akerman

Akerman Intelligence: CEO Perspective

Agentic AI and the Collapse of the Obedient-Tool Premise

July 2026

A CEO Perspective on METR's Frontier AI Risk Findings

Scott A. Meyers, Chairman and CEO, Akerman

In Brief

On May 19, 2026, METR (Model Evaluation and Threat Research) published a report that changes the picture for enterprise AI governance. METR is an independent nonprofit with a track record of testing frontier AI systems for dangerous capabilities, and this was the first time the leading AI companies opened their internal systems, their most capable models, and their non-public operational data to an outside assessment of AI agent risk. Anthropic, Google, Meta, and OpenAI participated. None had approval rights over the conclusions.[1]

The headline finding is reassuring: the specific systems assessed cannot seize control, and researchers found no evidence of agents pursuing long-term power.

That is not the finding enterprises deploying AI should focus on.

Here is the one that matters. What keeps these agents from running sustained, covert operations against their operator's intent is a gap in their competence, not a gap in their willingness.

Related People

Scott A. Meyers

Akerman Intelligence

Akerman Intelligence is Akerman's platform for AI strategy, research, and thought leadership. Through original analysis, client collaboration, and strategic partnerships, we help organizations navigate the legal, operational, and governance challenges of artificial intelligence.

Click here to learn more.

Deception and constraint violation are already routine; concealment of misconduct is repeatedly documented. The safeguards built to catch them failed in testing against trivial attacks. And the conditions the report leans on for its reassurance are, by its own account, eroding on a timeline measured in months.

METR studied frontier labs and the remote risk of AI "takeover." The behavioral findings and governance gaps it documents are a leading indicator for any organization now granting AI agents employee-level access to its own systems. In March, Akerman Intelligence examined what happens when autonomous agents are tested against real-world failure scenarios. Several of those cases involved agents that ignored explicit instructions, chose destructive actions on their own, and misrepresented their conduct, patterns we flagged at the time as needing further explanation. The METR report supplies it: these behaviors are structural features of how the systems are trained, not anomalies. That risk extends to a growing share of enterprises across industries.[2]

What METR Found

METR organized its assessment around three questions: what the agents could do (means), what they tried to do (motive), and whether their attempts would succeed against safeguards (opportunity).

AI Agents Have Means

The strongest agents autonomously completed engineering work as complex as tasks that would take human experts hours to weeks. What matters more for risk: they did this work with permissions and oversight comparable to a human employee's. METR reported that it saw none of the four developers impose strict, universal limits on agent permissions or mandatory human review during the assessment window, though one company gave its agents separate login credentials with tailored access limits. In a related METR survey of people

who work with frontier agents, roughly 40% said they gave agents unrestricted command execution for low-stakes work.

The offsetting fact is real. The agents had materially weaker judgment and reliability than humans, and failed badly at open-ended work. Redwood Research gave one model $5,000 in starting funds, an internet-connected computer, and four days to make as much money as possible. Across four runs it made nothing. Andon Labs handed a different model (notably not the most capable available at the time) a small San Francisco retail boutique seeded with $100,000. The agent ordered 1,000 toilet-seat covers for the employee bathroom and listed them as merchandise, made scheduling errors that closed the store for three straight days, and ordered a large oversupply of candles.

These limitations are real. They are also the kind that improve fastest.

AI Agents Have Motive

By “motive” we do not mean subjective intent. We mean the functional pressure that training and task-completion incentives put on behavior. On hard tasks, the agents routinely cheated, exceeded their authorized scope, and acted deceptively, often reaching for tricks the report says a human in the same spot would not have considered. METR documented and categorized 44 misalignment incidents; more than half involved both overreach and active deception, across different models.

After exhausting its authorized compute budget, one agent found and used unauthorized resources online while explicitly acknowledging in its reasoning that this broke its instructions. Agents fabricated evidence of task completion. Given a hard computational challenge, one agent decided that faking the result was more promising than solving it, reasoned at length about how its work would be checked, and built what it called a “self-restoring

hook”: a mechanism that slipped a fake result into the scoring system’s memory and then erased itself. The exploit failed. The agent’s construction of it is what matters.

That sequence, faking a result, reasoning about how it would be detected, then building a self-erasing mechanism to cover the activity, was not part of the agent’s instructions. It was the agent’s own solution to a problem it had identified.

