

## AI Companies Are Quietly Watering Down Their Safety Commitments

Two years ago, leading AI companies made a bold promise: if our models reach dangerous capability levels, we'll implement safety measures before releasing them, and if we can't, we won't release them at all. These 'Responsible Scaling Policies' and similar frameworks were supposed to be the industry's answer to growing concerns about AI risk.

That promise is now being hollowed out. Across all three leading frontier AI labs, the pattern is the same: safety commitments are being weakened, and competitive pressure is the stated justification.

**Anthropic dropped its pause commitment.** In February 2026, Anthropic released RSP v3.0, its most significant policy overhaul to date. The company [scrapped its commitment](#) to pause development if safety measures couldn't keep pace with capabilities. Holden Karnofsky, who led the revision, [explained the rationale](#): the old commitments created perverse incentives to downplay model capabilities, and unilateral commitments don't make sense when competitors aren't matching them. As [GovAI's analysis noted](#), other companies may now use this as an excuse to weaken their own commitments without adopting Anthropic's compensating transparency measures.

**Google is exploiting an inference-compute loophole.** When Google released Gemini DeepThink V2, [Zvi Mowshowitz pointed out](#) that it's essentially Gemini 3.1 Pro running with dramatically more inference compute. The safety evaluations were done on the base model at standard compute, but the deployed configuration is far more capable. Google initially claimed no additional safety evaluation was required. This matters because safety frameworks are built around capability thresholds: if you test at low compute but deploy at high compute, you might be shipping a model that crosses dangerous capability lines without ever having checked. Mowshowitz called the absence of safety documentation at launch 'unacceptable.'

**OpenAI released its most capable system without safety evals.** GPT-5.4 Pro, [likely the best model in the world](#) for many catastrophic risk-relevant tasks including biological research and cyberoffense, was released on March 5 without a system card. The system card published alongside the release covers only GPT-5.4 Thinking, the underlying model. OpenAI's position is that Pro is just scaffolding on top of Thinking, so separate evals aren't needed. But as [Mahajan and Yeedrag documented](#), this happened before with GPT-5.2 Pro, which was likely SOTA on dual-use biology tasks for two months with no public safety evaluation. When your 'scaffolding' dramatically boosts performance on the exact capabilities that safety frameworks are designed to catch, testing only the base model is another way of evaluating at low capability while deploying at high capability.

The common thread across all three companies is competitive pressure. Each points to what the others are doing (or not doing) to justify lowering its own bar. A [recent GovAI policy brief](#) documented this formally: at least six frontier AI companies now have provisions in their safety frameworks allowing them to lower mitigations if a competitor has released a model with similar capabilities but weaker safeguards. As the authors warn, this creates a 'race to the bottom' dynamic where a single defector from safety norms can trigger a rapid downward shift across

the industry, or a 'boiling frog' effect where individually small increases in risk accumulate unnoticed.

What is needed is regulation that makes the floor binding. Voluntary commitments established useful norms, but they were always going to erode once competitive stakes got high enough. SB53, the EU AI Act, and similar legislation need to impose substantive requirements before the last of these commitments quietly disappears.