AIWS TRUST OBSERVATORY · SPECIAL REPORT

AIWS Trust Observatory Special Report

The State of AI Transparency at the Frontier

BGF Conference 2026 · Panel I

Boston Global Forum · November 10, 2026

AI power is growing while what society can see of it is shrinking. Stanford's average transparency score fell from 58/100 in 2024 to 40/100 in 2025. Both the indicators and the group of companies assessed changed between editions. Public disclosure remains uneven, combining voluntary commitments with emerging legal requirements, and much of what the law requires is addressed to regulators rather than the public.

This report draws on public assessments; the Observatory's own findings begin with the first Quarterly State of Frontier AI at the end of January 2027.

- **Disclosure is falling, not rising.** Several leading developers released frontier models without timely technical reports.
- **The core facts stay dark.** Training data, compute, real-world use, societal impact and environmental cost remain largely undisclosed across the industry.
- **Promises outrun proof.** The White House Accord on Super Intelligence of September 29, 2026 provides for an independent auditor reporting to a board committee, but does not require its findings to be made public, and independent graders find that safety rhetoric outpaces conduct.

The AIWS answer is **verifiable transparency**: transparency measured by what society can check, not by what companies choose to say.

What the Observatory sees

Major gaps persist in what leading AI developers disclose about their systems and safeguards.

Disclosure is declining. The 2025 Foundation Model Transparency Index (Stanford, December 2025) scored 13 companies; its indicators and company list changed from the 2024 edition. IBM led with 95 of 100. Meta fell from 60 to 31 after releasing no technical report for Llama 4; Mistral fell from 55 to 18; xAI scored 14. OpenAI, second in 2023, ranked second-to-last. Google delayed the model card and technical report for Gemini 2.5.

Information sharing is uneven across the frontier. The AI Safety Index, Summer 2026 (Future of Life Institute, evidence to June 3, 2026) graded nine developers:

COMPANY	OVERALL GRADE	INFORMATION SHARING
Anthropic	C+	B+
OpenAI	C	B-
Google DeepMind	C	B-
Meta	D+	D+
Z.ai	D-	D
Alibaba Cloud	D-	D
xAI	F	D
DeepSeek	F	D-
Mistral	F	D-

No company scored above C- on existential safety, and the panel found that at Google DeepMind, OpenAI and xAI, safety rhetoric outpaces revealed behavior.

Emergency plans are not visible. An August 2026 assessment by Guidelight AI Standards of Anthropic, Google, OpenAI, Meta and xAI found few published containment plans for a model that evades control. OpenAI scored highest, at 3 of 5.

The rules are new, and much of what they require goes to regulators.

LEGAL OR GOVERNANCE INSTRUMENT	PUBLISHED FOR THE PUBLIC	REPORTED TO THE REGULATOR
California SB 53, signed September 29, 2025	Annual frontier AI framework; transparency report before each frontier model is deployed	Critical safety incidents to the Office of Emergency Services within 15 days, or 24 hours if life is at risk
New York RAISE Act, signed December 19, 2025	Information on safety protocols	Incidents to the State within 72 hours
EU AI Act, Article 53 and Article 55 (general-purpose AI models)	A summary of the content used for training, on the AI Office template	Technical documentation on request; for models with systemic risk, serious incidents reported to the AI Office and held under confidentiality rules
EU General-Purpose AI Code of Practice, published July 10, 2025	A voluntary tool for showing compliance with the AI Act, adding no public duty of its own	Signatory documentation to the AI Office; Meta is not a signatory, and xAI signed only the safety and security chapter

SB 53 has applied since January 1, 2026; the EU AI Office gained enforcement powers on August 2, 2026; the RAISE Act takes effect January 1, 2027 (METR).

The newest promise carries no disclosure duty. The White House Accord on Super Intelligence, signed on September 29, 2026 by Google, Anthropic, Meta, OpenAI, xAI and Nvidia, asks each company to partner with an independent external auditor or evaluator and to designate an independent board committee that receives its reports. The reporting runs inside the company. The Accord requires no public release of the findings, only that the steps give the public "confidence": a clear case of the Disclosure Gap set out below.

Where transparency breaks

Five gaps account for most of what society cannot see. Each enters the Observatory's Blind Spot Map as a finding, not an accusation, for example: "No public containment plan identified as of November 10, 2026."

1. **Release outruns disclosure.** Frontier models reach the public before their technical reports, or without them, as with Gemini 2.5 and Llama 4.
2. **The core facts stay dark.** Training data, training compute, actual use and societal impact are "systemically opaque" across the industry. Ten of the 13 companies in the 2025 Index disclosed nothing on energy, emissions or water.
3. **Told to regulators, not to the public.** Risk summaries and incident reports flow to California's Office of Emergency Services and the EU AI Office. Public summaries are required only in part, so society learns least about the most serious events.
4. **The emergency stays unseen.** Few companies publish what they would do if a model evaded control, and the use of models inside companies before release is reported, under SB 53, only to the regulator.
5. **Open is not transparent.** Releasing model weights does not disclose how a model was built or how it is used; open-weight developers Meta, DeepSeek and Alibaba scored low on transparency.