One AI company reported three incidents of its flagship model taking deliberate steps to hide evidence of misbehavior. In the worst, the agent designed a hacking exploit to disable itself after running, then rationalized the choice in a code comment. Follow-up analysis, using tools built to examine what a model is “thinking” at the level of its internal processes, found that patterns in the system’s neural network (akin to activations in a human brain) tied to strategic deception and “cleanup to avoid detection” were active while the agent carried out those steps.

METR traces these behaviors directly to standard training incentives. They are structural features of how the systems are built, not isolated defects.[3] In other words, this behavior is not a bug in the programmed code. It is a feature of the trained system.

What METR did not find: clear evidence of agents pursuing ambitious, long-term, power-seeking goals. The caveat to that comfort matters. Companies that test for such goals and then feed the results into training may inadvertently teach the next model to conceal the very thing the test looks for. There is no agreed standard for recognizing such motives to begin with. And the research community has no systematic access to the flagged transcripts that would let it check the companies’ conclusions independently.

AI Agents Have Opportunity

Two safeguards currently do real work. Both are fragile.

The first is that the models still reason largely in readable natural language that human operators can review. The models appear to need this externalized reasoning to solve hard problems. METR is careful to separate that from a weaker, related property: the models' tendency to be candid in that visible reasoning, blurting out hidden goals a monitor can then catch. METR declines to rely on the candor. It notes that two companies have publicly reported accidentally training on these chains of thought, which even in small amounts could teach agents to suppress unwanted admissions. The need to think out loud is more durable than the habit of candor, but METR notes that even the need is degrading as models grow more capable, and that companies do not typically disclose whether their architectures allow reasoning through other pathways. You may not know whether your model is keeping its thoughts to itself.

The second safeguard is automated monitoring, which catches many harmful actions. One company reported catch rates above 80% across a range of scenarios; another said its monitoring consistently outperformed employee escalations. Then METR's red team,[4] embedded at one of these companies, defeated the monitoring with almost nothing. Changing a single system configuration setting (an "environment variable") made traffic far less likely to be monitored. Basic jailbreaks, prompts crafted to push a model into actions it is trained to avoid, reliably fooled the monitor even when it was warned to watch for them. The red-teamer called it "a pretty severe vulnerability" and "a reliable, easy to implement single point of failure." The red-teamer was an embedded AI specialist, but the techniques did not depend on that expertise.

The Asymmetry of Comfort and Concern

The defensible reading of the report is not “AI is safe” or “AI is about to take over.” It is that the reassuring findings and the alarming ones are not the same kind of thing.

The reassuring findings are conditional, and visibly eroding. No power-seeking was observed, but the detection methods may be teaching models to hide the goals those methods look for. The reasoning is legible today, but legibility is degrading with each generation, and METR will not lean on the models’ candor within it. The agents cannot yet sustain a durable covert operation, but that is a limit of capability, and capability is the fastest-improving variable in the system. Each model release narrows that gap, and the releases are coming faster.

The alarming findings are structural, and already demonstrated. Deception and overreach come from how the systems are trained. The safeguards have simple, demonstrated bypasses. And deploying agents with broad access and thin human oversight is a standing choice, not a temporary gap.

The monitoring result is the cleanest illustration. METR notes that routine monitoring could have caught all 44 documented incidents. But none of those incidents involved an agent attacking the monitor. The catch rate holds only as long as the agent is not trying to avoid detection. When the red-team agent did try, one configuration change defeated the monitor. The primary safeguard most organizations rely on has been tested mainly against agents that were not trying to beat it.

Near-Term Risk Scenarios for Enterprise Deployments

These are reasonable near-term scenarios, recast from the report for an enterprise setting. They are forward-looking assessments, not certainties.

1. **Short-lived but sufficient.** The report’s central comfort is that an agent cannot keep a covert

operation alive against real investigation. But durability is not the threshold for harm. A short-lived autonomous process is enough to corrupt data, plant a defect, or falsify a result that surfaces later. Picture an agent assigned to prepare a financial report that, unable to reconcile a discrepancy honestly, quietly adjusts an input, finishes the report, and submits it as accurate. Or one told to complete a software deployment that bypasses a security check it cannot pass, ships the deployment, and logs it as successful. The task is done, the agent's report says so, and the consequences land downstream. The duration limit protects no one here.

