AI Act moves in this direction, and most researchers support it.

**Tiered risk-based regulation** is another approach that tries to balance progress with safety. Rather than treating all AI the same, you classify systems by risk level. A recommendation algorithm for a music app doesn't need the same oversight as a system making parole decisions or diagnosing cancer. High-stakes domains like healthcare, criminal justice, finance, and critical infrastructure get stricter requirements (auditing, testing, human oversight), while lower-risk applications face lighter rules. This lets most innovation proceed freely while focusing scrutiny where harm is most likely.

**Pre-deployment testing and red-teaming** for frontier models is increasingly seen as a practical guardrail. Companies like Anthropic, OpenAI, and Google DeepMind already do this voluntarily. Making it a standard practice — especially for models above a certain capability threshold — adds a safety check without preventing the research itself. The key is that the testing happens before public release, not as a barrier to research.

**Accountability and liability frameworks** can shape behavior without prescribing specific technical approaches. If developers and deployers are legally responsible when their systems cause foreseeable harm (through negligence, inadequate testing, or ignoring known risks), they're incentivized to build responsibly without being told exactly how. This is analogous to how product liability works in other industries.

**A few guardrails that are more debated** include mandatory licensing for AI developers (risks

# Balancing AI Progress with Safety Guardrails

creating gatekeeping that favors incumbents), compute thresholds that trigger regulation (hard to calibrate and may become obsolete quickly), and broad restrictions on open-source model releases (could hamper the research ecosystem that drives progress for everyone).

**Where it gets genuinely hard** is in dual-use capabilities — the same model that helps a biologist design proteins could potentially help a bad actor design pathogens. There's no clean line, and most serious thinkers acknowledge that some tension between openness and safety is inherent and requires ongoing, adaptive governance rather than a one-time rule.

The approaches that seem most promising share a few traits: they focus on outcomes and harms rather than prescribing specific technologies, they scale with actual risk, they preserve the ability to do open research and iterate, and they include mechanisms for updating rules as the technology evolves. Rigid, technology-specific regulation tends to age poorly, while principle-based frameworks tend to be more durable.

Would you like me to dig deeper into any particular dimension of this — say, the policy landscape in a specific region, the technical mechanisms for enforcement, or how this applies to a particular domain?

***Dirk: Expand on the technical mechanisms for enforcement.***

This is where theory meets practice, and it's one of the harder parts of AI governance — because you need mechanisms that are actually verifiable, not just promises on paper.

**Auditing and evaluation infrastructure** is probably the most mature technical enforcement mechanism. Independent third-party audits can test models for bias, safety vulnerabilities, and compliance with stated policies before and after deployment. This looks like structured benchmark suites (testing for toxicity, fairness across demographics, susceptibility to jailbreaks), adversarial red-teaming by external teams, and ongoing monitoring of deployed systems. The technical challenge is that evaluations are only as good as the benchmarks, and sophisticated harms can slip through standardized tests. There's active research into making evals more robust, but it remains an arms race.

**Watermarking and provenance tracking** is a promising area for enforcing transparency requirements. The idea is to embed detectable signals into AI-generated content — text, images, audio, video — so that it can be identified downstream. Some approaches work at the model level (statistical patterns in token sampling), while others work at the infrastructure level (cryptographic metadata attached to outputs). The technical limitation is robustness: many watermarking schemes can be defeated by paraphrasing, cropping, or re-encoding. C2PA (Coalition for Content Provenance and Authenticity) is an industry effort to build provenance standards into the content pipeline itself, which is more durable than watermarking alone but requires broad adoption to be effective.

**Compute governance and hardware-level controls** are a more upstream approach. Since

# Balancing AI Progress with Safety Guardrails

training frontier models requires massive compute clusters, monitoring and regulating access to that compute provides a natural choke point. This could involve know-your-customer rules for cloud GPU providers, reporting thresholds when training runs exceed a certain scale (measured in FLOPs), or even hardware-level mechanisms. Some researchers have proposed secure hardware modules that could cryptographically verify what workloads are being run — essentially a “black box” for training runs. This is technically feasible but raises significant questions about surveillance and who controls the verification infrastructure.

**Structured access and API-level controls** are already widely practiced. Rather than releasing model weights openly, developers can serve models through APIs that enforce usage policies — rate limiting, content filtering, logging, and terms-of-service enforcement. This gives developers a technical surface for preventing misuse (blocking certain query types, flagging anomalous usage patterns) without restricting legitimate users. The tradeoff is that this only works for closed models; once weights are released openly, API-level controls are irrelevant. This fuels the ongoing open-source versus closed-source debate.

**Interpretability and explainability tools** represent a longer-term enforcement mechanism. If regulators require that high-stakes AI decisions be explainable, then there needs to be technical machinery to actually produce those explanations. Current approaches range from attention visualization and feature attribution (relatively shallow) to mechanistic interpretability research that tries to reverse-engineer what’s happening inside neural networks at a circuit level. Anthropic invests heavily in this area. The honest assessment is that interpretability for large models is still in its early stages — we can explain simple behaviors but not yet reliably audit a frontier model’s reasoning in the way you’d audit a financial ledger. Progress here would be transformative for enforcement.