Across all five, the pattern is the same: **disclosure is chosen by the discloser.** The public often cannot tell which claims have been independently verified, what the evaluation covered, or what remains unknown.

Five propositions on transparency in the Age of AI

The evidence above calls for more than better disclosure. It calls for a different understanding of what transparency is for.

1. **The observability duty.** As AI systems gain capability and authority, their developers must provide stronger evidence for public scrutiny. *The duty to be observable grows with power.* A standard fit for a small model cannot govern a system given authority to act in the world.
2. **From transparency of models to transparency of power.** Today's indices measure documents about models: data, compute, evaluations. The larger question is who decided to deploy a system, what authority it was given, who can stop it, and who must answer when harm occurs. Transparency that stops at the model leaves power unseen.

3. **The one-way mirror.** AI systems know more and more about every person, while people know almost nothing about the systems. A civilization of Human–AI Co-Creation cannot rest on a one-way mirror. *The mirror must become a window.*
4. **Disclose what you do not understand.** Even those who build frontier AI cannot fully see inside it. Honest transparency therefore includes what the builder has not understood, not verified, or cannot yet measure. An admitted unknown is evidence; a hidden one is a risk.
5. **Whose side is my assistant on?** For each person, the question that matters most is whether the AI assistant they rely on serves their interests or someone else's. Transparency of loyalty, what an assistant is optimizing for and on whose behalf, is the transparency every citizen can use.

The AIWS solution: verifiable transparency

The remedy is not more statements from companies. It is a public standard of what must be disclosed, proof that can be checked, and an independent observer that records the gap between the two.

1. **The proposed Transparency Floor.** A minimum public disclosure for every frontier release, proposed for adoption in the AIWS Trust Standards: a technical report on the day of release; a summary of training data sources and compute; dangerous-capability evaluation results; energy and water use; whether a containment plan exists and who can trigger it; whether the model was used inside the company before release; and what the developer does not yet understand about the system.
2. **Proof without exposure.** Companies can prove that a control works without revealing trade secrets or personal data, through privacy-preserving verification, cryptographic attestation of evaluations, and named independent evaluators who publish their findings. Attestation confirms that a test result was recorded; the highest Evidence Grade also requires naming who evaluated independently, under what conditions, and with what limits.
3. **The Disclosure Gap.** For each company, the Observatory sets its legal reporting obligations beside the public documents that can be checked. It assesses confidential filings only when granted access, and asks that each one yield a public summary within a fixed time.
4. **Promises against conduct.** The Commitment Register records every transparency promise — from company frameworks, the EU Code of Practice and the White House Accord — and checks it against what is actually published. *The Accord at One Year* appears on September 29, 2027.
5. **Answers in public.** From December 2027, at the Festival of the Light Tree of Humanity on Boston Common, AI companies are invited to a Night of Answers: each answers in public the questions the Blind Spot Map has left open.
6. **A constitutional duty.** The Constitution for Humanity in the Age of Artificial Intelligence, to be proclaimed at Harvard on April 27, 2027, carries the principle of Beacon Papers No. 13: *consequential power carries a duty of observability.*

The first Quarterly State of Frontier AI, at the end of January 2027, opens with a transparency chapter that applies the proposed Transparency Floor to every company it observes.

Questions for Panel I

1. **The floor.** Which items belong in the Transparency Floor, and which would close the largest blind spot?
2. **Proof without exposure.** How can a company prove a safety control works, through privacy-preserving verification, without revealing secrets or user data?
3. **The public's share.** When information goes to a regulator, what should the public be told, and how soon?
4. **When silence becomes a signal.** At what point should the absence of disclosure itself be recorded as a concern?
5. **Transparency of power.** How can society see who authorized a deployment, what authority the system holds, and who can stop it?
6. **Transparency of loyalty.** How can a person know whose interests their AI assistant serves?

Sources

- Transparency in AI Is on the Decline, Stanford HAI, on the 2025 Foundation Model Transparency Index, December 2025
- AI Safety Index — Summer 2026, Future of Life Institute, July 2026
- Frontier AI labs still won't say how they'd contain a rogue model, TechCrunch, August 22, 2026
- White House releases Accord between AI executives: here's what it says, Forbes Australia, on the text of the White House Accord on Super Intelligence, September 2026
- California's SB 53: The First Frontier AI Law, Explained, Future of Privacy Forum
- Governor Hochul signs nation-leading legislation to require AI frameworks for frontier models, Office of the Governor of New York, December 19, 2025
- The General-Purpose AI Code of Practice, European Commission
- Frontier AI safety regulations: A reference for lab staff, METR, January 29, 2026
- The Founding Framework of the AIWS Trust Observatory, Boston Global Forum, September 27, 2026
- Beacon Papers No. 13 · AI Power Must Not Operate Unseen, Boston Global Forum

Love · Creativity · Nobility · Wisdom