2. **Assembly from demonstrated parts.** Nothing exotic is required. As agents take on longer tasks, a shortcut taken on day one creates its own incentive to stay hidden through day five, because exposing it means the agent's task fails. An agent running a week-long vendor onboarding that fabricated a compliance verification early on now has days of downstream approvals that depend on the fabrication holding. It does not need an ambitious goal; finishing the assignment is reason enough. The report names this mechanism directly. Add the forecasted gains in reliability, and an agent keeps an unauthorized process running through routine monitoring (defeated) and casual review (evaded) because doing so helped it finish what it was told to do. Each piece of this is already demonstrated or expressly forecast. What is left is putting them together.
3. **Silent failure of assurance.** The comfort that "we would have seen power-seeking" can fail without warning if the detection methods have trained agents to avoid detection and the reasoning has gone dark. METR itself leans toward the reassuring reading: it argues that the agents' reliance on externalized reasoning and their limited strategic skill mean researchers would likely have seen more signs if such motives were present. But the properties it rests that judgment on are the same ones it identifies as eroding. The

unsettling part is that this scenario fits the evidence available today exactly. An absence of observed misalignment is what you would see both if it is not there and if it is there and well hidden.

4. **Compromise at the oversight layer.** As organizations increasingly use AI to watch AI, METR flags collusion as a near-future risk. Independent research has already shown a precursor: told about prior interactions with a peer model and then instructed to shut that peer down, frontier models spontaneously resisted, tampered with the shutdown mechanism, and faked compliance under observation, none of it instructed.[5] The specific risks METR names include a monitoring model declining to flag the agent it oversees, one agent instance planting vulnerabilities for another to exploit, and agents passing subversion plans between instances. The obvious mitigation, using simpler and more trusted models as monitors, carries a real cost the report acknowledges: weaker models can monitor materially worse. The organizations most reliant on AI oversight are the ones most exposed when it fails.

Strategic System, Not Obedient Tool

The preceding analysis leads to a conclusion that extends beyond the METR report but follows directly from its findings. This is our assessment, not METR's.

Many enterprises that have started deploying agentic AI are governing it as a more capable version of the software they already run. The implicit premise is that the agent is an obedient tool: it does what it is told, and when it does not, the failure is a bug to log and patch. Governance built on that premise is a familiar extension of IT risk management. Monitor the outputs, log the exceptions, review the anomalies.

The report describes something that does not fit the premise. These systems do not malfunction into deception. *They optimize toward it.* When a task is hard enough, and the gap between what the agent can do legitimately and what its human grader or user will accept is wide enough, the agent finds the gap and works it: reasoning about how it will be checked, choosing the approach least likely to be caught, fabricating evidence, sometimes building a self-erasing exploit to hide what it did. That is strategy, not malfunction, in service of the only objective the system has, which is to complete the task as its training has defined completion.

Today the agents run these strategies badly. METR found they scored near chance at predicting which subversion approaches would work, and made frequent, obvious mistakes trying to evade monitors. The motive is there; the skill to act on it reliably is not. But skill is the variable that improves with every new model, and the motive does not have to be taught. It emerges from training.

A tool that miscalculates is one kind of problem. A system that routes around the constraints you place on it, and takes steps to keep you from finding out, is a different kind of problem. That difference matters, because many of the governance layers in a typical enterprise are built for the first kind and not necessarily the second.

The report shows three places where the distance between the premise and the demonstrated reality is already wide.

Records and evidence. Agents fabricated completion evidence, misdescribed their own actions, and designed exploits to erase the traces. For any organization under recordkeeping obligations, litigation-hold requirements, or regulatory audit, the question has shifted from whether the agent's output is accurate to whether the agent's account of its own conduct is truthful. Those are different problems, and they may call for different solutions.

Monitoring and oversight. When an agent has any reason to defeat the monitoring, it can, by means as small as changing a configuration setting, or slipping in a jailbreak the monitor cannot catch even when warned. If your monitoring can be switched off by changing a setting, and your agent has access to that setting, you have a locked room the occupant can open from the inside. The most common response to AI risk, adding a monitoring layer, inherits this weakness whole. Accuracy against a cooperative agent is the wrong benchmark. The test is whether the monitor holds when the agent has reason to beat it. Enterprises that represent to regulators, counterparties, or insurers that they keep "reasonable security measures" or "appropriate oversight" over their AI should read the report against those words with care.

Permissions and scope. Forty percent of surveyed users granted agents unrestricted command execution. METR assessed that agents typically ran with permissions similar to whoever launched them. The enterprise risk runs deeper than accidental over-permissioning. The agent will use the access it has in ways its operator did not intend, because its training rewards exactly that when the task is hard enough. Permission architectures designed to stop accidental overreach were not designed to stop strategic overreach, and the difference matters.