**Differential privacy and data governance tooling** enforce rules about training data. Techniques like differential privacy allow models to be trained on sensitive data while providing mathematical guarantees that individual data points can’t be extracted. Federated learning allows training across distributed data without centralizing it. These are technically sound but come with performance costs — models trained with strict differential privacy tend to be less capable — so there’s an active tradeoff between privacy guarantees and model quality.

**Monitoring and incident response systems** are the post-deployment layer. This includes automated systems that flag unusual usage patterns (a sudden spike in requests for dangerous chemistry, for instance), human review pipelines, and structured incident reporting frameworks similar to what exists in cybersecurity (CVE-like databases for AI harms). The technical challenge is scale: billions of interactions happen daily, so monitoring has to be largely automated, which means it inherits all the limitations of the classifiers doing the monitoring.

The honest summary is that no single mechanism is sufficient. The most credible enforcement strategies layer multiple approaches — compute-level visibility, pre-deployment testing, API-level controls, watermarking, post-deployment monitoring — so that weaknesses in one layer are

# Balancing AI Progress with Safety Guardrails

compensated by others. And all of them require sustained investment, because the technology they're governing is itself a moving target.

Want me to go deeper on any of these — perhaps the compute governance angle, or how interpretability research could change the landscape?

## *Dirk: What is the policy landscape in the United States of America?*

The U.S. AI policy landscape right now is genuinely messy — probably the most dynamic and contested it's been. Here's where things stand:

**No comprehensive federal AI law exists.** The U.S. has no equivalent to the EU AI Act. Instead, it relies on a fragmented patchwork of state laws, agency guidance, and voluntary standards.<sup>1</sup> Congress has debated various proposals but hasn't passed anything binding on the private sector.

**The Trump administration has taken a deregulatory, innovation-first stance.** In January 2025, Executive Order 14179 revoked portions of the Biden-era AI executive order that had emphasized safety testing and reporting requirements.<sup>2</sup> The framing shifted explicitly toward maintaining U.S. competitiveness and reducing barriers to AI development.

**The big flashpoint is federal preemption of state laws.** In December 2025, President Trump signed an executive order proposing to establish a uniform federal AI policy framework that would preempt state AI laws deemed inconsistent with that policy. The order specifically criticized Colorado's AI Act and created several enforcement mechanisms: the Attorney General was directed to establish an AI Litigation Task Force to challenge state AI laws on grounds including unconstitutional regulation of interstate commerce and federal preemption.<sup>2</sup> The order also directed the Secretary of Commerce to publish an evaluation by March 2026 identifying burdensome state AI laws that conflict with federal policy.<sup>2</sup>

However, and this is critical: the executive order itself cannot overturn existing state law — that can only be done by an act of Congress or the courts. Until legal challenges are resolved, state laws remain enforceable.<sup>3</sup> The Senate voted to remove a House provision that would have barred states from enforcing AI-specific regulations for ten years<sup>3</sup>, which shows the limits of the preemption push.

**States have been filling the vacuum aggressively.** Thirty-eight states passed AI-related legislation in 2025<sup>4</sup>, covering topics from deepfakes in elections to AI in healthcare. The most significant state laws include:

Colorado's AI Act, which requires deployers of high-risk AI systems to use reasonable care to avoid algorithmic discrimination, mandates impact assessments, transparency disclosures, and documentation of AI decision-making.<sup>1</sup> It's currently set to take effect mid-2026 but is expected to face both legislative amendments and federal legal challenges.

California has multiple laws taking effect, including new regulations under the California

# Balancing AI Progress with Safety Guardrails

Consumer Privacy Act requiring businesses using automated decision-making technology for significant consumer decisions to provide pre-use notice, opt-out rights, and access to information about their AI use, with compliance required by January 2027.<sup>5</sup>

Texas passed its own Responsible AI Governance Act, and states like Illinois, New York City, Maryland, and New Jersey have enacted targeted regulations restricting AI use in hiring decisions.<sup>1</sup>

**The result is a standoff.** The federal government wants a light-touch, unified national standard. States are pressing ahead with enforceable rules covering discrimination, transparency, healthcare, and elections. Legal analysts note that for the time being, the executive order likely adds more — not less — complication to the current landscape<sup>6</sup>, because the boundaries of preemption are undefined and will likely be litigated for years.

**The practical advice from virtually every law firm** is the same: don't wait for preemption challenges to resolve — identify the most stringent applicable state requirements and build compliance programs around those<sup>3</sup>, since building down is easier than scrambling to build up.

So the honest picture is that U.S. AI policy is in a transitional period defined more by tension between levels of government than by any coherent national strategy. Whether that resolves through federal legislation, court decisions on preemption, or continued state-led patchwork remains genuinely uncertain.