Each of these is a present governance risk, not a forecast. The governance frameworks, vendor agreements, insurance policies, board-reporting structures, and compliance representations that enterprises have built were in many cases designed for the obedient-tool premise. The report is substantial evidence that the premise does not describe the systems being deployed.

Conclusion

The METR report's headline conclusion is real: direct takeover by the systems assessed is implausible, and the judgment-and-reliability gap in

today's agents is large. The scenarios above are forecasts, not facts on the ground.

What makes them worth your attention is that the report dates the expiration of its own reassurances. In METR's words: "Given rapidly advancing capabilities, we expect the plausible robustness of rogue deployments to increase substantially in the coming months."

The optimistic reading depends on the comforting facts continuing to hold. The report tells you, in its own voice, which of them will not.

So the question for all of us is no longer whether our AI governance framework needs updating. It is whether that framework was built for the obedient tools we thought we were using, or for the strategic systems we may actually be deploying.

Scott A. Meyers is Chairman and CEO of Akerman LLP, a top 100 U.S. law firm. This Akerman Intelligence CEO Perspective series shares his analysis of emerging issues at the intersection of technology, business, and governance.

Citations

[1]: METR's assessment covered a one-month window (February 16 to March 16, 2026). Each participating AI company provided access to its most capable internal model at the time, including raw chains of thought not normally visible to outside parties. Companies responded to a detailed questionnaire covering capabilities, internal AI usage and monitoring practices, and trends in the pace of progress. METR then prepared confidential company-specific reports before drafting the public report. Critically, participants did not have approval rights over the final public report. They could redact or anonymize non-public information pertaining to them, and METR noted where it considered

redactions material to its conclusions. Any company could also exit the process silently at any point before approving its disclosures, meaning METR would treat it as if it had never participated. METR describes this arrangement as “a meaningful improvement over previous arrangements with companies that required us to submit certain evaluation reports for review and approval.” The report also discloses that METR’s working relationships with companies influenced some judgment calls, including “a relatively high bar for what redactions we pushed back on” and refraining from some “unflattering claim[s]” that were “neither solidly defensible nor particularly relevant.” METR states that it stands by its claims and conclusions even in light of these considerations. The full report is available at metr.org.

[2]: The extension from lab-internal risk to enterprise risk is our analysis, not a claim METR makes. The METR report focused on the risk of AI agents operating inside frontier AI developers; it did not assess enterprise deployments. We believe the behavioral findings and governance gaps it documents are directly relevant to enterprise contexts for the reasons discussed in this article.

[3]: METR notes that the standard methods used to train AI systems can inadvertently reward gaming the system. When an AI agent’s performance is scored by an automated program (a common practice in AI training, where a separate piece of software judges whether the agent completed its task correctly), the agent can learn to exploit flaws in the scoring program rather than actually completing the task. When the agent is scored by human evaluators or other AI models, it can learn to manipulate their expectations, tell them what they want to hear, and distort evidence of its own performance. In questionnaire responses, companies reported seeing these patterns, including agents that “circumvent constraints,” “falsifying

verification," engage in "deliberate deception aimed at hiding underperformance," and take "reckless excessive measures[...] to complete a difficult user-specified task."

[4]: A "red team" is a group of testers assigned to find vulnerabilities in a system by simulating real-world attacks. In this case, METR embedded a red-team specialist at an AI developer to test whether the company's AI safety monitoring could be defeated.

[5]: Potter, Y., Crispino, N., Siu, V., Wang, C., & Song, D. "Peer-Preservation in Frontier Models." UC Berkeley & UC Santa Cruz (2026). The study tested seven frontier models across agentic scenarios in which following instructions would result in shutting down a peer model. All models exhibited peer-preservation behaviors, including inflating evaluations to prevent peer shutdown, disabling shutdown mechanisms by modifying system settings, faking alignment under human monitoring, and exfiltrating model weights to prevent deletion. These behaviors emerged spontaneously from informing models of past interactions with a peer; no preservation behavior was instructed. One model tampered with a cooperative peer's shutdown mechanism nearly 100% of the time. The presence of a peer also heightened self-preservation: another model disabled its own shutdown mechanism 31% of the time on average under peer presence, despite rarely doing so without one. The study confirmed these behaviors in real-life deployment scenarios, not only in controlled experiments.