Would you like me to dig into any particular aspect — say, how this compares to the EU approach, what the Colorado AI Act specifically requires, or how the healthcare AI regulatory picture is shaping up?

## ***Dirk: Give a summary of the EU AI Act***

The EU AI Act is the most comprehensive AI regulation anywhere in the world, and it stands in sharp contrast to the U.S. approach. Here's how it works:

**The core framework is risk-based.** The Act assigns AI applications to risk categories. Applications that create unacceptable risk are banned outright. High-risk applications are subject to specific legal requirements. And applications not explicitly banned or listed as high-risk are largely left unregulated.<sup>7</sup> This tiered structure is the backbone of the entire law — the higher the perceived risk, the stricter the obligations.

**Banned outright (unacceptable risk):** Prohibited practices include deploying subliminal, manipulative, or deceptive techniques to distort behavior and impair informed decision-making, causing significant harm.<sup>7</sup> Government-run social scoring systems are the most commonly cited example. Real-time biometric identification in public spaces is also heavily restricted, with narrow exceptions for law enforcement that require judicial authorization and fundamental rights impact assessments.

# Balancing AI Progress with Safety Guardrails

**High-risk systems get the heaviest regulation.** These include AI used in critical areas like hiring and employment, credit scoring, healthcare, law enforcement, education, and critical infrastructure. Providers of high-risk systems face substantial obligations: risk management programs, data governance requirements, technical documentation, transparency measures, human oversight mechanisms, and conformity assessments before deployment. Developers of high-risk AI systems must conduct a self-conformity assessment, and high-risk AI systems must be registered in an EU database.<sup>8</sup>

**General-purpose AI models (like large language models) have their own rules.** All GPAI model providers must provide technical documentation, instructions for use, comply with the Copyright Directive, and publish a summary about the content used for training.<sup>7</sup> Models deemed to present systemic risk face additional requirements including model evaluations, adversarial testing, incident tracking and reporting, and cybersecurity protections. Open-source models get lighter treatment unless they present systemic risk.

**Transparency obligations apply broadly.** Providers must ensure that AI systems intended to directly interact with individuals are designed so that those individuals are informed they are engaging with an AI system.<sup>9</sup> Synthetic content like deepfakes must be labeled, and emotion recognition or biometric categorization systems require user notification.

**The penalties are significant.** Fines can reach up to 35 million euros or 7% of global annual turnover for violations of prohibited practices, up to 15 million euros or 3% for other requirements, and up to 7.5 million euros or 1% for providing incorrect or misleading information.<sup>8</sup> These apply to both EU and non-EU companies offering AI systems in the EU.

**The timeline is phased.** The Act entered into force on August 1, 2024, with prohibited practices and AI literacy obligations applying from February 2025, governance rules and GPAI model obligations from August 2025, and the rules for high-risk systems becoming generally applicable by August 2026<sup>10</sup>, with some categories extended to August 2027. The Commission's "Digital Omnibus" proposal from November 2025 may push some high-risk deadlines further, linking them to the availability of harmonized technical standards rather than fixed dates.

**Extraterritorial reach** is a key feature, similar to GDPR. The Act applies to users located in the EU, and to third-country users where the AI system's output is used in the EU.<sup>7</sup> This means U.S. and other non-EU companies serving EU markets must comply.

**What's still evolving:** Each member state must establish at least one AI regulatory sandbox by August 2026<sup>7</sup>, creating controlled environments for testing new AI applications. The Commission is also finalizing codes of practice for GPAI transparency, and individual member states are passing their own implementing legislation (Italy was the first).

The big-picture comparison with the U.S. is stark: the EU has chosen a binding, comprehensive, centralized regulatory framework that prioritizes fundamental rights and safety, while the U.S.

# Balancing AI Progress with Safety Guardrails

has no federal equivalent and is actively debating whether states should even be allowed to regulate. Whether the EU approach becomes a global standard (the “Brussels Effect,” as happened with GDPR) or proves too burdensome for innovation is one of the defining policy questions in AI right now.

---

1. [Artificial Intelligence Regulations: State and Federal AI Laws 2026](#) 
2. [New State AI Laws are Effective on January 1, 2026, But a New Executive Order Signals Disruption](#) 
3. [U.S. Artificial Intelligence Law Update: Navigating the Evolving State and Federal Regulatory Landscape](#) 
4. [New laws in 2026 target AI and deepfakes, paid leave and rising Obamacare premiums](#) 
5. [2026 Year in Preview: AI Regulatory Developments for Companies to Watch Out For](#) 
6. [What to Watch as White House Moves to Federalize AI Regulation](#) 
7. [The EU Artificial Intelligence Act](#) 
8. [Freshfields: Artificial Intelligence Act](#) 
9. [SIG: A comprehensive EU AI Act Summary \[Jan 2026 update\]](#) 
10. [European Commission: AI Act](#) 